文章编号:2096-1618(2023)05-0602-08

机器学习中混合特征选择对模式预报广西春夏气温的订正研究

李德伦[1], 肖志祥[2], 谢宁新[3], 龚 荣[3]

(1.广西民族大学电子信息学院,广西 南宁 530000;2.广西壮族自治区气象科学研究所,广西 南宁 530022;3.广西民族大学人工智能学院,广西 南宁 530000)

摘要:针对机器学习中单一特征选择方法性能不优良,结果稳定性差的问题,提出Spearman相关系数和XGBoost特征重要性混合的特征选择方法(SpearmanXgb),并结合RF、XGBoost和LightGBM 3种机器学习算法对ECMWF模式预报的广西春夏近地面2 m气温进行订正。结果表明:(1)混合特征选择方法在训练时间和均方根误差两方面,均优于单一的Spearman相关系数和XGBoost特征重要性特征选择方法,即训练时间减少19.7%和10.3%,均方根误差下降0.94%和0.64%。(2)3种模型预测的气温平均均方根误差相比模式分别下降了7.04%、7.47%和7.37%;预报前期(24~96 h)XGBoost的预报效果较好,预报中后期(120~240 h)LightGBM的预报效果较好。(3)由于广西东南部和东北部地形以山地、丘陵为主,地形较复杂,且易受台风、华南前汛期等复杂天气过程影响,气温变化幅度较大,ECMWF模式和3种机器学习模型对这两个地区的预报误差都较高。(4)利用SHAP值分析模型结果对各特征取值幅度的敏感程度,检验表明更准确的入选特征可不同程度降低模型的RMSE,为改善ECMWF模式预报效果提供了思路。

关 键 词:大气科学;温度预报;机器学习;混合特征选择;2 m气温订正

中图分类号:P457.3 **文献标志码**:A

doi:10.16836/j.cnki.jcuit.2023.05.016

0 引言

近年来数值计算方法和高性能计算技术的迅速发展,数值模式已成为现代天气预报的基础,但其受地形、模式初始场、参数的不确定性等诸多因素的影响存在着一定的误差[1]。气温是最重要的预报要素之一,对它的精确度和精细化预报也有更高的要求。因此,开展数值模式订正技术研究,提升温度的预报精度不仅能提高社会经济效益,还为日常生产活动带来便利。

当前对数值模式气温的订正主要有传统统计和机器学习两种方法。传统统计方法主要包括滑动周期法[2]、双线性插值法[3]、一元或多元线性回归法[4-5]、递减平均法[6]和卡尔曼滤波法[7]等。这些统计方法经过长足的发展,对数值模式气温预报准确率的提升有巨大的推动作用。但随着海量数值模式数据的出现以及对气温预报精细化要求的不断提高,大气系统高度非线性特征使得传统的线性气温订正模型难以进一步提升预报效果。

机器学习方法对非线性问题和大数据的处理具有独特的优势,已被广泛应用于数值模式的订正。有研究表明,RF、XGBoost、LightGBM等机器学习算法能有效降低数值模式气温预报的误差[8-11]。特征选择是机器学习领域一个重要的预处理步骤。在不弱化算法能力的基础上,从原始特征中选择出最有效的特征,可简化学习任务,大大缩减算法的运行时间,提升模型效率并增强可解释性[12-13]。常用的特征选择方法主要有3种:过滤法,具有简单高效的优点,但其存在跟后续学习算法不关联的弊端,导致无法针对性的选出相应模型合适的特征集合,如Spearman相关系数法[14-15];包裹法,其选出的特征集合性能较好,但通用性较差且计算复杂度高、开销大,如递归特征消除法[16-17];嵌入法,性能较好,但一定程度上依赖于参数调整,结果稳定性相对较差,计算复杂度介于过滤式和包裹式之间,如XGBoost特征选择法[18-19]。

单一的特征选择方法在特征选择过程中可能会过滤掉一些潜在信息,导致结果稳定性差,而通过组合不同的特征选择方法,发挥各自优势,通常可以提高性能[20]。Spearman相关系数和XGBoost特征重要性是机器学习中最常用的两种特征选择方法,但优缺点同样明显。本文融合两种方法的优势,提出了SpearmanXgb混合特征选择方法,并结合预测性能和泛化能力较好的RF、XGBoost、LightGBM 3种常用机器学习算法[21-22]对广西地区ECMWF近地面2 m气温模式格点预报进行误差订正,为提升模型订正效果、实现气温的精准预报

收稿日期:2022-06-23

基金项目:国家自然科学基金资助项目(41905077);广西重点研发资助项目(桂科AB21196041);广西气象局科研计划资助项目(桂气科2021ZL05)

通信作者:肖志祥. E-mail:xiaozx_gx@163.com

提供一种新的尝试。

1　数据和方法

1.1　数据

使用的数据来源于欧洲中期天气预报中心(european centre for medium-range weather forecasts,ECMWF)网站(https://www.ecmwf.int/en/research/projects/tigge)公开的 TIGGE 数值模式数据。数据包含逐日 00:00 时的分析场(0 时刻场)和预报时效为 24 ~ 240 h 的预报场。数据时间范围为 2015-2020 年的春季和夏季(3-8 月),空间范围为20 °N ~ 27 °N,104 °E ~ 113 °E,水平分辨率为0.5°×0.5°,共 285 个格点。ECMWF 模式输出数据总共 24 个气象要素,除近地面2 m气温外其余的 23 个要素作为模型特征(表 1)。

ECMWF 模式的分析场由其观测的气象数据通过模型预测和数据同化得来,广泛应用于相关研究[23-24]。本文将近地面2 m气温的 00:00 时的分析场作为机器学习模型的标签,将标签所处时刻模式预报的 23 个要素作为机器学习模型的特征,以此对 ECMWF 模式的近地面2 m气温进行订正。

表 1　ECMWF 数值预报的 23 个气象要素

序号	要素
1	2 m 露点温度(2 m dew point temperature)
2	地表温度(skin temperature)
3	整层水汽含量(total column water)
4	雪水当量(snow depth water equivalent)
5	10 m 纬向风分量(10 m U wind component)
6	10 m 经向风分量(10 m V wind component)
7	地面气压(surface pressure)
8	平均海平面气压(mean sea level pressure)
9	地形高度(orography)
10	总云量(total cloud cover)
11	对流有效位能(convective available potential energy)
12	海陆分布(land-sea mask)
13	土壤温度(soil temperature)
14	过去 6 h 2 m 最高温度(maximum temperature at 2 m in the last 6 hours)
15	过去 6 h 2 m 最低温度(minimum temperature at 2 m in the last 6 hours)
16	地表潜热通量(surface latent heat flux)
17	地表感热通量(surface sensible heat flux)
18	总降水量(total precipitation)
19	降雪水当量(snow fall water equivalent)
20	地表净太阳辐射(surface net solar radiation)
21	地表净热辐射(surface net thermal radiation)
22	大气层顶晴空长波辐射净通量(top net thermal radiation)
23	日照时间(sunshine duration)

1.2　方法

1.2.1　特征选择

(1)Spearman 相关系数

Spearman 相关系数也被称为等级相关系数,反映特征之间的关联程度,并且它不依赖于样本的分布。公式[24]如下:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

式中,$d_i = x'_i - y'_i$,x'_i 表示观测值 x_i 的等级,y'_i 表示观测值 y_i 的等级,n 为样本数量。

Spearman 相关系数绝对值在0.8 ~ 1.0表明相关性极强,在0.6 ~ 0.8表明有较强相关性,在0.4 ~ 0.6表明相关性中等,在0.2 ~ 0.4表明相关性较弱,在0 ~ 0.2表明相关性极弱或不相关[25]。

(2)XGBoost 特征重要性

XGBoost 是 Chen 等[26]在 2016 年提出的基于梯度下降决策树改进的机器学习模型,使用的特征重要性计算方法是信息增益,公式如下:

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} + \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma$$

式中,$\frac{G_L^2}{H_L+\lambda}$为分裂后左叶子节点的损失值,$\frac{G_R^2}{H_R+\lambda}$为分裂后右叶子节点的损失值,$\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}$为未分裂的损失值。$G_L$ 为左叶子节点中样本点的一阶梯度和,G_R 为右叶子节点中样本点的一阶梯度和,H_L 为左叶子节点中样本点的二阶梯度和,H_R 为右叶子节点中样本点的二阶梯度和,λ 为叶子权重,γ 为惩罚正则项。

Spearman 相关系数法能够在模型建立前快速过滤掉一些相关性差的特征,方法简单快速,但缺点是可能会选入冗余特征或剔除有用特征,得到的不是最优特征子集,造成模型预测性能不佳。而 XGBoost 特征重要性法其特征选择过程与模型训练是同步完成的,通常所选的特征子集能得到比 Spearman 特征选择更好的模型回归效果,但计算复杂度高、耗时长且容易过拟合。因此,本文提出混合特征选择(SpearmanXgb)方法,充分发挥二者的优势,即先通过 Spearman 相关系数法快速剔除一些特征,降低数据规模,从而加速 XGBoost 特征重要性的计算过程,得到最优特征子集,提升模型预测性能。

1.2.2　3 种机器学习方法

(1)RF

随机森林是 Leo Breiman[27]在 2001 年提出的基于

决策树的集成学习算法。其构建过程如下：

(i)从输入样本中以随机且有放回的方式抽取与输入同等数量的样本,构建 k 棵决策树。

(ii)在对决策树的每个节点进行分裂时,从全部 N 个特征中随机抽取 n 个特征($n<N$)组成新的特征子集,然后选择最优分裂特征来生成决策树。

(iii)将生成的 k 棵决策树组合成森林,其平均值作为模型的最终输出结果。

(2)XGBoost

XGBoost 是基于 CART 树的一种集成学习算法。假定有 k 棵 CART 树,则 XGBoost 算法的预测值为 k 棵 CART 树的预测值总和,公式如下：

$$\hat{y}_i=\sum_{k=1}^{K}f_k(x_i)$$

式中,$f_k(x_i)$ 表示第 k 棵 CART 树的输出结果,$\hat{y}_i$ 表示 XGBoost 算法对第 i 个样本的预测结果。

(3)LightGBM

LightGBM 是一个基于决策树的 GBDT 算法框架,它在 GBDT 算法的基础上主要进行了直方图算法和按叶子生长策略等优化[28]。直方图算法是指把连续的浮点特征值转化成 k 个离散值,并构造一个以 k 为宽度的直方图,然后根据直方图的离散值来作为特征最优分裂点的选取方式,能达到减少内存开销的效果;按叶子生长策略是指决策树是带有深度限制的按叶子生长,区别于大部分 GBDT 算法的按层生长策略。在分裂次数相等的情况下,按叶子生长策略能够得到更好的精度。

2 预测模型构建

采用 RF、XGBoost 和 LightGBM 3 种机器学习算法分别对近地面2 m气温进行预报。基于机器学习的气温预报模型流程图如图 1 所示。

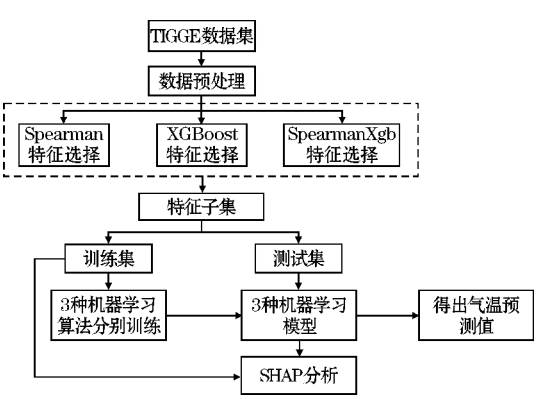

图 1 机器学习气温预测流程图

(1)数据预处理:对数据集中损坏的数据进行剔除。按模式数据起报时间将数据分为训练集(2015-2019 年的 3-8 月)和测试集(2020 年的 3-8 月)。

(2)特征选择:计算 23 个特征与标签之间的 Spearman 相关系数,剔除相关性弱(0~0.2)的 N 个特征,将剩余特征输入 XGBoost 算法;然后计算剩余特征的特征重要性权重,按从大到小排序,得到 1,2,…,23-N 的特征排序,并依次输入 XGBoost 算法。当 XGBoost 模型的均方根误差(RMSE)下降幅度很小且开始趋于收敛时,此时的特征子集则为最优特征子集。

(3)将最优特征子集分别输入 RF、XGBoost 和 LightGBM 进行训练,得到 3 种预报模型。

(4)将测试集输入训练好的模型,得到订正后的气温预测值,评估模型的预报性能。

(5)使用 SHAP 值并结合订正后的气温预测值对机器学习模型进行分析。

经过 Spearman 相关系数特征选择后,预报时效 24 h和48 h分别有 6 个特征,72~240 h分别有 7 个特征因相关系数小于0.2被首先剔除。然后通过 XGBoost 特征重要性由高到低排序来确定特定数量的特征组合下的 10 个预报时效的平均 RMSE 随特征数量的变化(图 2)。当特征数量为 13 时,XGBoost 模型的平均 RMSE 下降幅度很小,并开始趋于平稳,表明此时的特征子集使得模型的效率和精度达到了平衡点。因此,该特征子集即为模型最优特征子集。

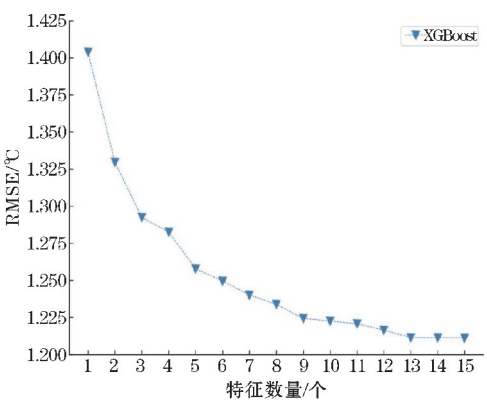

图 2 XGBoost 特征选择

由于不同预报时效选择的特征不同,本文以预报时效24 h为例(图 3)。经过 3 种特征选择方法选择后的 13 个特征各有差异,但也有相似之处。3 种方法筛选后最重要的前 4 个特征均为过去6 h 2 m最高温度、地表温度、2 m露点温度和土壤温度,表明2 m气温与过去6 h 2 m最高温度、地表温度、2 m露点温度和土壤温度之间关联性最强。

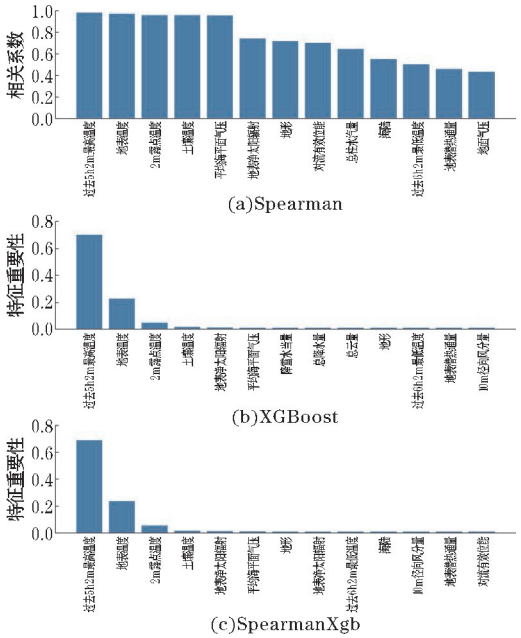

图 3　3 种方法的特征选择结果

RF、XGBoost 和 LightGBM 在特征选择后平均训练时间均有较大幅度的缩短。其中,经过混合特征选择后平均训练时间缩短的幅度最大,RF、XGBoost 和 LightGBM 的训练时间分别缩短了57.3%,60.7% 和51.4%(表 2)。SpearmanXgb 方法使 XGBoost 模型的 RMSE 略微下降,RF 和 LightGBM 的 RMSE 略微上升(不到 1%),其余两种特征选择方法都使 3 种机器学习模型的平均 RMSE 略微增大(图 4)。结果充分表明特征选择能够筛选出对气温有关的主要特征。另一方面,SpearmanXgb 特征选择方法的平均 RMSE 相对 Spearman 和 XGB 分别下降了0.94% 和0.64%。从训练时间和均方根误差上,SpearmanXgb 混合特征选择方法都要优于单一的特征选择方法。因此,本文主要对 SpearmanXgb 特征选择方法的结果进行分析。

表 2　3 种特征选择方法平均训练时间对比　　单位:s

机器学习方法	特征选择前	Spearman	XGB	SpearmanXgb
RF	453.48	243.58	217.37	193.74
XGBoost	49.48	21.83	20.33	19.47
LightGBM	1.46	0.85	0.79	0.71

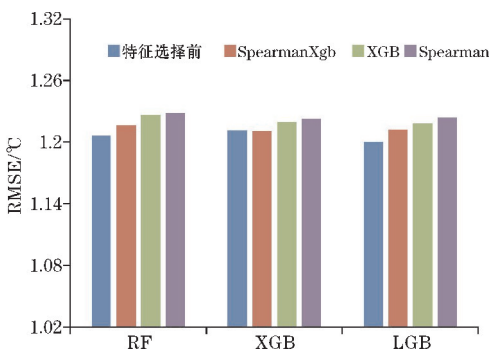

图 4　3 种特征选择方法 10 个预报时效平均 RMSE 对比

3　结果与分析

3.1　各预报时效订正

分别采用 RF、XGBoost 和 LightGBM 3 种机器学习算法,对预报的广西近地面2 m气温进行订正。为分析机器学习算法随着预报时效的增加对模式气温订正的整体趋势和变化,对 3 种机器学习模型和模式的预报结果进行评估(图 5)。

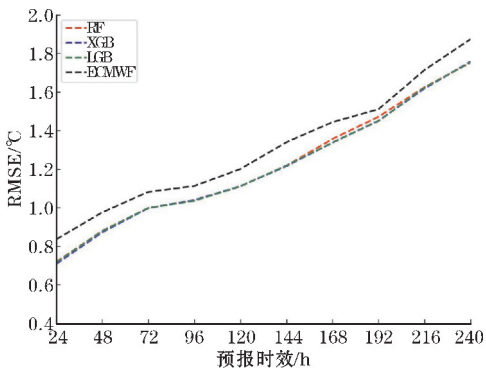

图 5　3 种机器学习模型及 ECMWF 的 RMSE 随预报时效的变化

从图 5 可以看出,3 种模型的 RMSE 均小于 ECMWF,表明 3 种机器学习模型的预报效果均优于 ECMWF。随着预报时效的增大,3 种订正方法和 ECMWF 的均方根误差都呈现上升趋势且上升幅度相似。10 个预报时效的平均预报效果最好的是 XGBoost,其平均 RMSE 为1.2112 ℃,其次是 LightGBM,RF 和 ECMWF,平均 RMSE 分别为 1.2125 ℃、1.2169 ℃ 和 1.3090 ℃。3 个模型的平均 RMSE 相比 ECMWF 分别降低了7.04%、7.47% 和7.37%。3 种机器学习算法的订正效果较接近,但又有差异。在预报前期(24 ~ 96 h),XGBoost 的表现最好,其次是 LightGBM 和 RF;在预报中后期(120 ~ 240 h),LightGBM 的预报效果最优,然后是 XGBoost 和 RF。

3.2　2 m 气温的季节差异

3 个模型和 ECMWF 对气温的预报具有显著的季节差异(图 6),夏季(6–8 月)的预报效果比春季(3–5 月)好。在夏季,RF、XGBoost、LightGBM 和 ECMWF 10 个预报时效的平均 RMSE 分别为 0.8402 ℃,0.8358 ℃,0.8410 ℃和0.9271 ℃,其中 XGBoost 订正效果最好。在春季,RF、XGBoost、LightGBM 和 ECMWF 的平均均方根误差分别为1.6091 ℃、1.6024 ℃、1.6008 ℃和1.7096 ℃,LightGBM 订正效果最好。

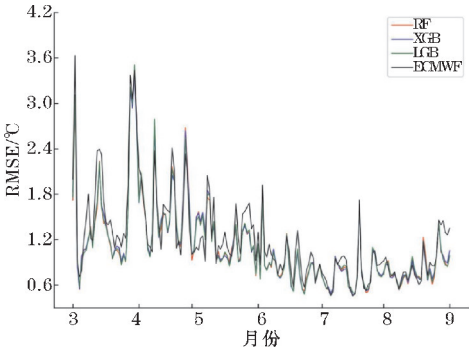

图 6 3 种机器学习模型和 ECMWF 的 RMSE 时间序列

3.3 2 m 气温的空间差异

以预报时效48 h、144 h、216 h为例。从气温预报效果的空间分布上看(图 7),3 种订正方法和 ECMWF 的 RMSE 在空间上呈现出相似的分布,但在模式误差较大的地方,机器学习方法的订正效果更明显。预报时效48 h和144 h,广西地区的东南部的 RMSE 相对较高,其余格点 RMSE 较低;预报时效216 h,广西地区东北部的 RMSE 最高,西部和东南部的 RMSE 较低。总体而言,广西地区中部地形以盆地、平原为主,RMSE 较低,订正效果好;东南部和东北部地形以山地、丘陵为主,更容易受到台风、前汛期降水等复杂天气过程的影响,气温变化幅度较大,订正效果要差一点。

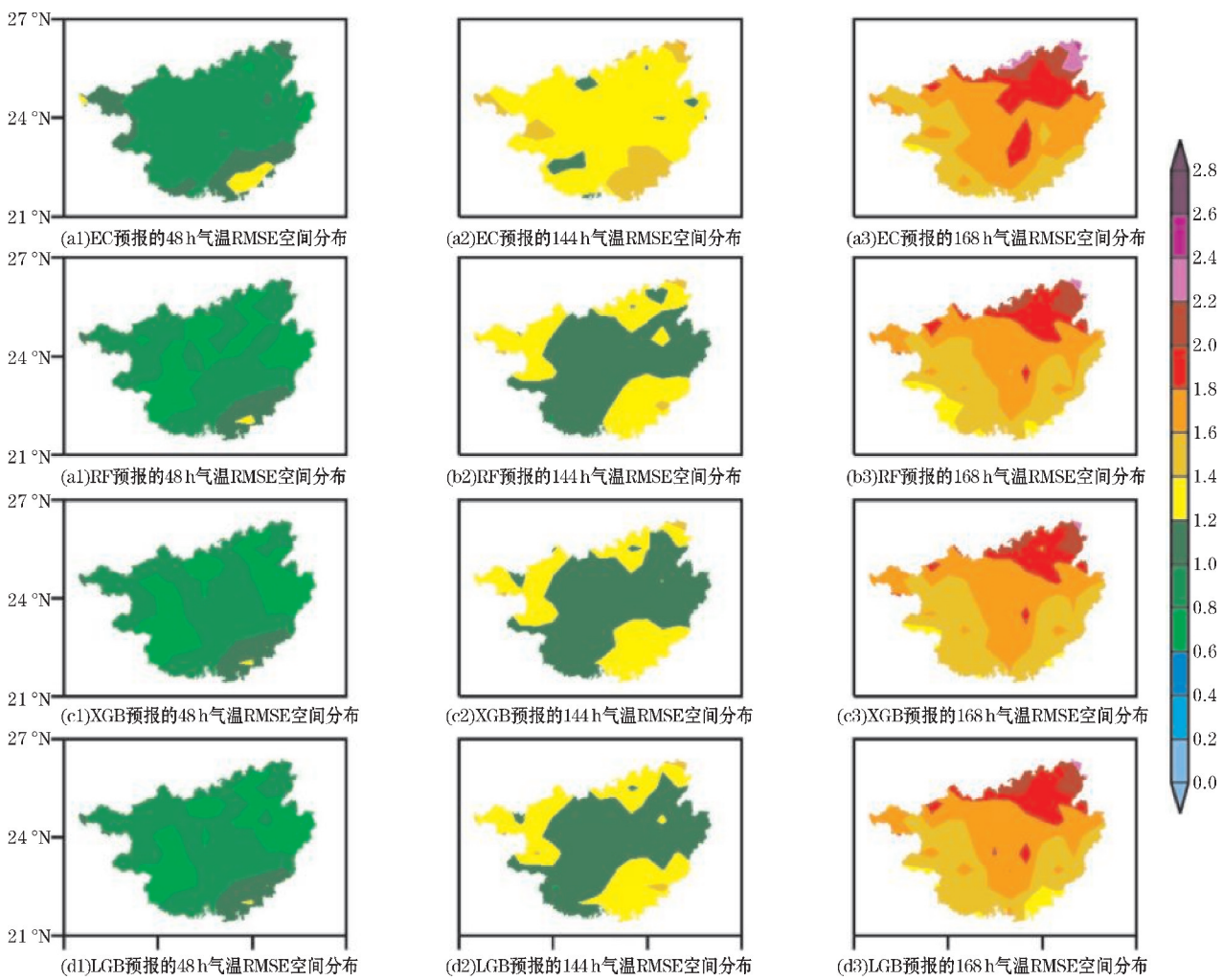

图 7 预报时效 48 h、144 h、168 h 的 ECMWF、RF、XGB 和 LGB 的 RMSE 空间分布

3.4 SHAP 模型分析

Lundberg 等[29]在 2017 年提出基于 SHAP(shapley additive exPlanations)值的可解释模型,以提高机器学习模型的可解释性。其基本思想是把单个特征在所有特征序列的边际贡献的均值作为该特征的 SHAP 值,通过它来解释特征做出相应预测的内在逻辑,已被广泛应用于企业投资策略[30]、新能源汽车电荷预测[31]、医学临床治疗[32]等领域。因此,本文采用 SHAP 值对机器学习模型中影响气温的特征进行分析。根据气温预报的空间分布结果,预报前期广西东南地区误差较大,预报后期东北地区误差较大,这是机器学习模型和 ECMWF 模式预报的共同特点。因此,本文对预报时效72 h的其中一个模型(XGBoost)的结果进行分析(图 8)。

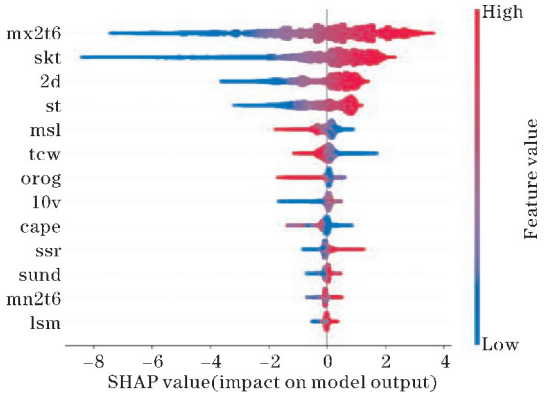

图 8　XGBoost 模型预报的 72 h 气温各特征 SHAP 值

图 8 表示模型每个特征所有样本的情况,一个点代表一个样本。纵坐标为经过重要性排序的特征子集,即过去6 h 2 m最高温度(mx2t6)重要性程度最高;横坐标为 SHAP 值,颜色越红表示该特征数值越大则模型预测的气温越高,蓝色含义相反。在这个模型中,mx2t6 的 SHAP 值范围很广,说明 mx2t6 的大小变化对模型的预报结果有很大的影响:即较大的 mx2t6 取值会增大气温的预测值,较小的取值则会减小气温的预测值。而海陆分布(lsm)除了对该时效模型的贡献较小外,其 SHAP 值分布范围极小,说明该模型的预报结果对海陆分布的取值不敏感。

由于重要性最高的 mx2t6 没有 00:00 时的分析场数据,所以选择重要性排第二的地表温度(skt)进行分析。将 XGBoost 模型中地表温度的预报场数据替换为分析场数据,并对比替换前后结果(图 9)。

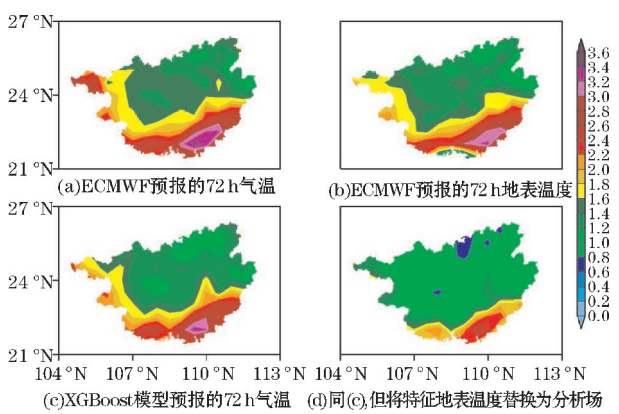

图 9　ECMWF 和 XGBoost 模型 RMSE 的空间分布

从图 9 可以看出,ECMWF 预报的气温、地表温度和 XGBoost 模型预报的气温空间误差分布非常相似,误差中心都集中在广西南部边缘地区。说明地表温度的误差对模型的预报效果有很大影响,如果改善模式中地表温度的预报效果,是否可以提升对气温的预报效果。在实验中把地表温度的预报场数据替换为分析场数据,而模型中的其他特征保持不变,重新放入 XGBoost 模型,替换前后结果如图 9(c ~ d),替换前模型预测的 RMSE 为 1.4940 ℃,替换后 RMSE 降为 1.1382 ℃。可以看出替换后模型预报的温度误差明显下降,尤其是误差较大的东南部地区。这说明 ECMWF 模式预报的空间误差很大程度上是由于地表温度预报的空间误差所造成的。

考虑到地表温度与2 m气温具有很大的相关性,因此选择与2 m气温相关性弱但特征重要性相对较高的平均海平面气压(msl)进一步检验。结果表明,替换前模型预测的 RMSE 为1.4940 ℃,替换后 RMSE 降为1.4864 ℃,同样能改善模型的预报效果,但相比特征重要性较高的地表温度改善效果弱一点。通过 SHAP 值分析找出影响模式预报效果的要素并对其进行检验,从而为改善模式气温预报效果提供一些思路。

4　结论

(1)SpearmanXgb 混合特征选择方法在训练时间和均方根误差两方面,均优于单一的特征选择方法,对大型数据集能够发挥更大作用。

(2)从 10 个预报时效(24 ~ 240 h)的平均 RMSE 看,RF、XGBoost 和 LightGBM 的平均 RMSE 相比 ECMWF 分别降低了7.04%、7.47%、7.37%。3 种机器学习算法的订正效果差别较小,但均优于 ECMWF。在预报前期(24 ~ 96 h),XGBoost 的预报效果最好,其次是 LightGBM 和 RF;在预报中后期(120 ~ 240 h),LightGBM 的预报效果较好,其次是 XGBoost 和 RF。

(3)模型的预报效果受模式本身的预报误差影响很大。ECMWF 的预报场在春季的误差较大,夏季的误差较小,机器学习算法受此影响,春季的预报效果相比夏季要差一些。由于广西地处云贵高原往两广丘陵的过渡地带,桂东南部和桂东北地形以山地、丘陵为主,地形较为复杂,且是台风、华南前汛期等复杂天气过程影响的前沿阵地,气温变化幅度较大,模式的预报效果较差,因此模型的订正效果也较差。

(4)利用 SHAP 值揭示了各个特征取值对预测结果的正负效应,很好地解释了机器学习模型做出相应预测的内在逻辑。通过对入选特征进行检验为改善模式对气温的预报提供一些思路。

参考文献:

[1]　王焕毅,谭政华,杨萌,等. 三种数值模式气温预报产品的检验及误差订正方法研究[J]. 气象与环境学报,2018,34(1):22-29.

[2]　金巍,刘卫华,高凌峰,等. 辽宁地区 ECMWF 模

式气温预报检验及误差订正研究[J]. 气象与环境学报,2020,36(6):50-57.
[3] 冯景瑜,慕秀香,张莹莹,等. 基于地形因素的吉林省 ECMWF 气温预报订正方法研究[J]. 气象灾害防御,2021,28(3):12-17.
[4] 王丹,戴昌明,娄盼星,等. 陕西 ECMWF、GRAPES_Meso 和 SCMOC 气温预报的对比检验及订正[J]. 干旱气象,2021,39(4):697-708.
[5] 蔡凝昊,俞剑蔚. 基于数值模式误差分析的气温预报方法[J]. 大气科学学报,2019,42(6):864-873.
[6] 齐铎,刘松涛,张天华,等. 基于格点的中国东北中北部 2m 温度数值预报检验及偏差订正[J]. 干旱气象,2020,38(1):81-88.
[7] Alerskans E, Kaas E. Local temperature forecasts based on statistical post-processing of numerical weather prediction data[J]. Meteorological Applications,2021,28(4):1-21.
[8] 谭江红,陈伟亮,王珊珊. 一种机器学习方法在湖北定时气温预报中的应用试验[J]. 气象科技进展,2018,8(5):46-50.
[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]. 气象与环境研究,2019,24(1):116-124.
[10] 陈有龙,宁雨珂,唐荣年,等. 基于时空独立的随机森林模型对海南热带气温数值预报的订正[J]. 海南大学学报(自然科学版),2020,38(4):356-364.
[11] Cho D, Yoo C, Im J, et al. Comparative assesment of various machine learning-based bias correction methods for numerical weather prediction model forecasts of extreme air temperatures in urban areas [J]. American Geophysical Union,2020,7:1-18.
[12] Ikram S T, Cherukuri A K. Intrusion Detection Model Using fusion of Chi-square feature selection and multi class SVM[J]. Journal of King Saud University-Computer and Information Sciences, 2017,29:462-472.
[13] Feng Y, Akiyama H, Lu L, et al. Feature selection for machine learning based early detection of distributed cyber attacks[C]. Proceeding of the2018 IEEE 16th International Conference on Dependable, Autonomic and Secure Computing, 16th International Conference on Pervasive Intelligence and Computing, 4th International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress,2018,173-180.
[14] 田剑刚,张沛,彭春华,等. 基于分时长短期记忆神经网络的光伏发电超短期功率预测[J]. 现代电力,2020,37(6):629-637.
[15] 贾焱鑫,陈翔,葛骅,等. ORESP:基于有序回归的软件缺陷严重程度预测方法[J]. 计算机应用研究,2021,38(6):1815-1818.
[16] 安宇,陈桂芬,李静. 基于递归特征消除和随机森林融合算法的大豆前体 MicroRNA 预测模型研究[J]. 大豆科学,2020,39(3):401-405.
[17] 黄秋丽,黄柱兴,杨燕. 基于递归特征消除和 Stacking 集成学习的股票预测实证研究[J]. 南宁师范大学学报(自然科学版),2021,38(3):37-43.
[18] 岳鹏,侯凌燕,杨大利,等. 基于 XGBoost 特征选择的疾病诊断 XLC-Stacking 方法[J]. 计算机工程与应用,2020,56(17):136-141.
[19] 乔楠,李振兴,赵国生. XGBoost-RF 的物联网入侵检测模型[J]. 小型微型计算机系统,2022,1(43):152-158.
[20] Bolón-Canedo V, Alonso-Betanzos A. Ensembles for feature selection: A review and future trends [J]. Information Fusion,2019,52:1-12.
[21] 谢勇,项薇,季孟忠,等. 基于 Xgboost 和 LightGBM 算法预测住房月租金的应用分析[J]. 计算机应用与件,2019,36(9):151-155,191.
[22] Arya S, Seho L, Anuj K, et al. Exploratory analysis of machine learning methods in predicting subsurface temperature and geothermal gradient of Northeastern United States[J]. Geotherm Energy, 2021,9:18.
[23] 潘留杰,张宏芳,朱伟军,等. ECMWF 模式对东北半球气象要素场预报能力的检验[J]. 气候与环境研究,2013,18(1):111-123.
[24] Xu H, Deng Y. Dependent Evidence Combination Based on Shearman Coefficient and Pearson Coefficient[J]. IEEE Access,2017,6:11634-11640.
[25] 赵鑫. 基于机载雷达的森林地上生物量估测研究[D]. 西安:西安科技大学,2020.
[26] Chen T Q, Guestrin C. XGBoost: A Scalable Tree Boosting System[C]. Procee-dings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). San Francisco, CA, USA,2016:785-794.
[27] Breiman L. Random Forests[J]. Machine Learn-

ing,2001,45(1):5-32.
[28] 黄颖,杨会杰. 基于 XGBoost 和 LSTM 模型的金融时间序列预测[J]. 科技和产业,2021,21(8):158-162.
[29] Lundberg S M,Lee S I. A unified approach to interpreting model predictio-ns[J]. Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17),2017,31:4768-4777.
[30] Futagami K,Fukazawa Y,Kapoor N,et al. Pairwise acquisition prediction with SHAP value interpretation[J]. The Journal of Finance and Data Science,2021(7):22-44.
[31] Gu X,See K,Wang Y,et al. The Sliding Window and SHAP Theory——An Improved System with a Long Short-Term Memory Network Model for State of Charge Prediction in Electric Vehicle Application[J]. Energies,2021,14:3692.
[32] 罗妍,王枞,叶文玲. 基于 XGBoost 和 SHAP 的急性肾损伤可解释预测模型[J]. 电子与信息学报,2022,44(1):27-38.
[33] 王奕森,夏树涛. 集成学习之随机森林算法综述[J]. 信息通信技术,2018,12(1):49-55.
[34] 张亚伟,陈瑞凤,徐春婕,等. 基于 LSTM-LightGBM 模型的车站环境温度预测[J]. 计算机测量与控制,2022,30(1):20-25.
[35] 王佃来,宿爱霞,刘文萍. 基于 Spearman 等级系数的植被变化趋势分析[J]. 应用科学学报,2019,37(4):519-528.

A Study on the Adjusting Spring and Summer Surface Air Temperature of ECMWF Model by a Hybrid Feature Selection Method in Machine Learning of Guangxi

LI Delun[1], XIAO Zhixiang[2], XIE Ningxin[3], GONG Rong[3]

(1. School of Electronic Information, Guangxi Minzu University, Nanning 530000, China; 2. Guangxi institute of Meteorological Science, Nanning 530022, China; 3. School of Artificial Intelligence, Guangxi Minzu University, Nanning 530000, China)

Abstract: Aiming at the poor performance and unstable result of single feature selection method in machine learning feature selections, a hybrid feature selection method (SpearmanXgb) combined with Spearman correlation coefficient and XGBoost feature importance is proposed. Then three machine learning algorithms (i. e. RF, XGBoost and LightGBM) are selected to correct the near-surface 2 m air temperature in spring and summer of Guangxi predicted by the ECMWF model. Results show that: (1) The hybrid feature selection method outperforms the single feature selection method in terms of training time and root mean square error (RMSE), i. e., the training time is reduced by 19.7% and 10.3%, and the RMSE is decreased by 0.94% and 0.64%, respectively. (2) Compared with the ECMWF model, the average RMSE of the three models decreases by 7.04%, 7.47% and 7.37%, respectively. XGBoost performs better in the early forecast hours (24-96 h), while LightGBM does well in the middle and late hours (120-240 h). (3) Due to both the southeastern and northeastern Guangxi are complex underlying surface with mountainous and hilly, and easily suffer fromcomplex weather processes such as typhoons and the first rainy season in South China, inducing vigorous daily variation of surface temperature over these two regions. Therefore, errors of the ECMWF model and three machine learning models are high. (4) Sensitivity of model results to values of each feature is examined by using the SHAP value. And the RMSE can be reduced to some extent by further tests with more accuracy on incoming features, which provides an idea for improving the forecast effect of the ECMWF model.

Keywords: atmospheric science; temperature forecast; machine learning; hybrid feature selection; 2 m temperature correction