罗威,罗烨泓,王威. 基于机器学习的韶关地区短期日平均气温研究[J]. 陕西气象,2022(6):21-25.
文章编号:1006-4354(2022)06-0021-05

基于机器学习的韶关地区短期日平均气温研究

罗 威[1],罗烨泓[2],王 威[3]
(1. 兴宁市气象局,广东兴宁 514500;2. 韶关市气象局,广东韶关 512028;
3. 深圳市气象局,广东深圳 518000)

摘 要:利用1965—2017年韶关地区8个站点的日平均气温观测资料,发现将过去连续7 d的日平均气温分别作为逐步多元线性回归、LightGBM(light gradient boosting machine)以及BP-NN(back propagation neural network)算法(机器学习)的自变量可最准确地预报出未来1~3 d的日平均气温。据此分别构建了三种短期气温预报模型,并系统地探讨了其适用性。主要结果如下:(1)三种模型的预报准确度均较高,机器学习方法在预报正确率(绝对误差(absolute error)小于2 ℃的天数占比)、相关系数(R)和平均绝对误差(mean absolute error, MAE)上均要优于逐步多元线性回归法。其中LightGBM的预报效果最优,其1~3 d的预报正确率分别为84.38%、69.86%、61.37%,对应预报值与测量值之间的R分别为0.98、0.94、0.93,MAE分别为1.17、1.76、2.00 ℃。(2)冬春季、秋冬季日平均气温较大的波动性是导致该时期三种模型MAE总体偏大的主要因素,但LightGBM仍具有最优的预报稳定性,其绝对误差的方差最低。

关键词:机器学习;气象业务;短期预报;气温
中图分类号:P457.3 **文献标识码**:A

自工业革命以来,人类活动所导致的大气二氧化碳排放量剧增,更多的热量被截留在大气层内,致使地球气温增高。受此影响,全球平均表面温度自工业革命以来表现出显著的上升趋势。全球变暖会引起冰雪融化、冻土消融、海平面上升、极端天气频发等,其严重威胁了全球自然生态系统,乃至人类的生存[1-6]。为此,开展针对全球气温的精细化预报研究具有重要现实意义。

近年来,监督类机器学习(树模型、支持向量机、神经网络等)和深度学习算法等各类机器学习算法已在气象短临[1]、中短期[2]乃至长期预报[3]等领域发挥了积极的重要作用,其在相关领域中的表现要显著优于统计和主观经验等传统方法。相较于其他方法而言,监督类机器学习算法能够更有效综合应用来自观测、数值模式等多源数据,据此可更有效地提取大气的非线性演化特征,进而提升数值模式的天气预报效果。然而,受地形、模式参数化方案等不确定性因素的共同影响,目前数值预报模式对气温的预报尚存在一定的偏差,尤其对于中小城市、偏远以及具有复杂地形的地区而言,预报方法通常仅依赖于数值模式,且缺乏有效的补充与优化方法。韶关地区地处南岭山脉南麓,属于亚热带季风气候区,夏无酷暑,冬无严寒,雨量充沛,日照温和,气候条件优越,生态与旅游资源丰富,并且是广东省重要的粮食蔬菜供应地、农业大市。因此,针对韶关地区等中小城市气温预报业务算法的开发对推动气象业务的精细化预报具有重要意义。为此,仅利用1965—2017年韶关地区8个站点的日平均气温观测资料,构建了基于历史数据驱动与机器学习方法的短期气

收稿日期:2021-07-26
作者简介:罗威(1987—),男,广东兴宁人,学士,助理工程师,主要从事气象服务工作。
通信作者:罗烨泓(1988—),男,广东丰顺人,硕士,高级工程师,主要从事气象服务工作。
基金项目:四川省教育厅气象灾害预报预警与应急管理研究中心科研项目(ZHYJ19-YB03)

温预报模型,以期为中小城市乃至偏远地区的气温业务预报的改进提供参考,为当地生态环境的改善、社会经济的发展提供科学决策依据。

1 资料和方法

1.1 资料介绍

基于1965—2017年韶关地区8个台站(南雄、曲江、乐昌、仁化、乳源、始兴、翁源和新丰)的日平均气温,求取上述8个站点日平均气温的平均值。台站的基本信息见表1。

表1 韶关地区各个台站基本信息

站号	站名	类别	海拔高度/m
57996	南雄	基准气候站	149.7
57988	乐昌	一般站	143.2
57989	仁化	一般站	112.7
59081	乳源	一般站	131.1
59082	曲江	基本气象站	121.3
59090	始兴	一般站	143.9
59094	翁源	一般站	184.1
59097	新丰	基本气象站	269.3

1.2 方法介绍

本文用于短期气温预报的算法分别有传统回归方法——逐步多元线性回归法和机器学习方法——LightGBM(light gradient boosting machine)和BP-NN(back propagation neural network)。短期日平均气温预报的流程主要可概括如下:(1)将连续n d的日平均气温(T_1, T_2, ……, T_n)作为自变量,将第$n+1$,$n+2$,$n+3$天所分别对应的日平均气温T_{n+1}, T_{n+2}, T_{n+3}作为因变量来构建预报T_{n+1}, T_{n+2}, T_{n+3}的模型。(2)n是决定T_{n+1},T_{n+2},T_{n+3}预报精度的重要参数,因此在1~365 d的范围内对n进行遍历,最终确定了$n=7$时可使T_8,T_9,T_{10}的预报效果达到最优。即将过去连续7 d的日平均气温(T_1, T_2, ……, T_7)作为自变量,以未来3 d日平均气温(T_8,T_9,T_{10})作为因变量,来构建相应的日平均气温预报模型。

1.2.1 逐步多元线性回归 逐步多元线性回归[4]是基于最优的自变量来构建回归模型,其较好地解决了传统多元线性回归法中所存在的共线性问题[5],从而有利于回归模型获得更加精确的计算效果。目前,多元逐步线性回归法已被广泛应用于气象领域,其详细流程可详见文献[4]。

1.2.2 LightGBM(light gradient boosting machine)算法 LightGBM[6]是微软在2017年提出的基于GBDT的算法。相较于传统树模型而言,LightGBM预报精度更高、模型泛化性更强、计算效率更快,适用于大规模数据集的并行计算。LightGBM算法的原理参照文献[6]。

1.2.3 BP-NN(back propagation neural network)算法 BP-NN算法的原理参照文献[7]。为避免权重参数过多而引起过拟合,本文的BP-NN仅5层,由1层输入层、3层隐藏层、1层输出层组成,其中隐藏层的特征维度为100。为了增强BP-NN的训练/预报效果,采用如下优化机制。(1)Kaiming初始化方案[8];(2)Relu激活层[9];(3)L2正则化方案[10],权重衰减系数为0.000 1;(4)Adam优化算法;(5)余弦退火的学习率衰减策略:学习率随训练迭代次数的增大而呈现余弦的周期形态变化,变化的周期为100次迭代,学习率最大值为0.001,最小值为0.000 01;(6)均方误差(mean square error, MSE)的损失函数;(7)自变量与因变量均采用最大最小值归一化。

2 结果与分析

2.1 T_{n+1}, T_{n+2}, T_{n+3}与$T_{1\sim n}$之间的相关性

表2为T_{n+1}, T_{n+2}, T_{n+3}与$T_{1\sim n}$($n=7$)之间的相关性,从表2可见,随着自变量与因变量之间时间间隔的增大,其对应的相关性逐渐降低,但总体仍十分显著。因此,其显著的相关性为以历史日平均气温作为自变量来预报未来短期内的日平均气温奠定了基础。

表2 未来日平均气温与历史日平均气温之间的相关性

历史气温	T_7	T_6	T_5	T_4	T_3	T_2	T_1
T_8	0.97	0.92	0.88	0.86	0.85	0.84	0.83
T_9	0.92	0.88	0.86	0.85	0.84	0.83	0.83
T_{10}	0.88	0.86	0.85	0.84	0.83	0.83	0.82

2.2 三种短期日平均气温预报模型的构建

取1965—2014年的日平均气温作为逐步多

元线性回归法建模数据集,并作为 LightGBM 与 BP-NN 模型的训练集,2015—2016 年的作为两种机器学习模型的验证集,2017 年的则作为上述三种模型适用性分析的测试集。其中训练集用于训练构建上述三种模型,验证集用于监控机器学习模型的训练情况。当 LightGBM 与 BP-NN 超过 100 次的训练迭代而验证集误差不再下降时则停止训练,以防过拟合。两种模型均训练迭代 1 000次,训练结束后保存验证集误差最低的模型。

2.2.1　基于逐步多元线性回归的模型　日平均气温之间显著的相关性(表 2)极易带来共线性问题,进而导致普通线性回归方法存在计算的不稳定性问题。为此,本文采用了逐步线性回归方法来构建短期日平均气温预报模型。

通常情况下,可认为方差膨胀因子≤10 时不存在明显的共线性[5]。据此,预报未来 1～3 d 短期日平均气温的多元逐步线性回归方程如下。

$$T_8=0.656+0.967T_7, \quad (1)$$

$$T_9=1.024+0.757T_7+0.192T_2, \quad (2)$$

$$T_{10}=1.551+0.645T_7+0.278T_3。 \quad (3)$$

通过逐步多元线性回归分析发现,在满足方差膨胀因子≤10 的条件下,T_8 的自变量因子仅为 T_7,也表明 $T_{1\sim6}$ 与 T_7 之间存在显著的共线性;T_9 的自变量因子为 T_7 和 T_2,其对应的方差膨胀因子均为 3.545;T_{10} 的自变量因子则为 T_7 和 T_3,其对应的方差膨胀因子均为 3.876。可见时间间隔更久远的气温反而可能是未来气温的重要影响因子。建模数据集的拟合结果表明,式(1)、式(2)和式(3)的计算值与实测值之间的拟合相关系数分别为 0.97、0.93、0.90,拟合平均绝对误差(mean absolute error, MAE)分别为 1.32、2.04、2.42 ℃。

2.2.2　基于 LightGBM 的模型　利用 LightGBM 算法计算的训练集 T_8, T_9, T_{10} 与实测值之间的拟合 R 分别为 0.98、0.95、0.93,MAE 则分别为 1.09、1.68、2.00 ℃。

2.2.3　基于 BP-NN 的模型　利用 BP-NN 算法计算的训练集 T_8, T_9, T_{10} 与实测值之间的拟合 R 分别为 0.97、0.93、0.91,MAE 分别为 1.22、1.94、2.30 ℃。

2.3　三种模型的适用性

从 2.2 节针对三种短期日平均气温预报模型的建模结果可知,LightGBM 的预报效果最优,BP-NN 次之,逐步多元线性回归最差。将上述三种算法应用于 2017 年的测试集,就各自在短期日平均气温预报中的适用性展开系统分析。

首先分别绘制了上述三种算法日平均气温预报值与实测值之间的时间序列图(图 1)。从图 1 可知,针对 T_8 的预报效果而言,逐步多元线性回归、LightGBM、BP-NN 的预报值与实测值之间的 R 分别为 0.97、0.98、0.97,MAE 分别为 1.25、

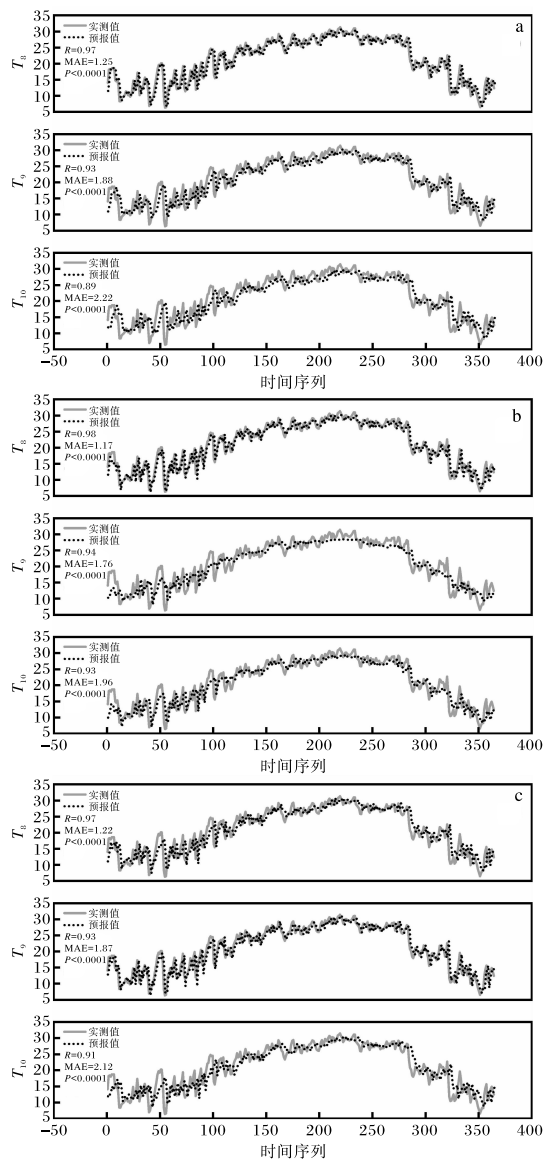

图 1　三种模型日平均气温预报值与实测值的时间序列(a 逐步多元线性回归;b LightGBM;c BP-NN)

1.17、1.22 ℃；T_9 的 R 分别为 0.93、0.94、0.93，MAE 则分别为 1.88、1.76、1.87 ℃；T_{10} 的 R 分别为 0.89、0.93、0.91，MAE 则分别为 2.22、1.96、2.12 ℃。可见三种算法在测试集上的预报效果与训练集表现基本一致，具有优良的泛化性，其中 LightGBM 的预报效果最优，BP-NN 次之，逐步多元线性回归最差。

在实际的气温预报业务中，当气温的预报误差小于 2 ℃时可认为预报正确。为此，分别统计了三种算法日平均气温预报值的准确率及其与实测值之间的 MAE(表 3)。从表 3 可知，就 T_8 的预报结果而言，三者预报准确率可分别高达 83.29%、84.38%、82.73%，MAE 为 1.25、1.17、2.22 ℃；就 T_9 而言，三种模型的差异性明显体现，三者预报准确率分别为 64.38%、69.86%、63.56%，MAE 为 1.88、1.76、1.87 ℃；就 T_{10} 而言，三者预报准确率分别为 56.44%、61.37%、59.18%，MAE 为 2.22、1.96、2.12 ℃。

表 3 三种模型日平均气温的预报结果评价指标

预报结果评价指标	逐步多元线性回归			LightGBM			BP-NN		
	T_8	T_9	T_{10}	T_8	T_9	T_{10}	T_8	T_9	T_{10}
预报准确日数/d	304	235	206	308	255	224	302	232	216
相关系数	0.97	0.93	0.89	0.98	0.94	0.93	0.97	0.93	0.91
平均绝对误差/℃	1.25	1.88	2.22	1.17	1.76	1.96	2.22	1.87	2.12

综上可见，相较于逐步多元线性回归法和 BP-NN 而言，LightGBM 不仅在相关系数以及精确度上更占优势，并且具有更高的预报正确率。尤其随着预报时效的增大，LightGBM 具有更优的预报效果，而 BP-NN 与逐步线性回归法的预报效果则均急剧下降，说明 LightGBM 具有最优的预报稳定性。推测样本数量较少可能是三种模型预报效果存在显著差异的最主要原因。传统的机器学习模型，如 LightGBM、Xgboost、Catboost 等更适用于百万级以下的样本量。

明确模型预报误差的时间分布情况对于提高气温的预报精度具有重要意义。为此，绘制了三种模型日平均气温预报值与实测值之间绝对误差(absolute error，AE)的时间序列图(图 2)。从图 2 可见，相同预报时效，三种模型所表现的 AE 波动形态基本一致。但总体而言，LightGBM 的 AE 及其波动幅度最小，其预报未来 3 d 气温绝对误差的标准差(standard deviation，STD)分别为 1.09、1.53、1.65 ℃；BP-NN 次之，STD 分别为 1.15、1.66、1.97 ℃；逐步线性回归最大，STD 则分别为 1.23、1.73、2.00 ℃。此外，逐步多元线性回归以及 BP-NN 的 AE 及其波动幅度均随着预报时效的增大而显著增大，但 LightGBM 则相对最为稳定。另外从图 2 不难看出，三种模型 AE 的大值区基本位于 0～100 日以及 240～365 日，即处于冬春季以及秋冬季。结合图 1 可推测，该时期气温较大的波动性会加大机器学习模型的训练难度，进而导致 AE 的总体偏大。

综合上述分析可知，LightGBM 和 BP-NN 机器学习方法在预报准确性，拟合效果(R)以及稳定性(AE)方面均要优于逐步多元线性回归法。

3 结论和讨论

(1)通过对过去 1～365 d 的日平均气温进行遍历测试，确定将过去连续 7 d 的日平均气温分别作为逐步多元线性回归、LightGBM 以及 BP-NN 算法的自变量可最准确地预报出未来 1～3 d 的日平均气温，据此构建了短期气温预报模型。该最优自变量的确定方法是以结果为导向，其中所表征的科学背景仍有待进一步探索。

(2)从预报准确率(绝对误差小于 2 ℃的天数占比)，相关系数和绝对误差来看，三种模型均能较准确地预报出未来 1～3 d 的短期日平均气温，其中 LightGBM 最优，BP-NN 次之，逐步多元线性回归最差。以 LightGBM 为代表的传统机器学习模型适用于非图像领域百万级左右的数据集，而对于雷达回波外推以及空间降尺度等图像领域，则要以神经网络为代表的深度学习方法更

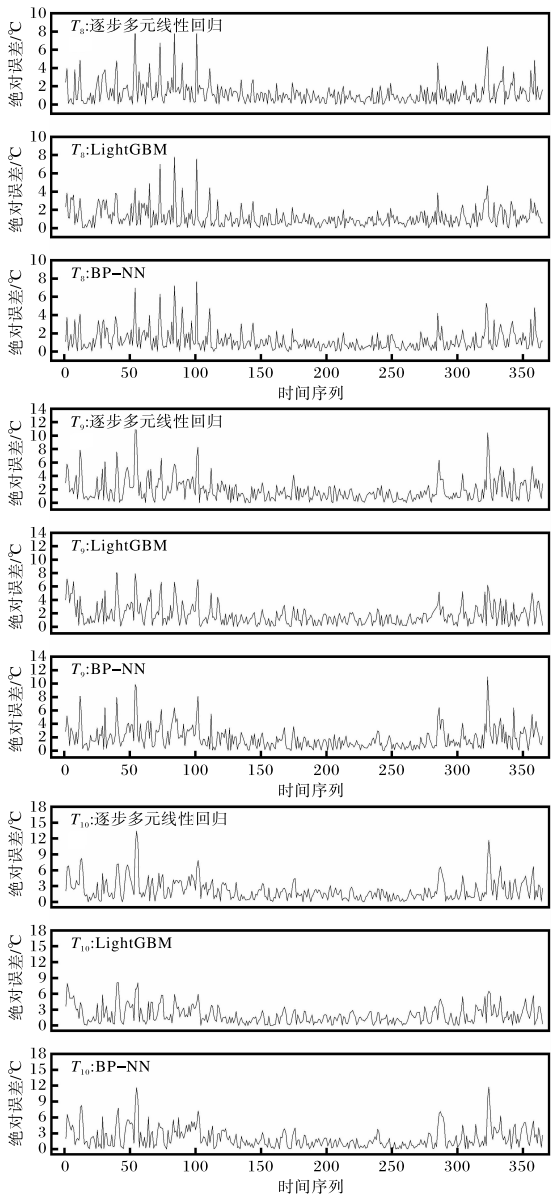

图 2　三种模型针对未来 1～3 d 日平均气温预报的绝对误差(AE)时间序列

为适用。

(3)数据的质量决定了预报效果的上限,而模型只是协助逼近该上限。因此,增加更多的观测与模式预报资料,通过采用特征工程等方法,将有助于进一步提升算法的预报性能。

参考文献:

[1] 陈泮勤. 全球增暖对自然灾害的可能影响[J]. 自然灾害学报,1996,5(2):95-101.

[2] 张加春,饶灶鑫,陈丽珍,等. 1884～2006 年西北太平洋热带气旋活动特征[J]. 广东气象,2008(2):24-26.

[3] HOUGHTON J T. The science of climate change [R]. Climate change 1995, IPCC. Cambridge: Cambridge University Press, 1995.

[4] 丁一汇,戴晓苏. 中国近百年来的温度变化[J]. 气象,1994,20(12):19-26.

[5] 王绍武,叶瑾琳,龚道溢,等. 近百年中国年气温序列的建立[J]. 应用气象学报,1998,9(4):9-18.

[6] 陈隆勋,邵永宁,张清芬,等. 近四十年我国气候变化的初步分析[J]. 应用气象学报,1991,2(2):164-174.

[7] 李文娟,赵放,郦敏杰,等. 基于数值预报和随机森林算法的强对流天气分类预报技术[J]. 气象,2018,44(12):1555-1564.

[8] 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.

[9] 覃卫坚,廖雪萍,陈思蓉. 延伸期暴雨过程的神经网络预报技术应用初探[J]. 气象研究与应用,2018,39(4):1-4.

[10] 俞晓景,董晓敏. 多元回归和逐步回归法的比较试验[J]. 气象,1984,12(12):8-11.

[11] TU J, XIA Z G, WANG H, et al. Temporal variations in surface ozone and its precursors and meteorological effects at an urban site in China[J]. Atmospheric Research, 2007, 85(3-4):310-337.

[12] 王志宇. 基于 LightGBM 框架的上海市大气能见度预报订正研究[D]. 上海:华东师范大学,2019.

[13] 金龙. 神经网络气象预报建模理论方法与应用[M]. 北京:气象出版社,2004.

[14] HE K M, ZHANG X Y, REN S Q, et al. Delving deep into rectifiers: surpassing human-level performance on imagenet classification[J]. arXiv: 1502.01852, 2015.

[15] 蒋昂波,王维维. ReLU 激活函数优化研究[J]. 传感器与微系统,2018,37(2):50-52.

[16] 吕国豪,罗四维,黄雅平,等. 基于卷积神经网络的正则化方法[J]. 计算机研究与发展,2014,51(9):1891-1900.