

# 一种机器学习方法在湖北定时气温预报中的应用试验

谭江红<sup>1, 2</sup> 陈伟亮<sup>1, 2</sup> 王珊珊<sup>3</sup>

(1 荆州市气象局, 荆州 434020; 2 江汉平原生态气象遥感监测技术协同创新中心, 荆州 434025;  
3 武汉中心气象台, 武汉 430074)

**摘要:** 利用2015—2017年湖北89个气象站地面观测温度、欧洲中心再分析资料和0~12 h预报资料回归模式输出要素与地面气温之间的关系, 建立了LightGBM模型, 并在2018年数据集上进行测试。结果表明, 定时气温平均绝对误差由模式本身的1.8 °C下降到1.1 °C, 2 °C以内预报准确率由65.9%上升至86.6%, 决定系数(拟合优度)高达0.97。该模型已经在武汉中心气象台业务化, 初步选取定时气温中的极值进行2018年2—6月预报评分, 24 h高、低温预报准确率分别为76.9%和91.4%, 在客观产品中排名前列, 较数值预报模式产品提升明显, 低温预报准确率超过预报员水平。LightGBM作为一个年轻的机器学习框架, 在气象要素预报方面具备良好的应用前景。

**关键词:** 机器学习, 决策树, LightGBM, 气温, 预报试验

**DOI:** 10.3969/j.issn.2095-1973.2018.05.006

## Using a Machine Learning Method for Temperature Forecast in Hubei Province

Tan Jianghong<sup>1, 2</sup>, Chen Weiliang<sup>1, 2</sup>, Wang Shanshan<sup>3</sup>

(1 Jingzhou Meteorological Bureau, Jingzhou 434020 2 Collaborative Innovation Center of Remote Sensing Technology in Ecological and Meteorological Monitoring in the Jiangnan Plain, Jingzhou 434025 3 Wuhan Central Meteorological Observatory, Wuhan 430074)

**Abstract:** A prediction model using LightGBM was established according to a regression analysis using ground observations of temperature, ERA-Interim data and 0-12 h forecast data. Tests on the data set from January to April in 2018 showed that the mean absolute error of the temperature decreased from 1.8 °C to 1.1 °C with the EC model, the prediction accuracy increased from 65.9% to 86.6% and the coefficient of determination was 0.97. This model has been operationalized by the Wuhan Central Meteorological Observatory. The forecast score from February to June in 2018 showed that the accuracy of the highest and lowest temperature forecasts were 76.1% and 91.5%, respectively, which was significantly higher than that of the numerical weather prediction, and the minimum temperatures were better than current forecasters. As a young machine learning frame, LightGBM has strong prospects for application in meteorological forecasting.

**Keywords:** machine learning, decision tree, LightGBM, temperature, forecast experiment

### 0 引言

机器学习是人工智能的重要解决方案, 包括监督学习和非监督学习<sup>[1]</sup>。天气预报中的很多问题都可以转化为标签已知的监督学习问题: 标签即待预报量, 机器学习中的特征即预报因子。机器学习模型具有一定黑盒性, 但其使用效果经常出乎意料。

经验表明, 代表全球数值模式最高水平的欧洲中期天气预报中心(以下简称“欧洲中心”)在形势预报方面日益进步, 达到了非常高的可用水平, 预报员在订正形势预报方面操作空间很小, 但数值预报的要素误差仍有较大订正空间。随着人工智能浪潮的来临, 客观化智能化预报是未来天气预报的发展方向。2017年湖北省气象局开始开展智能网格预报业务。在湖北地区, 黄治勇等<sup>[2]</sup>使用了带海拔高度的距离权重温度插值方法和灰色预测模型来预报湖北温度。另外一些见诸文献的方法还包括卡尔曼滤波订正模式温度预报<sup>[3]</sup>、神经网络方法中的BP网络<sup>[4]</sup>、Barnes插值<sup>[5]</sup>、时效偏差消除<sup>[6]</sup>、统计降尺度<sup>[7]</sup>等。

收稿日期: 2018年5月10日; 修回日期: 2018年8月5日  
第一作者: 谭江红(1987—), Email: 343017729@qq.com  
通信作者: 陈伟亮(1991—), Email: ouchenweiliang@163.com  
资助信息: 中国气象局预报员专项基金(CMAYBY2016-045);  
湖北省气象局科技发展基金重点项目(2016Z02)

在温度预报中使用的线性方法主要包括多元线性回归、逐步线性回归、最优子集回归，但由于天气演变的非线性特征、预报量与预报因子关系的非线性特征，线性方法的局限性较大。除了神经网络外，非线性方法还包括支持向量机、相似预报理论等<sup>[8]</sup>。预报员的思维一般是基于历史经验，本质上也是一种相似预报的主观方法，只不过是基于抽象经验，难以通过生物神经系统准确记住海量历史个例，故难以客观量化，而通过机器学习方法进行数据挖掘正是从历史数据中挖掘规律，本质是一种基于历史相似样本的客观定量方法。数据挖掘技术也早已在精细化温度预报中得到探索<sup>[9]</sup>，前人研究多为BP神经网络、时间序列数据挖掘方法，近年来改进并兴起的树类机器学习方法暂未得到广泛使用。本文探索和引入了数据挖掘中比较年轻的树类集成数据挖掘方法在历史数据集上进行预报应用，并在预报实践中检验其效果。

## 1 资料与方法

本文使用的资料为2015—2017年湖北89个气象站地面观测温度，时间为每天逐3 h正点观测时间（北京时间20, 23, 02, 05, 08, 11, 14, 17, 20时），数据来源为自动站数据库，欧洲中心（EC）再分析资料、0~12 h预报场（每天08、20时2次）。最优训练期方案<sup>[10-11]</sup>证明，样本时间尺度的选择对于温度预报的影响较大。选取的时间尺度足够长，才能优于传统的季节固定期分类，才能更好地衡量不同天气条件下的各类不同情形。通俗地说，机器学习需要“大数据”支持，足够多的样本才能使模型“见多识广”，例如，建模数据不包括台风过程，在预报时段出现台风，模型很可能无法识别罕见低气压的意义，无从知道台风低压会造成何种结果。

LightGBM是微软于2016年开源的一种以决策树作为基学习器的梯度提升（boosting）机器学习框架，较以往的树类集成学习方法（如随机森林、XGBoost<sup>[12]</sup>）有明显优势，LightGBM算法相关文章<sup>[13]</sup>有详细介绍。与之前的梯度提升类决策树相比，LightGBM具有更快的训练效率、低内存使用、更高的准确率、支持并行化学习等很多优点，其基学习器决策树<sup>[14]</sup>的思想本质是一系列if-then条件判断的嵌套集合，与预报员的思路不谋而合：预报员正是基于各种因子（比如灾害性天气的形成的水汽、动力、热力条件）对结果进行分析，从而判断各条件的组合是否会产生特定的天气现象。

由于决策树具有高度非线性的特点，所以可以解决非线性很强的天气预报问题。单一的决策树往往效

果一般，但包括随机森林、XGBoost在内的树类集成机器学习方法在结构化数据挖掘方面有非常广泛的应用，特别是近年来，XGBoost、LightGBM在数据挖掘竞赛中应用非常广泛，被誉为冠军选手的“杀器”，同时这些方法在工业预测领域也有了大量相关应用<sup>[15-25]</sup>，应该借鉴到气象预报中。而在预报领域，集合预报与集成预报的思路同样广为应用，已经有大量温度预报客观方法采用了集成MOS方法<sup>[3, 26-28]</sup>，随机森林、XGBoost、LightGBM等集成方法正是构建“多棵树”作为基预测器实现集成效果，相当于拥有多个预报成员。

大多数机器学习工具都无法直接支持类别特征作为输入，一般需要进行one-hot码，转换成多维特征。LightGBM增加的针对类别特征的决策规则在处理温度预报时很有用，因为预报实践证明地形气候、观测环境在气象要素预报中是不可忽略的，正是地形因素使模式温度预报在山区经常与站点观测存在较大差异（模式很好地处理了物理过程，而对地形的处理仍有缺陷），因而需要分站点建模以区分局地因子的差异，这里以站号来区分不同站点，作为预报因子输入，使模型自动学习不同站点，相当于考虑了不同站点的预报要素与因子之间的统计关系差异。由于该方法是最新开源的树类机器学习框架，暂未见集成决策树类相关模型在温度预报中应用的中文文献（黎光智<sup>[29]</sup>在“透过网页内容预测新闻热门程度”研究中使用了LightGBM方法），因此在天气预报领域的应用具有一定示范性。

之所以选取模式0~12 h预报进行建模是由于以下两个原因：1）模式时效越近，模式的预报效果越好，更容易反映大气的真实情况。这本质上是一种更接近PP法的MOS预报方法，回归出的预报关系比较可靠，这样可以尽量避免模式重大的性能调整带来的回归关系误差。2）使用模式要素进行MOS建模，相比PP法，其预报因子更加丰富：模式同化了包括卫星资料在内的大量气象观测资料，弥补了建模的时空分辨率，也有大量二次计算加工的物理量，同时可以更方便地直接利用模式输出资料进行任意时效的预报应用，由于相同的数据格式，无需额外的数据处理，对于一线业务预报非常实用。

## 2 特征工程

特征工程相当于描述了机器学习模型的内在构成因子，所以特征工程与数据质量决定了机器学习模型效果的上限，此步骤尤为重要，需要确保不丢失重要特征的基础上避免冗余特征、无关特征和维

数灾害<sup>[1]</sup>，即选取与待预测量相关性最强的影响因子是必要的，否则无法学习到数据之间的内在客观规律，而选取的因子过多也可能会引起机器的无效学习或陷入过拟合，增加计算开销和学习难度，需要一定理论知识和业务经验。根据天气学原理<sup>[30]</sup>，局地温度的变化主要取决于温度平流（大气内部的热量交换）和非绝热因子（大气与外部的热量交换）。本项目选取的特征如下：风场包括10 m、925 hPa、850 hPa、700 hPa，一定程度上体现了影响某地的天气系统和冷暖空气活动状态（例如吹南风和吹北风的热力性质是不一样的）。实践经验表明，相对湿度越大，水汽凝结倾向越高，相对湿度是否超过80%与该层的云量状况直接相关，例如工作经验表明，模式预报地面相对湿度超过95%时，大概率有雾出现，故选取1000~100 hPa（1000，925，850，700，600，500，400，300，200，100 hPa）各层相对湿度，用于体现当地上空的天空状况以衡量辐射因子。温度方面，选取了1000~850 hPa温度，以体现大气的基本冷暖状态，不仅考虑了天气尺度的时空差异，也相当于考虑了季节因素，而已经有研究表明，聚类分析方法进行季节划分可改善极端温度预报效果<sup>[11]</sup>。这些层次接近地面，与地面温度相关性高。此外还选取了海平面气压以及变压，用于衡量冷空气的强度及其活动。选取模式本身输出的2 m温度作为基准量，

选取站号和时间作为分类变量。矢量场分解为 $u$ ， $v$ 分量，共计26个因子。数值预报格点的经纬距分辨率为 $0.125^\circ \sim 0.25^\circ$ ，取值方式采用站点最邻近格点方式。

该特征工程选取温、压、湿、风等基本气象要素来衡量大气的基本状态（其他派生量实际上也均可通过基本量计算得到），从物理上体现出了影响温度的主要因子，其基本思想是，建立统计机器学习回归模型，用于学习大气不同状态下模式本身的2 m温度预报误差，其本质是基于历史相似数据分布或排列组合进行相似订正从而做出最终预报。通过数据处理，形成了预报因子和实况数据对应的建模样本矩阵数据库。

### 3 模型建立与评估

本文利用Python进行机器学习建模，为避免过拟合采用5折CV交叉验证，给定参数列表字典通过遗传进化算法进行参数搜索完成机器学习超参数寻优，最终形成了LightGBMRegressor模型，并保存模型用于预测。相关使用方法和原理可参考LightGBM<sup>[12]</sup>、scikit-learn<sup>[31]</sup>、tpot<sup>[32]</sup>等项目文档，最终使用的模型参数见表1。使用scikit-learn随机选取模型1%的样本数进行拟合效果评估，模型的误差见表2，可以看出模式的均方根误差RMSE和平均绝对误差MAE较小，而衡量回归效果的决定系数很高，正负误差样本数基本相同，表明了模型预报能力的平衡性。

表1 模型主要参数

Table 1 Main parameters of LightGBM

n_estimators	learning_rate	num_leaves	lambda_l2	boosting_type	min_child_samples	colsample_bytree	subsample
1000	0.05	250	0.001	gbdt	1	0.7	1.0

表2 模型拟合效果评价

Table 2 Goodness of fitting using regression

RMSE	MAE	决定系数	正误差样本数	负误差样本数
0.82	0.63	0.99	3657	3796

图1是模型给出的特征重要性排名，从大到小（图1从右至左）依次为：站点、模式2 m温度、100 hPa相对湿度、海平面气压、925 hPa温度、700 hPa相对湿度、200 hPa相对湿度、850 hPa温

度、时间、700 hPa经向风速 $v$ 、700 hPa纬向风速 $u$ 、600 hPa相对湿度、850 hPa相对湿度、1000 hPa相对湿度、300 hPa相对湿度、500 hPa相对湿度、1000 hPa温度、400 hPa相对湿度、850 hPa  $v$ 风速、850 hPa  $u$ 风速、925 hPa相对湿度、3 h变压、925 hPa  $v$ 风速、925 hPa  $u$ 风速、10 m  $v$ 风速、10 m  $u$ 风速。

可见在温度预报方面，站点的因素排位第一，即需首要考虑观测环境不同导致的模式预报误差的不同，充分证明了特征工程设计的合理性。模式2 m温度排名第二，意味着温度预报最重要的参考资料仍然是模式输出的2 m温度，排名第三位的是100 hPa相对湿度，当100 hPa湿度较大的时候，往往意味着云层的伸展高度较高（较厚），这是影响温度的辐射因子的重要体现。排名第四位的是海平面气压，与冷空气的活动相关性很大。可以看出，机器学习的结果与天气学原理和实际预报经验是一致的。排名三位以下的因

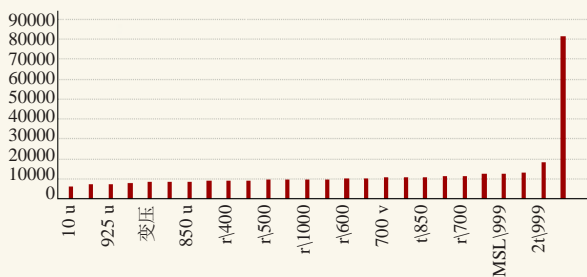


图1 特征重要性排名

Fig. 1 The feature importances

子的特征重要性在数量级上相差不大，但并不一定意味着这些特征并不重要，因为这是以单一因子进行度量，而实际情况是这些因子排列组合起来的结果，如上所述，各层相对湿度就与云高云厚有对应关系，进而影响光照或长波辐射。

#### 4 预报检验与预报实例

利用2018年独立样本进行预报检验，并与现行智能网格预报结果进行对比，其检验效果如图2。可以看出模型的预报和实况变化趋势一致，经计算其决定系数高达0.97。平均绝对误差1.1℃，而相同样本欧

洲模式2 m温度预报误差为1.8℃。2℃以内的预报准确率由欧洲数值预报本身的65.9%提高到86.6%，而同期省台智能网格客观产品12 h以内的定时气温预报准确率约为70.1%。例如，对57453站2018年1月9日08时的预报，欧洲中心数值预报为-12.3℃，而模型预报为1.6℃，实况为0.4℃，预报绝对误差由欧洲中心的12.7℃下降到模型的1.2℃。又如表3，相比欧洲中心模式，2018年3月10日05时，LightGBM模型大部分站点预报误差都在2℃以内（现行业务准确率评价标准），相比数值预报本身模型降低预报误差的趋势明显，仅仅57256站出现了预报错误。

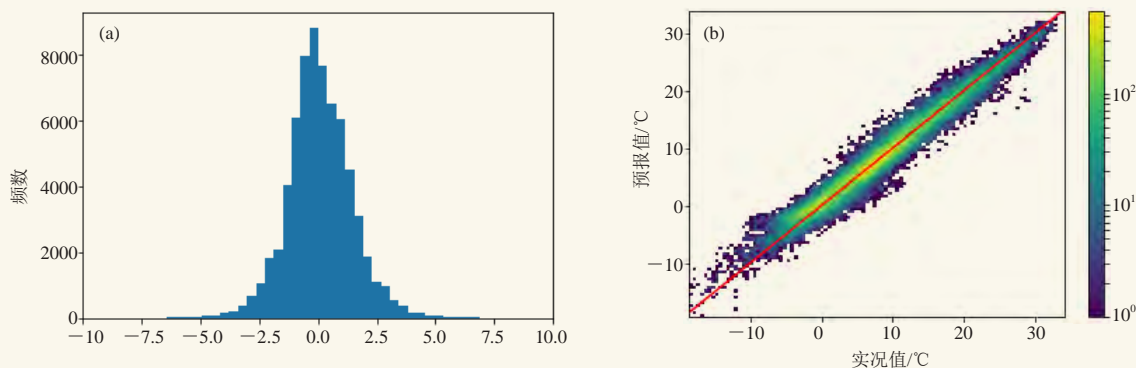


图2 预报与实况对比检验 (a: 误差频数分布直方图; b: 散点及其密度和趋势图)

Fig. 2 Inspection of forecast in contrast to the observed data  
(a: frequency distribution; b: scatterplot and its density and trend line)

表3 2018年3月10日05时模型预报示例(单位:℃)  
Table 3 Examples of model forecast at 5:00AM on March 10, 2018 (°C)

站号	实况	LightGBM	EC	LightGBM误差	EC误差
57249	3.6	5.3	7.6	1.7	4.0
57251	3.5	4.3	5.3	0.8	1.8
57253	4.4	5.2	3.8	0.8	-0.6
57256	4.6	7.1	6.6	2.5	2.0
57257	5.6	5.9	4.5	0.3	-1.1
57259	4.1	3.7	6.5	-0.4	2.4
57260	7.9	5.9	4.6	-2.0	-3.3
57264	7.1	6.1	5.0	-1.0	-2.1
57265	4.3	3.9	4.2	-0.4	-0.1
57268	7.7	6.1	5.4	-1.6	-2.3
57278	6.5	7.4	5.9	0.9	-0.6
57279	7.5	8.0	5.8	0.5	-1.7
57351	6.7	6.8	4.5	0.1	-2.2
57352	6.8	8.6	7.8	1.8	1.0
57353	8.0	8.2	4.5	0.2	-3.5
57354	-1.9	-1.3	-0.9	0.6	1.0
57355	8.2	9.0	6.0	0.8	-2.2
57358	7.3	7.2	4.5	-0.1	-2.8
57359	6.9	7.0	4.4	0.1	-2.5
57361	5.8	5.1	4.1	-0.7	-1.7

本模型现已接入了武汉中心气象台智能网格预报平台进行业务化应用，日常业务一般检验的是最高、最低气温，由于本模型预报的是定时气温，所以初步产品只是简单地从24 h内8个定时气温中挑选出最高、最低温度作为高、低温预报。对2018年2—6月预报产品进行评分（表4），该模型的高、低温预报位居所有客观预报产品前列（其中省台产品是预报员最后主观订正后的产品），特别是24 h低温预报准确率高达91.4%，位居所有预报产品第一位，但分析其平均误

表4 2018年2—6月高、低温预报评分  
Table 4 Maximum and minimum temperature forecast score from February to June, 2018

产品	24 h高温 准确率/%	24 h低温 准确率/%	24 h高温 平均误差/℃	24 h低温 平均误差/℃
EC	56.2	79.1	-1.7	-0.3
SCMOC	66.2	73.9	-0.1	0.2
GRAPES	57.2	65.6	-1.5	0.7
T639	56.7	42.1	-0.5	-2.3
NCEP	63.0	67.8	0.3	0.1
省台产品	77.8	86.1	0.1	-0.3
MOS订正	76.4	83.8	0.4	-0.3
LightGBM	76.9	91.4	-0.7	0.3
集合优选	68.2	82.8	-0.2	0.1

差,发现高温预报有系统性偏低,低温预报有系统性偏高,这是不难理解的:由于建模和预报使用的是定时气温,从定时气温中挑选出的最大、最小值实际上仍不能完全代表日极端值,例如,在晴朗天气下,高温常常比14时气温要高,经常出现在15时左右,可以推测,针对日最高、最低气温直接建模应该可以获得更大提升空间,后期可以继续探索,优化模型。

## 5 结论与讨论

1) 机器学习模型的决策过程有一定的黑盒性,但使用效果好(0~24 h气温预报模型相比数值模式准确率可提高10%以上),可以将预报员的主观预报经验进行高效地客观定量化,具备广阔的应用前景。

2) 由于不同时间辐射状况不同,模式误差不同;不同站点小气候特点(例如海拔)不同,模式误差也不同,需要将相对标准化的模式输出进行系统订正,而LightGBM可以直接处理数值本身没有意义的分类变量,无需离散编码,从而相当于针对不同时段和站点进行建模,效率有很大的进步。

3) 由于LightGBMRegressor叶子数、学习率等超参数寻优需要一定的经验,更好的模型效果需要更多的试验,由于超参数寻优的专业性和计算密集型,具备一定的软硬件要求和难度,需要进一步探索和专业软硬件支持。

4) 由于不同时段的数据格式相同,业务化简便易行,在模式分析场质量不剧烈变化、模式性能、模式不同时段基本形势的预报没有明显改变,即预报稳定时,该模型不同时段的气温预报能力是相同的。

### 参考文献

- [1] 周志华. 机器学习. 北京: 清华大学出版社, 2016.
- [2] 黄治勇, 张文, 陈璇, 等. 湖北省乡镇温度预报方法初探. 气象, 2011, 37(12): 1578-1583.
- [3] 罗聪, 曾沁, 高亭亭, 等. 精细化逐时滚动温度预报方法及检验. 热带气象学报, 2012, 28(4): 552-556.
- [4] 赵声蓉. 多模式温度集成预报. 应用气象学报, 2006(1): 52-58.
- [5] 邱学兴, 王东勇, 朱红芳. 乡镇精细化最高最低气温预报方法研究. 气象与环境学报, 2013, 29(3): 92-96.
- [6] 李佰平, 智协飞. ECMWF模式地面气温预报的四种误差订正方法的比较研究. 气象, 2012, 38(8): 897-902.
- [7] 陈国华. 基于统计降尺度的中国地面气温精细化预报方法研究. 南京: 南京信息工程大学硕士学位论文, 2013.
- [8] 薛志磊, 张书余. 气温预报方法研究及其应用进展综述. 干旱气象, 2012, 30(3): 451-458, 464.
- [9] 段文广, 周晓军, 石永炜. 数据挖掘技术在精细化温度预报中的应用. 干旱气象, 2012, 30(1): 130-135.
- [10] 吴启树, 韩美, 郭弘, 等. MOS温度预报中最优训练期方案. 应用气象学报, 2016, 27(4): 426-434.
- [11] 车钦, 赵声蓉, 范广洲. 华北地区极端温度MOS预报的季节划分. 应用气象学报, 2011, 22(4): 429-436.
- [12] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. arXiv:1603.02754v3.
- [13] Ke G, Meng Q, Finley T, et al. LightGBM: A highly efficient gradient boosting decision tree. //31st Conference on Neural Information Processing Systems(NIPS 2017).
- [14] 张桢, 曹健. 面向大数据分析的决策树算法. 计算机科学, 2016, 43(S1): 374-379, 383.
- [15] 叶倩怡, 饶泓, 姬名书. 基于Xgboost的商业销售预测. 南昌大学学报(理科版), 2017, 41(3): 275-281.
- [16] 王焱. 基于随机梯度提升决策树的行人检测算法设计与实现. 杭州: 浙江大学硕士学位论文, 2017.
- [17] 陈子之. 基于GBDT的地方政府债务风险评级和预警研究. 上海: 上海师范大学硕士学位论文, 2017.
- [18] 郑凯文, 杨超. 基于迭代决策树(GBDT)短期负荷预测研究. 贵州电力技术, 2017, 20(2): 82-84, 90.
- [19] 肖琴. 基于互联网数据的个人信用风险评估的研究与应用. 成都: 电子科技大学硕士学位论文, 2017.
- [20] 凌筱玥. 基于XGBoost算法的上证指数预测方案设计研究. 上海: 上海师范大学硕士学位论文, 2017.
- [21] 霍荣瑞. 柳州市手足口病发病率预测模型的建立与评估. 南宁: 广西医科大学硕士学位论文, 2017.
- [22] 叶倩怡, 饶泓, 姬名书. 基于Xgboost的商业销售预测. 南昌大学学报(理科版), 2017, 41(3): 275-281.
- [23] 乐明明. 数据挖掘分类算法的研究和应用. 成都: 电子科技大学硕士学位论文, 2017.
- [24] 张昊, 纪宏超, 张红宇. XGBoost算法在电子商务商品推荐中的应用. 物联网技术, 2017, 7(2): 102-104.
- [25] 王尉同, 潘毅群, 黄治钟. 梯度渐进回归树在建筑能耗短期预测中的应用. 建筑节能, 2018, 46(3): 112-116.
- [26] 张秀年, 曹杰, 杨素雨, 等. 多模式集成MOS方法在精细化温度预报中的应用. 云南大学学报(自然科学版), 2011, 33(1): 67-71, 76.
- [27] 熊敏谏. 基于集合预报系统的日最高和最低气温预报. 气象学报, 2017, 75(2): 211-222.
- [28] 林春泽, 智协飞, 韩艳, 等. 基于TIGGE资料的地面气温多模式超级集合预报. 应用气象学报, 2009, 20(6): 706-712.
- [29] 黎光智. 透过网页内容预测新闻热门程度. 台北: 台湾大学电机工程学研究所学位论文, 2017.
- [30] 朱乾根, 林锦瑞, 寿绍文, 等. 天气学原理和方法. 北京: 气象出版社, 2007.
- [31] Pedregosa F, Gramfort A, Michel V, et al. Scikit-learn: machine learning in Python. Journal of Machine Learning Research, 2012, 12(10): 2825-2830.
- [32] Olson R, Moore J. TPOT: A tree-based pipeline optimization tool for automating machine learning. Workshop on Automatic Machine Learning. 2016(64): 66-74.