文章编号：2095-2163(2024)03-0187-05 中图分类号：TP391 文献标志码：A

基于机器学习和时间因子的气温变化模型研究

汪礼原[1]，李 全[1]，龚莹洁[2]，王 颖[2]
（1 湖北师范大学 计算机与信息工程学院，湖北 黄石 435002；2 湖北师范大学 数学与统计学院，湖北 黄石 435002）

摘 要：随着全球气温变暖的加剧，解决全球变暖问题愈加紧迫，本文综合分析了全球气温变化趋势，并针对传统的气温预测模型的局限性进行了研究，在此基础上建立了一种基于机器学习和时间因子的温度预测模型。该模型从时间层面研究气温变化，对时序性温度数据进行数据分析，根据数据变化趋势得出时间和温度数据的变化呈现周期性，并计算不同时间因子对温度模型的影响程度，通过影响程度对时间数据进行细腻化。将不同时间细腻度下的时间数据分别导入 XGBoost 等机器学习模型中进行训练和预测，并对结果进行对比分析。实验结果表明，时间细腻度等于 5 时均方根误差最小，XGBoost 模型的均方根误差在 0.26 左右，精度达到 99.18%。

关键词：时间因子；机器学习；气温变化；数据分析；XGBoost

Research on temperature change based on machine learning and time factor

WANG Liyuan[1], LI Quan[1], GONG Yingjie[2], WANG Ying[2]
(1 School of Computer and Information Engineering, Hubei Normal University, Huangshi 435002, Hubei, China;
2 College of Mathematics and Statistics, Hubei Normal University, Huangshi 435002, Hubei, China)

Abstract: The need to address global warming has increased in importance as the pace of global temperature change has accelerated. As a result, this research conducts a thorough analysis of the global temperature change trends and discusses the shortcomings of conventional temperature forecast models. To address the problem, the temperature prediction model based on machine learning and time factor is built in this study. The model analyzes temperature change from a temporal perspective, conducts data analysis of time-series temperature. According to the trend of data change, the changes in time and temperature data are periodic, and the degree of influence of different time is calculated on the temperature model, further the level of influence is used to refine the temporal data. The temporal data with various degrees of refinement are imported into XGBoost and other machine learning models for training and prediction, respectively. The experimental results shows that when various degrees of refinement is equal to 5, the root mean square error is minimal. The root mean square error of the XGBoost model is about 0.26, and the accuracy reaches 99.18%.

Key words: time factor; machine learning; temperature change; data analysis; XGBoost

0 引 言

21 世纪以来，全球气候变暖问题日益突出。全球气候变暖与人类活动密切相关，人口数量的激增、大量燃烧化石燃料、排放工业废气废水、汽车尾气和砍伐树木等都造成了二氧化碳等温室气体含量的增加。全球气候变暖又会导致海平面上升、冰川融化等问题出现，从而影响人类的健康及居住环境[1]。研究全球气温变化的趋势，将对采取有效措施来保障人类与自然界和谐发展具有一定的参考意义。

目前，全球气候变化预测主要包括贝叶斯预测模型[2]、神经网络、时间序列[3]、随机森林[4]，等方法。2006 年，薛宇峰等学者[5]利用人工神经网络预测了全球气温变化趋势，得到了全球气温将持续偏高的结论。2021 年，侯惠清[6]建立基于 BP 神经网络的全球气候变化预测模型，预测全球平均气温将呈现缓慢上升的趋势。在 2022 年，寇露彦等学者[7]建立了多个因素的向量自回归模型，预测未来 25 年加拿大的平均气温和平均降水量。赵成兵等学者[8]、Strobach 等学者[9]和白雪[10]通过季节性时间建立了预测模型。这些研究都在气温预测上取得了不错的成果。但是部分没有考虑到气候变化与时间因子的密切联系。为解决部分气温变化模型与时间因子相关性不高而导致的预测不精确的问题，本文建立了基于机器学习

作者简介：汪礼原(2002-)，男，本科生，主要研究方向：机器学习、CV、数据挖掘；龚莹洁(2003-)，女，本科生，主要研究方向：数学与应用数学；王 颖(2001-)，女，本科生，主要研究方向：数学与应用数学。
通讯作者：李 全(1982-)，男，副教授，主要研究方向：机器学习、数据挖掘。Email：56322268@qq.com
收稿日期：2023-04-14

和时间因子的温度预测模型,在时间层面上研究温度的变化趋势,并结合 Ke 等学者[11]提出的 XGBoost 等一系列机器学习模型对数据集进行了训练和预测,得到了较高精度的优化结果。

本研究使用的数据来源为 Berkeley Earth,数据涵盖了地表温度记录中最权威的 3 家机构所观测的数据,分别是:英国的 HadCrut、美国的 NASA 和 NOAA。数据包括 1833~2012 年的 3 239 个月中,48 个国家和 100 个城市的月平均气温和修正气温,以及各城市的坐标。

本文研究观测到过去 30 年来全球气温上升的趋势,全球变暖已无法避免。由于全球各国的气候变化数据比较丰富,因此本文以中国从 1838 年到 2012 年的月平均温度为例,构建气温预测模型。

1 基于时间因子的气温预测模型

1.1 XGBoost 模型

XGBoost 模型的主要思想是使用多个弱分类器、即决策树逐步提高模型的拟合能力。在时间序列预测中,可以使用多个决策树学习过去一段时间的数据,预测未来的数值。每一轮迭代中,XGBoost 模型计算目标函数的梯度和二阶导数,并使用这些信息更新决策树叶子节点的权重。通过迭代,XGBoost 模型可以不断提高预测准确性。

1.2 模型评价标准

1.2.1 均方根误差

均方根误差(*RMSE*)[12]计算预测值与实际值之间平均差的平方根,是衡量预测结果与实际结果差异的指标之一。在回归问题中,*RMSE* 常用于评价预测模型预测准确性,定义公式为:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

其中,n 表示样本数量;y_i 表示实际值;$\hat{y}_i$ 表示预测值。

RMSE 的值越小,预测值越接近真实值,模型的预测性能越好。

1.2.2 K 折交叉验证

由 Geisser 等学者提出 K 折交叉验证(K-Fold Cross Validation)[10,13]是机器学习领域中广泛使用的一种交叉验证方法,将数据集分成 k 个子集,并进行 k 次模型训练和测试;每次训练时,使用 $k-1$ 个子集进行训练,用未使用的子集进行模型测试;最终,将 k 次测试结果求平均值,得到模型性能的评估指标。使用 K 折交叉验证方法可以有效避免数据集分割不合理引起的模型性能不稳定问题,提高模型泛化性能,对此可表示为:

$$Loss = \frac{1}{k}\sum_{i=1}^{k} loss_i \tag{2}$$

其中,k 表示将数据集划分为 k 个子集;每个子集都可以用作测试集;其余 $k-1$ 个子集是训练集;$loss_i$ 表示在第 i 个子集上的测试损失。

本文结合 *RMSE* 和 K 折交叉验证作为模型评价标准,将 *RMSE* 的值作为 $loss_i$,综合分析得到最后的评价值。

1.3 模型分析

根据 1838~2013 年的温度数据,观察到数据在这 175 年间呈现出一定的周期性变化,随着时间的推移,整体振幅也在逐渐上升,各年度气温值的变化呈现出相似的波动规律,如图 1 所示。

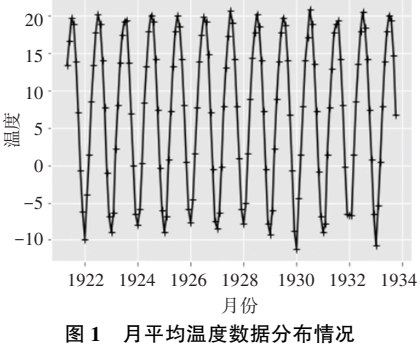

图 1 月平均温度数据分布情况

Fig. 1 Distribution of monthly average temperature data

结合数据特征,分析得出时间因子和温度有明显的周期性,基于此本文提出基于 XGBoost 和时间因子的温度预测模型。算法流程如图 2 所示。

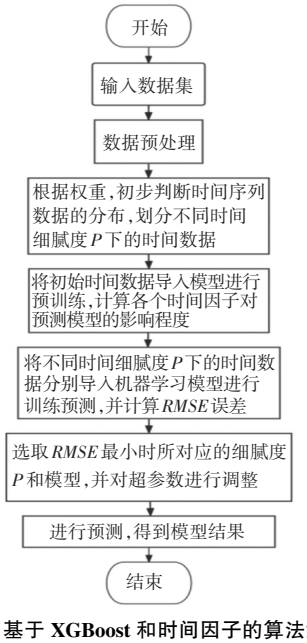

图 2 基于 XGBoost 和时间因子的算法流程图

Fig. 2 Algorithm flow chart based on XGBoost and time factor

在时间序列预测中，通常需要将时间序列数据转换成适合 XGBoost 模型的格式，即将时间序列数据按照时间顺序划分为多个时间段，每个时间段包含过去一段时间的数据以及对应的目标变量值、即待预测的值，这样就可以将时间序列数据转换成监督学习问题的数据集，然后使用 XGBoost 模型来拟合数据集，并预测未来的数值。本文将对时间因子进行细腻度划分，通过求解各个时间因子对模型的影响程度和权重，综合分析出最佳细腻度组合，从而提高模型精度。

本文使用 1838～2012 年的年平均温度，并对时间数据集进行初步细腻化，同时将数据集划分为训练集和测试集，划分比例为 7∶3。

将原始数据集进行拆分，提取各年份时间里的年、月、日、周等一系列时间因子，通过各个时间因子去综合分析，将时间数据进一步细腻化，时间序列训练数据集部分数据见表 1。

表 1　时间序列训练数据集

Table 1　Time series training data set

时间	星期几	季度	月份	年份	一年中的日期	月份日	一年中的星期
18380101	0	1	1	1838	1	1	1
18380201	3	1	2	1838	32	1	5
18380301	3	1	3	1838	60	1	9
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
20011201	5	4	12	2001	335	1	48
20020101	1	1	1	2002	1	1	1

2　模型训练与分析

2.1　细腻度划分

使用时间序列数据的训练集来训练机器学习模型，学习率设置为 0.3，并生成温度预测模型。同时，使用 XGBoost 工具包计算各个时间单位对温度的影响程度和权重。XGBoost 根据结构分数的增益情况计算出选择哪个时间因子作为分割点，而一个时间因子的重要程度则等于其在模型的所有决策树中出现的次数之和，见表 2。

表 2　各时间因子对温度的影响程度

Table 2　Degree of influence of each time factor on temperature

特征	特征影响程度
年份	1 530.0
星期几	678.0
月份	441.0
一年中的星期	218.0
一年中的日期	197.0
季度	16.0
月份日	2

由表 2 可知，在时间序列数据集中，年份的影响因素最大，星期几的影响因子次之，季度和月份日的影响程度非常小。为了筛选出时间因子最佳细腻化组合，本文提取了影响程度排名前 5 的数据，时间细腻度 P 分别为 7、6、5、4、3，根据权重将时间数据进行细腻度划分，并再次进行训练。

2.2　模型训练

本文使用了 Lasso[14]、ElasticNet[15]、Ridge[16]、Gradient[17]、LightGBM[11] 和 XGBoost 模型进行训练预测，并对不同模型得到的结果进行对比分析。在 10 折交叉验证方式下，分别计算不同 P 值下的 *RMSE* 值，结果见表 3。

表 3　不同细腻度 *P* 下的 *RMSE*

Table 3　*RMSE* at different refinement *P*

模型	*RMSE* (P = 7)	*RMSE* (P = 6)	*RMSE* (P = 5)	*RMSE* (P = 4)	*RMSE* (P = 3)
Lasso	9.442	9.307	9.360	9.301	9.774
ElasticNet	9.443	9.308	9.361	9.298	9.774
Ridge	8.362	8.341	8.401	8.343	9.774
Gradient	0.601	0.594	0.582	0.591	0.656
LightGBM	0.678	0.670	0.673	0.674	0.721
XGBoost	0.560	0.506	0.490	0.512	0.546

由表 3 可见，基于时间因子的预测模型随着细腻度 P 的变化，精度也随之变化，说明时间因子细腻化对于温度预测有明显的影响。当 P = 5 和使用 XGBoost 模型的情况下，*RMSE* 最小，精度最高，其中时间因子分别是年、星期几、月、一年中的星期、一年中的日期。因此本文选用了细腻度 P = 5 的时间因子，并采用 XGBoost 模型进行模型搭建。为了进一步验证 P = 5 的时间因子的可靠性，又计算了 P = 5 下的各时间因子影响程度，见表 4。由表 4 可以明显看出各个时间因子的影响程度对模型相对平衡，

年对模型的影响力依旧最大,其他因子不存在影响程度过小的情况。

表 4 筛选后时间因子对模型的影响程度

Table 4 The degree of influence of the time factor on the model after screening

特征	特征影响程度
年份	25 185
星期几	12 465
月份	10 755
一年中的星期	8 998
一年中的日期	7 269

2.3 参数优化

为了对模型超参数进行调整和分析,计算出最优超参数配比,本文使用了网格搜索和随机搜索两种方法对超参数进行寻优。

(1)网格搜索优化(GridSearchCV)。网格搜索[9]采用穷举搜索算法,在超参数空间中定义一组网格点,遍历所有可能的组合以寻找最优超参数组合。网格搜索易于理解并实现,但在高维参数空间中的计算成本很高,适用于低维参数空间,或用于筛选超参数空间的候选。

(2)随机搜索优化(RandomizedSearchCV)。随机搜索[19]是一种随机采样搜索算法,通过在超参数空间中随机采样一组超参数组合,并通过迭代不断更新采样分布,以寻找最优超参数组合。与网格搜索相比,随机搜索的计算成本较低,能够有效处理高维参数空间,通常能够找到比网格搜索更好的超参数组合。但是为了达到理想的性能,随机搜索需要更多的采样次数,同时也不能很好地处理离散参数空间。

本文研究中,使用 Sklearn 的 *GridSearchCV* 和 *RandomizedSearchCV* 函数分别进行参数寻优,网格搜索和随机搜索最佳参数配比见表 5。

表 5 网格搜索和随机搜索最佳参数配比表

Table 5 Optimal parameter ratios for *GridSearchCV* and *RandomizedSearchCV*

超参数	网格搜索参数选择	网格搜索最佳参数	随机搜索参数选择	随机搜索最佳参数
n_estimators(树的节点)	20,50,200,300,3 000	50	20/40/…/480/500	280
learning_rate(学习率)	0.1,0.01,0.05	0.1	0.01/0.02/…/1.99/2	0.42
max_depth(树的最大深度)	5,6,7,8	5	2/3/…/14/15	6
Min_child_weight(最小权重)	1,2,3,4	1	1/2/…/8/9	6
Subsample(样本采样比例)	0.6,0.7,0.8,0.9	0.6	0.6/0.61/…/0.8	0.61
colsample_bytree(特征采样比例)	0.5,0.6,0.7,0.8,0.9	0.6	0.5/0.55/……/0.95	0.82

根据表 5,网格搜索在树的节点为 50 时得到了最佳结果,此时模型的 *RMSE* 为 0.42,准确率为 98.22%。而随机搜索则在树的节点为 280,学习率为 0.42 时找到了最佳参数组合,此时 *RMSE* 为 0.26,准确率为 99.18%,模型精度进一步提高。

因此,本文最终选择了时间细腻度 $P = 5$ 的情况下,用 XGBoost 模型对预测集进行预测,并对模型预测结果和真实值的分布情况进行了可视化展示,2002 年以后为预测值部分,如图 3 所示。

通过图像可以看到预测值和真实值的曲线趋势基本拟合,再放大测试集的部分数据,如图 4 所示。

图 4 中,圆点代表预测值,曲线代表实际值。可以看出测试集的预测值基本上都在曲线上,拟合程度很高,模型鲁棒性非常好。

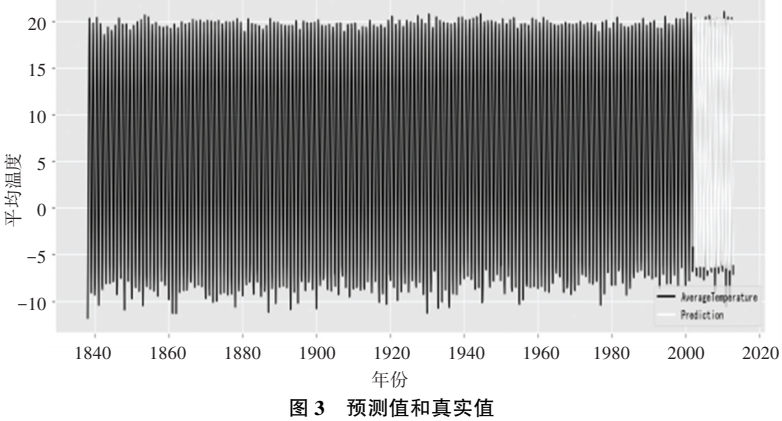

图 3 预测值和真实值

Fig. 3 Predicted values and true values

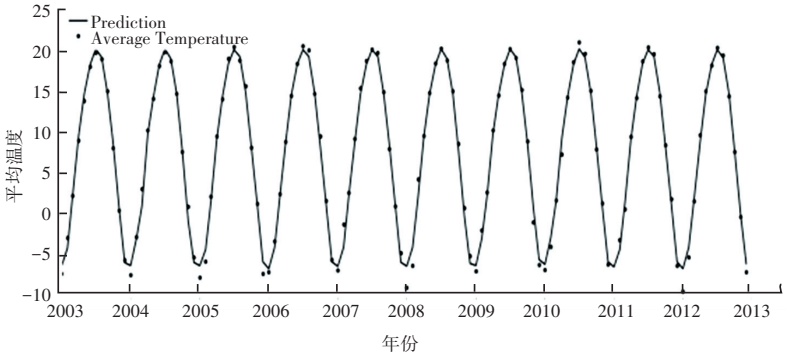

图 4　测试集的预测值和真实值

Fig. 4　Predicted values and true values of the test set

综上，基于 XGBoost 和时间因子的温度预测模型的 *RMSE* 值更小。时间因子与温度的相关性较强，相比其他模型，具有更强的鲁棒性，能更好地预测温度的变化和趋势。

3　结束语

在全球气候变暖问题日益严峻的背景下，本文提出了一种基于时间因子和机器学习的温度预测模型，通过综合分析时间数据，判断时间数据分布，计算各个时间因子对温度模型的影响程度和权重，再通过划分时间细腻度 P 来选择最优结果；采用机器学习模型对数据进行拟合预测，再和其他模型的结果进行对比分析。实验结果表明，在最优超参数组合下，基于机器学习和时间因子的气温变化模型在一定程度上提高了预测精度，且 *RMSE* 只有 0.26 左右，精确度能够达到 99.18%，具有更强的鲁棒性。在未来会通过增加更多的时间因子进一步综合分析。

本文的研究范围仅限于数据，模型在实际应用中的泛化性仍需提高。在后续研究当中将寻找更充足的数据，引入更多的时间因子并结合最新模型去综合分析，并结合最新模型展开后续研究。

参考文献

[1] 热伊莱·卡得尔，伊卜拉伊木·阿卜杜吾普，陈刚. 全球气候变化及其影响因素研究进展[J]. 农业开发与装备，2020(9)：81-82.

[2] BERNARDO J M, SMITH A F M. Bayesian theory[M]. Hoboken, USA：John Wiley & Sons, 2009.

[3] 张迎春，肖冬荣，赵远东. 基于时间序列神经网络的气象预测研究[J]. 武汉理工大学学报(交通科学与工程版)，2003，27(2)：237-240.

[4] SEGAL M R. Machine learning benchmarks and random forest regression[EB/OL]. [2004-04-14]. http://scholarship.org/uc/item/35x3v9t4.

[5] 薛宇峰，杨超梅. 近百年全球气温变化及其趋势预测[J]. 四川气象，2006(3)：16-19.

[6] 侯惠清. 基于 BP 神经网络的全球气候变化预测模型[J]. 科技与创新，2021(9)：10-11.

[7] 寇露彦，廖竞，李学俊，等. 基于 VAR 模型的加拿大气候变化预测[J]. 计算机与现代化，2022(10)：13-18.

[8] 赵成兵，刘丹秀，谢新平，等. 基于时间序列的季节性气温预测研究[J]. 安徽建筑大学学报，2022，30(3)：83-89.

[9] STROBACH E, BEL G. Decadal climate predictions using sequential learning algorithms[J]. Journal of Climate, 2016, 29(10): 3787-3809.

[10] 白雪. 基于时间序列模型和 XGBoost 组合的气温预测与高温热浪预警分析[D]. 南京：南京信息工程大学，2021.

[11] KE Guolin, MENG Qi, FINLEY T, et al. LightGBM: A highly efficient gradient boosting decision tree[C]//31st Conference on Neural Information Processing Systems(NIPS 2017). Long Beach, USA：NIPS Foundation, 2017: 3146-3154.

[12] CHAI T, DRAXLER R R. Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature[J]. Geoscientific Model Development, 2014, 7(3): 1247-1250.

[13] RODRIGUEZ J D, PEREZ A, LOZANO J A. Sensitivity analysis of k-fold cross validation in prediction error estimation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 32(3): 569-575.

[14] ZOU Hui. The adaptive lasso and its oracle properties[J]. Journal of the American Statistical Association, 2006, 101(476): 1418-1429.

[15] ZOU Hui, HASTIE T. Regularization and variable selection via the elastic net[J]. Journal of the Royal Statistical Society Series B: Statistical Methodology, 2005, 67(2): 301-320.

[16] HOERL A E, KENNARD R W. Ridge regression: Biased estimation for nonorthogonal problems[J]. Technometrics, 1970, 12(1): 55-67.

[17] FRIEDMAN J H. Stochastic gradient boosting[J]. Computational Statistics & Data Analysis, 2002, 38(4): 367-378.

[18] SYARIF I, PRUGEL-BENNETT A, WILLS G. SVM parameter optimization using grid search and genetic algorithm to improve classification performance[J]. TELKOMNIKA (Telecommunication Computing Electronics and Control), 2016, 14(4): 1502-1509.

[19] BERGSTRA J, BENGIO Y. Random search for hyper-parameter optimization[J]. Journal of Machine Learning Research, 2012, 13(10): 281-305.

[20] CHEN Tianqi, GUESTRIN C. XGBoost: A scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA：ACM, 2016：785-794.