

基于划分工作方式的中央空调达标时间预测

王 可¹, 卢焕达², 郑军红¹, 何利力¹

(1. 浙江理工大学信息学院, 浙江 杭州 310018)

(2. 浙大宁波理工学院计算机与数据工程学院, 浙江 宁波 315100)

[摘要] 为了同时满足中央空调温湿度控制工艺要求和企业节能降耗要求, 解决中央空调达标时间预测问题, 提出了一种在划分空调工作方式基础上的组合预测模型. 在加温加湿工作方式下, 采用随机森林算法构建分类预测模型, 用深度特征提取后的高级特征作为模型输入, 解决了小样本分类预测的过拟合问题. 为进一步降低算法时间复杂度, 利用改进粒子群方法对模型参数寻优. 在降温除湿工作方式下, 使用 K 近邻算法动态划分类别区间, 并利用密度峰值改进 SMOTE 算法解决类别不平衡问题, 采用极限梯度提升算法构建分类模型. 考虑到空调延迟开启或提前开启对企业效益造成的损失不同, 采用多角度综合评价方法对模型进行评估. 通过与支持向量机(现用模型)等多种预测模型的对比实验, 验证了组合模型的有效性和实用性. 实验表明组合模型平均绝对误差为 3.2 min, 与现用模型相比, 组合模型折标能耗降低了 14.71%.

[关键词] 中央空调达标时间, 组合预测模型, 随机森林, 深度特征提取, 过拟合, 样本不平衡

[中图分类号] TP181 **[文献标志码]** A **[文章编号]** 1672-1292(2023)01-0056-10

Forecasting Model of the Time Required for Central Air Conditioning to Achieve Control Effect Based on Divided Working Mode

Wang Ke¹, Lu Huanda², Zheng Junhong¹, He Lili¹

(1. School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

(2. School of Computer and Data Engineering, NingboTech University, Ningbo 315100, China)

Abstract: In order to solve the prediction problem of the time for central air conditioning to achieve the control effect, while meeting central air conditioning temperature-humidity control process requirements and enterprise energy saving reduction requirements, a combined prediction model is proposed in accordance with different working methods of air conditioning units. In the mode of Heating-Humidifying, the classification prediction model is constructed by Random Forest algorithm, and the advanced features extracted from depth features are used as the input of the model, which solves the over-fitting problem of small sample classification prediction. IPSO optimization parameters are used to reduce the time complexity of the algorithm. In the working mode of Cooling-Dehumidifying, the algorithm uses KNN to dynamically divide the category interval and uses density peak to improve SMOTE algorithm, so as to solve the problem of unbalanced sample. Then, the XGBoost algorithm is used to build the classification model. Considering that the air conditioning delay is turned on to generate a large loss to the workshop operation, a multi-angle comprehensive evaluation method is adopted to evaluate the model. Finally, the effectiveness and practicality of the combined model is verified by comparative experiments with a variety of predictive models. Experimental results show that the mean absolute error of the combined model is 3.2 minutes, and that the energy consumption of the combined model is 14.71% lower than that of the current model.

Key words: time required for central air conditioning to achieve control effect, combined prediction mode, random forest, depth feature extraction, over-fitting problem, sample unbalanced

中央空调系统作为企业生产链中重要一环, 是保证车间能够正常生产的基础. 中央空调系统属于多变量、大滞后的非线性复杂系统, 在监控管理空调设备的同时产生了大量运行数据, 如何从中挖掘出有效信息并加以利用, 是实现中央空调节能降耗目标的关键. 在许多工业企业中, 中央空调系统往往需要提前

收稿日期: 2021-08-27.

基金项目: 浙江省重点研发计划项目(2021C01110).

通讯作者: 郑军红, 博士, 讲师, 研究方向: 数据挖掘、人工智能. E-mail: zdzhengjh@sohu.com

开启来保证在某一时刻满足生产中对温湿度的要求,提前开启空调虽然保证了企业的正常运转,但也存在空调提前运行时间远大于空调达标时间的情况. 如何既保证满足车间生产环境温湿度要求,又使得能耗尽可能小,已成为一个亟需解决的问题^[1-3].

空调达标时间即从空调开启直至在空调作用下环境温湿度稳定在设定区间内的时间长度,由于其受到外界气象、建筑结构、生产任务等众多因子影响,结构关系复杂,因此机理模型不适用于解决该问题. 现阶段主要应用在空调达标时间方向的预测技术多为黑盒模型. 文献[4]采用一种基于 Elman 神经网络的空调最佳启停控制的预测系统,并应用于家庭智能控制器,有效减少了空调的预冷时间,降低了建筑物能耗. 文献[5]提出采用长短期记忆网络对空调启动时间进行预测,考虑了当前系统的状态可能受很长时间之前系统状态的影响,处理和分析了长延迟时间序列数据的内在联系,提高了预测精度. 文献[6]用聚类和 HCMAC 神经网络组合预测算法,建立空调系统最优启停模型,在满足用户舒适度的前提下,实现了最大限度节能.

由于影响空调达标时间的因素复杂,需要根据具体项目的运行特点和工业标准,有侧重地选择预测方法,对其进行改进以满足企业需求. 本文即针对某大型企业工作日中空调达标时间预测问题展开研究.

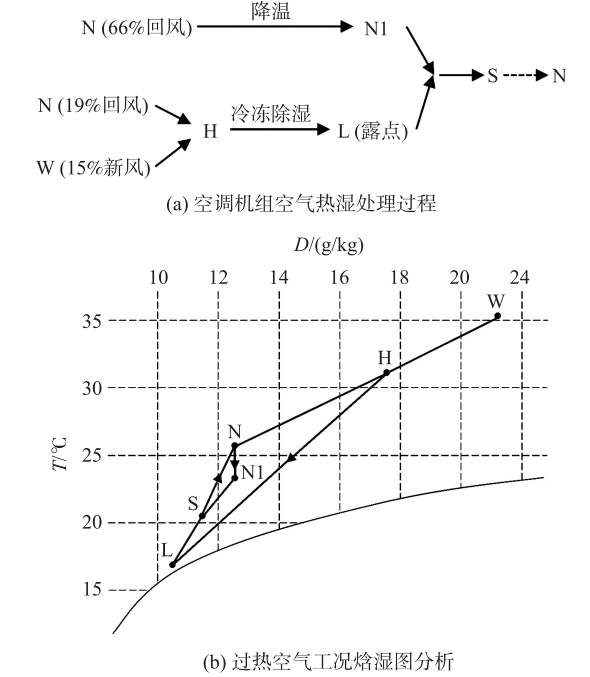
1 预测对象描述

以空调启动时刻为计时起点,在空调机组作用下,生产环境温湿度稳定在要求区间的时刻为计时终点,这个时间长度记做空调达标时间. 空调机组对空气的处理方式直接影响着空调达标时间.

空调机组对空气的处理方式分为两种. 在夏季,初始环境空气的温湿度通常高于生产工艺要求的温湿度,过热空气和部分回风经初中效过滤器过滤为干净空气,由主表冷盘管冷冻除湿处理至机器露点来保证生产环境湿度恒定,经副表冷盘降温来保证生产环境的温度恒定. 其处理过程如图 1 所示.

在冬季,通常初始环境空气的温湿度低于生产工艺要求的温湿度,过冷空气和回风经由初中效过滤器过滤干净后,利用干蒸汽加热盘管进行加热,再通过干蒸汽加湿器进行等温加湿后送入生产环境中. 其处理过程如图 2 所示.

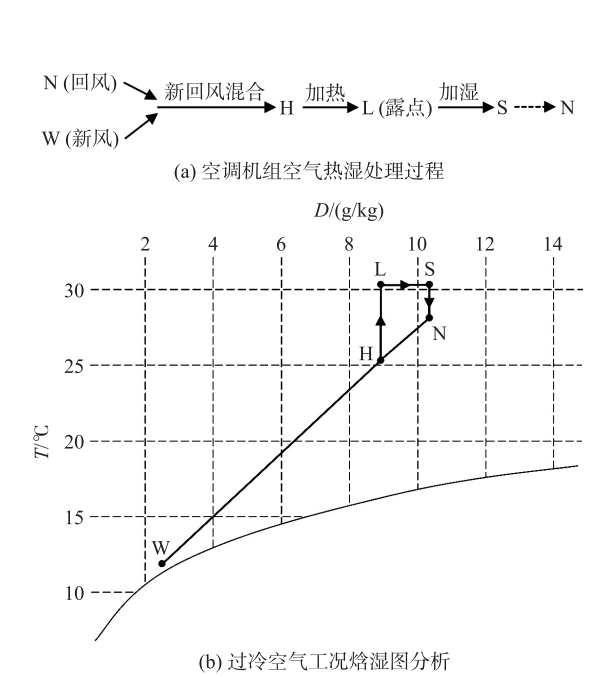
由于图 1 和图 2 的空调机理不同,研究空调达标时间预测问题要在划分空调机组工作方式的基础上,结合中央空调系统存储的历史数据,来构建预测模型.



W, 室外可能出现的状态点; C, 新回风混合状态点; S, 加热器和加湿器间的状态点; L, 夏季露点; N、N1, 车间恒温恒湿状态点; H, 冷源和热源相互抵消的热能焓值.

图 1 空调机组处理过热空气流程

Fig. 1 Process of air conditioning unit handling superheated air



W, 室外可能出现的状态点; C, 新回风混合状态点; S, 加热器和加湿器间的状态点; L, 夏季露点; N, 车间恒温恒湿状态点; H, 冷源和热源相互抵消的热能焓值.

图 2 空调机组处理过冷空气流程

Fig. 2 Process of air conditioning unit handling supercold air

2 数据描述与预处理

2.1 数据描述

采集中央空调系统存储的卷包车间空调机组和制丝生产区域空调机组工作日产生的数据集,数据集产生时间为 2018 年 1 月 2 日至 2021 年 3 月 19 日. 通常情况下,空调机组每天在早上 6:00—8:00 时间段内启动一次,系统每天会采集到一条如表 1 所示的数据. 其中,空调达标时间是启动时间与温湿度稳定时间的差值.

表 1 数据样例
Table 1 Data sample

编号	变量名称	样例值	编号	变量名称	样例值	编号	变量名称	样例值
1	空调机组编号	K2	6	风速	2.9 m/s	11	室内湿度稳定值	60.1 hPa
2	启动时间	2021/03/15 07:05:00	7	露点	12.3 ℃	12	温湿度稳定时间	2021/03/15 07:10:30
3	启动时室内温度	25.55 ℃	8	室外温度	13.67 ℃	13	空调达标时间	330 s
4	启动时室内湿度	58.5 hPa	9	室外湿度	77.5 hPa	14	空调模式	1:加温加湿;2:除温除湿
5	送风频率	16.799 Hz	10	室内温度稳定值	25.68 ℃	15	写入时间	2021/03/15 07:19:42

2.2 数据预处理

2.2.1 特征初步筛选

根据工作人员相关经验和专业知识的判断,数据样例中某些特征与达标时间相关性不大,可初步筛去. 例如,系统写入时间对空调达标时间无影响,可直接去除. 最终保留 8 个特征,包括启动时间、启动时室内温度、启动时室内湿度、送风频率、风速、露点、室外温度、室外湿度.

2.2.2 样本筛选和缺失值处理

以达标时间作为筛选样本条件. 当样本的达标时间在 10~180 min 范围内,视为有效样本. 在数据采集过程中,传感器故障会导致某些数据存储空间失败,造成单值缺失和多值缺失. 单值缺失采用上下平均值填补法进行处理. 在多值缺失情况中,若缺失特征较少,则采用样条插值法进行处理,在给定区间 $[a, b]$ 上划分 $a = x_0 < x_1 < \dots < x_n = b$,若函数 $S(x)$ 在每个区间 $[x_i, x_{i+1}]$ ($i=0, 1, \dots, n-1$)上是 m 次多项式,且在 $[a, b]$ 上具有 $m-1$ 阶连续导数,则 $S(x)$ 划 m 次样条函数用于处理连续缺失值;若数据异常或缺失超过 60%,很难估算出缺失数据,则直接删除. 因数据集缺失数据不足 1%,因此删除操作不会影响空调达标时间预测精度,也不会对其他数据造成影响.

2.2.3 相关性分析与特征选择

由于特征集分布存在两种情况,对于不满足正态分布或为定序变量的特征,计算其斯皮尔曼等级相关系数(Spearman's rank correlation coefficient),并筛选出相关系数大于 20%的特征,计算公式为:

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

若特征值为连续变量,利用皮尔森相关系数(Pearson correlation coefficient)计算并筛选出相关系数大于 20%的特征,计算公式为:

$$\rho(X, Y) = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sqrt{\sum_{i=1}^n (X_i - \mu_X)^2} \sqrt{\sum_{i=1}^n (Y_i - \mu_Y)^2}}, \quad (2)$$

式中, $\rho(X, Y)$ 是自变量 X 与因变量 Y 的相关系数值; μ_X 是自变量 X 的平均值; μ_Y 是因变量 Y 的平均值.

最终共筛选出 6 个输入变量,分别为启动时室内温度、启动时室内湿度、室外温度、室外湿度、露点和风速.

2.2.4 无量纲化处理

采用 Z-score 标准化对样本进行处理,原始数据集 $x = \{x_1, x_2, \dots, x_n\}$ 标准化公式为:

$$\hat{x} = \frac{x - \mu}{\sigma}, \quad (3)$$

式中, μ 为原始数据均值; σ 为原始数据标准差.

3 模型构建与算法原理

预处理后的有效数据为 1 162 条,将处理后的数据集按照空调机组工作方式划分. 加温加湿工作方式下的样本属于均衡非线性小样本数据,样本量为 339 条;降温除湿工作方式下的样本共 823 条且具有不平衡、非线性和多维数的特点. 针对不同工作方式的样本特点,设计对应算法来构建预测模型.

3.1 中央达标时间整体预测流程

本文提出一种组合预测模型. 在加温加湿工作方式下,采用一种改进粒子群优化参数的深度特征提取随机森林算法 (IPSO-DFERF) 来构建预测模型. 在降温除湿工作方式下,构建基于 K 邻近-极端梯度提升算法 (KNN-XGBoost) 的预测模型. 考虑到空调延迟启动和提早启动对企业运作产生的效益损失不同,从预测精度、能耗和企业效益损失 3 个方面对模型进行综合评估. 本文的整体预测流程如图 3 所示.

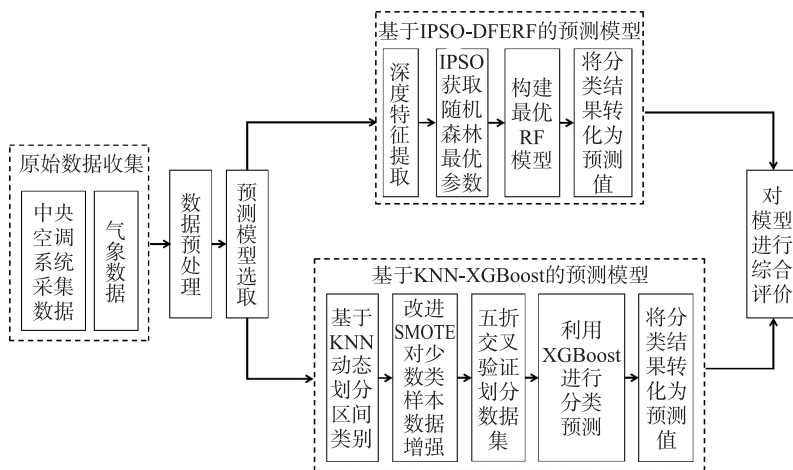


图 3 中央空调达标时间预测流程

Fig. 3 Prediction process of the time required for central air conditioning to achieve control effect

3.2 加温加湿工作方式下的模型构建

对于加温加湿工作方式下的均衡非线性小样本数据,采用随机森林 (random forest, RF) 算法构建达标时间分类预测模型. 利用深度特征提取 (depth feature extraction, DFE) 方法将原始特征抽象提取为高级特征作为输入,解决 RF 对小样本进行分类预测时出现的过拟合缺陷. 为了降低算法的时间复杂度,使用 IPSO 对算法参数寻优. 为避免某个类别有过少的样本量,按照达标时间升序排列样本集,以 2 min 为步长均匀划分样本集,把划分的每个区间记为一类. 据预测分类结果,选取其区间最大值作为最终预测值.

3.2.1 算法原理

空调达标时间受多种特征因素影响,特征间相互作用、关系复杂. 样本集 $X = \{x_1, x_2, \dots, x_m\}$ 中的每个样本都有特征集 $x_i = \{a_1^i, a_2^i, a_3^i, a_4^i, a_5^i, a_6^i\}$. 为得到各特征间的不确定关系并使分类预测模型能有较好的精度,将 6 个特征组成的原始向量乘以其转置得到 6×6 的高维矩阵作为输入,利用卷积过程中的共享权值和共享偏差获取深度特征^[7-8],提取流程如图 4 所示.

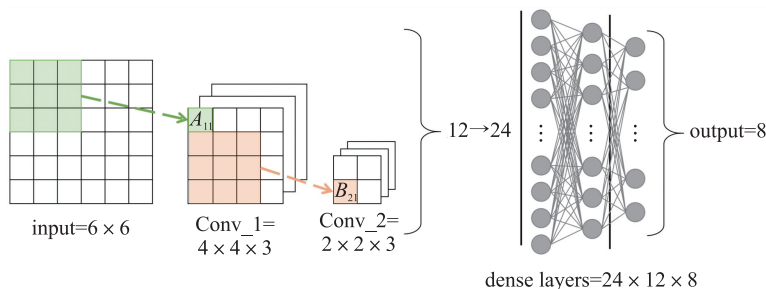


图 4 深度特征提取流程

Fig. 4 Depth feature extraction process

第一层卷积使用 3 个大小为 3×3 的过滤器,步幅设置为 1,使用 valid padding. 以 A_{11} 为例,第一次的卷积运算公式为:

$$A_{11} = F\left(\sum_{i=1}^3 \sum_{j=1}^3 Z_{ij} W_{ij}^{L_1} + D_{L_1}\right), \quad (4)$$

式中, F 为修正线性单元(rectified linear unit, ReLU); Z_{ij} 表示输入矩阵第 i 行第 j 列参数; L_k 表示输出矩阵的通道深度, Conv_1 的通道深度用 L_1-L_3 表示; $W_{ij}^{L_k}$ 为 L_k 层的共享权值,每个过滤器都有 $3 \times 3 = 9$ 个共享权值; D_{L_k} 表示 L_k 层的共享偏差.

第二层卷积同样使用 3 个大小为 3×3 的过滤器, Conv_2 的通道深度用 L_4-L_6 表示,以 $B_{21}^{L_4}$ 为例, Conv_2 运算公式为:

$$B_{21}^{L_4} = F\left(\sum_{i=2}^4 \sum_{j=1}^3 \sum_{k=1}^3 A_{ijk} W_{ijk}^{L_4} + D_{L_4}\right). \quad (5)$$

将卷积后的矩阵拉伸为向量送至全连接层,最后提取出高级特征 $x_i = \{A_1^i, A_2^i, \dots, A_8^i\}$. 将处理过的数据集和高级特征作为随机森林分类预测模型的输入.

使用深度特征提取随机森林算法(DFERF)构建基础分类模型. 对含有 m 个样本的训练集,通过自助采样法有放回地采集出 T 个含 m 个样本的采样集,保证样本子集的差异性. 以决策树为基学习器,从基决策树中结点的特征集合(共有 d 个特征)中随机选择一个有 k 个($k < d$)特征的子集用于训练基学习器,每个样本子集的特征也具有差异. 最终集成 T 个基学习器,使用简单投票法得出预测结果^[9-10].

为进一步降低算法的时间复杂度,利用改进粒子群算法(IPSO)寻优参数 T 和 k ^[11-14]. 以位置和速度作为自变量组合,根据自身最优解 $pbest_i$ 和群体最优解 $gbest_i$ 不断调整来寻找最优参数,在迭代过程中第 i 个粒子自身最佳位置和粒子群最佳位置为:

$$\begin{cases} pbest_i = (p_{i1}, p_{i2}, \dots, p_{iD}), i = 1, 2, \dots, N; \\ gbest_i = (g_{i1}, g_{i2}, \dots, g_{iD}), i = 1, 2, \dots, N. \end{cases} \quad (6)$$

其速度更新和位置更新公式分别为:

$$v_i^{(n+1)} = wv_i^n + c_1 r_1 (pbest_i - x_i^n) + c_2 r_2 (gbest_i - x_i^n), \quad (7)$$

$$x_i^{(n+1)} = x_i^n + v_i^{(n+1)}, \quad (8)$$

式中, w 为惯性权重; c_1 和 c_2 为加速因子; r_1 和 r_2 是随机函数; n 为迭代次数; x_i^n 和 v_i^n 分别为第 n 次迭代中第 i 个粒子的位置参数和速度参数.

为保证前期遍历完整个空间,要求粒子提升自身多样性,后期增强局部寻优能力,同时加强粒子的收敛能力,采用线性微分递减策略来调节惯性权重和加速因子:

$$w = w_{\max} - \frac{(w_{\max} - w_{\min})n^2}{n_{\max}^2}; \quad (9)$$

$$\begin{cases} c_1 = c_{1\max} - \frac{(c_{1\max} - c_{1\min})n^2}{k_{\max}^2}, \\ c_2 = c_{2\min} + \frac{(c_{2\max} - c_{2\min})n^2}{k_{\max}^2}; \end{cases} \quad (10)$$

式中, w_{\max} 为最大惯性权重; w_{\min} 为最小惯性权重; n 为当前迭代次数; n_{\max} 为最大迭代次数; $c_{1\max}$ 、 $c_{2\max}$ 分别为 c_1 、 c_2 的最大值; $c_{1\min}$ 、 $c_{2\min}$ 分别为 c_1 、 c_2 的最小值.

3.2.2 模型构建

构建基于 IPSO-DFERF 算法的模型,预测流程如图 5 所示.

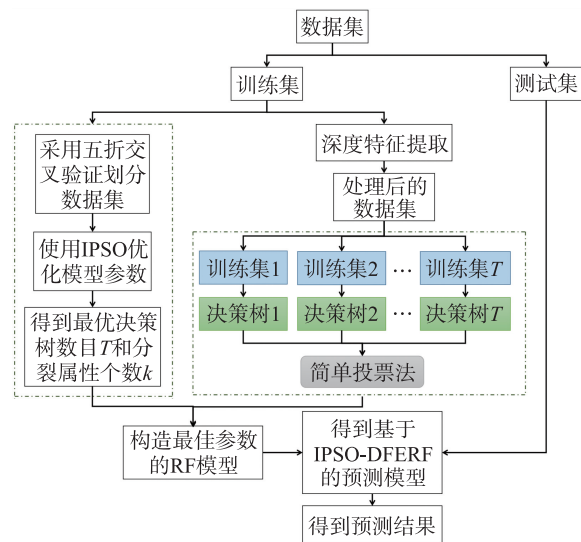


图 5 加热加湿工作方式下的预测模型

Fig. 5 Prediction model under heating-humidification working mode

3.3 降温除湿工作方式下的模型构建

针对降温除湿工作方式下的样本具有不均衡、非线性和多维数的特点,构建基于 KNN-XGBoost 算法的预测模型.为解决现类别中样本不均衡问题,先利用 KNN 重新划分类别区间,以此减弱类别间的样本不平衡程度,再利用密度峰值改进 SMOTE 算法解决此问题.将构建后的数据样本作为模型输入,基于 XGBoost 算法构建分类预测模型.

3.3.1 算法原理

首先,利用 KNN 动态划分类别区间.若某个原类别的样本数大于总样本数的 20%,则标为一类;剩余样本按达标时间升序排列,使用 KNN 动态划分每个类别的区间^[15],以 0.5 min 为初始步长,当类别区间样本数大于总样本数的 20%,则标为一类;若不满足条件,则步长以 0.5 min 递增,最大不超过 2 min,直至满足条件时划为一类.具体流程如算法 1 所示.

算法 1 动态划分类别区间

已知特征集 X ,不同达标时间升序排列组成集合 $Y = \{y_1, y_2, \dots, y_n\}$,每个达标时间对应的样本数为 $\text{num} = \{\text{num}_{y_1}, \text{num}_{y_2}, \dots, \text{num}_{y_n}\}$;类别数为 L ,初始值设定为 0;总样本数目为 NUM .

```

1: for  $i = 1$  to  $n$ 
2:   if  $\text{num}_{y_i} > 0.2 * \text{NUM}$ 
3:      $L++$ 
4:   else
5:     更新集合  $Y^* = \{y_1^*, y_2^*, \dots, y_k^*\}$ ,则新集合对应的样本数为  $\text{num}^* = \{\text{num}_{y_1^*}, \text{num}_{y_2^*}, \dots, \text{num}_{y_k^*}\}$ 
6:   end if
7: end for
8: 令  $j = 1, R = \{r_1, r_2, r_3, r_4\} = \{30, 60, 90, 120\}$ 
9: for  $m = j$  to  $k$ 
10: 以  $R$  为步长利用 KNN 动态划分区间
11:  if  $\sum_{i=j}^m \text{num}_{y_i}^* \geq 0.2 \text{NUM}^*$ 
12:     $L++$ 
13:  end if
14:   $j = m + 1$ 
15: end for

```

通过上述步骤,将数据集划分为 L 个类别,每个类别都有对应达标时间区间.

为了避免增加少量数据的噪点,采用利用密度峰值(DPCA)对少数类样本进行聚类操作,再进行 SMOTE 构建新样本^[16].算法流程为:

- Step 1** 采用 DPCA 对少数类样本进行聚类,选取 m 个类簇中心;
- Step 2** 将多数类样本划分 n 组,与 m 个类簇中心构成新样本;
- Step 3** 对于各组样本,通过计算新样本中少数类样本 X 与所有样本的距离,得到 K 近邻;
- Step 4** 确定各组采样倍率 N ,在 X 的 K 近邻中随机选取若干样本,设定选择近邻为 b ;
- Step 5** 根据公式(11)构建各组的新样本:

$$c = a + \text{rand}(0, 1) * |a - b|. \quad (11)$$

极端梯度提升算法(eXtreme Gradient Boosting, XGBoost)的核心思想为:通过不断添加树去学习一个新函数 $f(x)$ 来拟合上次预测的残差,不断进行特征分裂生长一棵树;训练完成后得到 k 棵树,预测样本的特征在每棵树都会对应叶子节点,一个叶子节点则对应一个分数;将每棵树对应的分数相加得到样本预测值^[17-19].算法流程为:

- Step 1** 选择梯度提升树作为弱评估器;
- Step 2** 构建 XGBoost 目标函数,采用 softmax 损失函数进行计算;
- Step 3** 求解目标函数,通过将目标函数泰勒展开为二阶进行简化;

Step 4 寻找最佳树结构,求解出叶子数量 T 和每片叶子权重 w ,最优叶子节点分数公式和目标函数公式如下:

$$w_j = -\frac{G_j}{H_j + \lambda}, \quad (12)$$

$$Obj^{(t)} = -\frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \lambda} + \gamma T; \quad (13)$$

Step 5 寻找最佳分支,利用贪婪算法控制局部最优来达到全局最优,使用目标函数定义结构优劣,计算分支前与分支后结构分数之差并记为 Gain,选择 Gain 最大特征上的分支点进行分支;

Step 6 设定复杂性控制变量 γ 和 λ ,其中 γ 控制叶子节点个数, λ 控制叶子节点分数,当 Gain 值小于某值时,令树停止生长,防止过拟合.

3.3.2 模型构建

构建基于 KNN-XGBoost 算法的分类预测模型,预测流程为:

Step 1 基于 KNN 动态划分区间类别;

Step 2 利用改进 SMOTE 对少样本类进行数据增强;

Step 3 采用五折交叉验证对数据集进行划分;

Step 4 利用 XGBoost 算法进行五次分类预测,采用简单投票法得出最后分类结果;

Step 5 根据分类结果,选取对应区间最大值作为最终预测的达标时间值.

3.4 模型综合评价方法

模型从预测精度、能耗和效益损耗 3 方面进行综合评价. 模型预测精度的评价指标选用均方根误差 (root-mean-square error, RMSE) 和平均绝对误差 (mean absolute error, MAE), 计算公式为:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (14)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|,$$

式中, y_i 为实际达标时间; \hat{y}_i 为预测达标时间; n 为总样本数.

空调提前开启会在规定时间前生产环境就已到达工艺生产温湿度要求,会导致电量的过度损耗. 空调启动的能源损耗主要由用电量、冷量构成,其中耗电量 = 冷量/COP. 电力折标煤系数为 0.123,折标能耗^[20-21]计算公式为:

$$E = 0.123 \times (Q_1 + Q_2), \quad (15)$$

式中, E 为折损能耗; Q_1 为用电量; Q_2 为冷量的耗电量.

空调延迟开启会在规定时间内生产环境无法到达工艺生产的温湿度要求,导致工作任务不能完成,原料也会有一定程度的损坏,给企业效益造成很大损失. 效益损失计算公式为:

$$Cost = C \times T, \quad (16)$$

式中, Cost 为生产物料损耗造成的效益损失; C 为空调延迟开启情况下每分钟的效益损失; T 为预测值小于实际值的误差时间.

4 实验与结果分析

该企业卷包车间空调机组共有 10 套空调系统,空调区域要求全年温度控制在 $(24 \pm 3)^\circ\text{C}$,湿度控制在 $(58 \pm 5)\%$,新回风比例为新风 15%、回风 85%. 制丝车间共有 4 套空调系统,要求全年温度控制在 $(25 \pm 3)^\circ\text{C}$,湿度控制在 $(65 \pm 5)\%$,新回风比例为新风 25%、回风 75%. 空调延迟开启每分钟造成的物料损坏按 0.1 kg 计算,每千克物料价值按 100 元计算.

4.1 模型训练

4.1.1 加热加湿工作方式的模型参数设定

将处理过的小样本数据作为模型输入. 为防止过拟合,限制树的最大深度为 3. 随着随机森林训练次

数增多,模型的准确率能稳定在 96% 以上,MAE 为 3.1 min, RMSE 为 6 min. 最终模型的主要参数设定如表 2 所示.

4.1.2 降温除湿工作方式的模型参数设定

将该工作方式下处理后的数据作为模型输入,不断调整模型参数,最终模型的主要参数设定如表 3 所示.

表 2 加热加湿工作方式下模型的主要参数
Table 2 Main model parameters under heating-humidification working mode

模型参数	最终设定值
类别 L	9
限制树的最大深度	3
IPSO 寻优迭代次数	200
决策树数目 T	60
分裂属性 k	2

表 3 降温除湿工作方式下模型的主要参数
Table 3 Main model parameters under cooling-dehumidification working mode

模型参数	最终设定值
类别 L	12
叶子节点个数 γ	20
叶子节点分数 λ	36
迭代速率	0.05

利用动态划分类别区间并结合改进 SMOTE 算法处理类别不平衡问题,结果如图 6 所示.

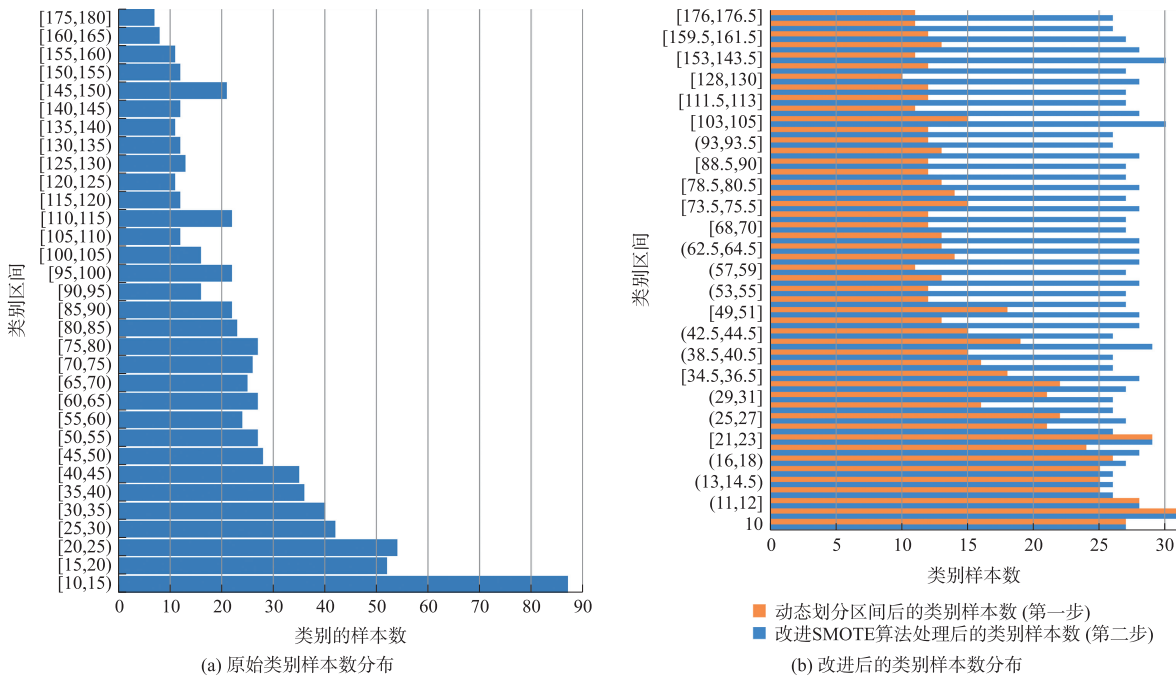


图 6 类别不平衡问题处理结果
Fig. 6 Result of handling category imbalance problems

五折交叉验证的误差结果如表 4 所示.

表 4 五折交叉验证下的实验结果误差

Table 4 Error of experimental results under 5-Fold cross validation

K (第 K 次交叉验证)	1	2	3	4	5
RMSE/min	3	3.3	3.4	3.2	3.2
MAE/min	6.1	6.3	6.5	6.2	6.2
平均误差值	RMSE: 6.3 min; MAE: 3.2 min				

4.2 对比实验与结果分析

为验证本文组合模型的有效性,分别选用 IPSO-DFERF、KNN-XGBoost、支持向量机 (support vector machine, SVM)、朴素贝叶斯 (naive Bayesian model, NBM) 预测模型与本文组合模型进行对比实验. 选用 IPSO-DFERF 模型和 KNN-XGBoost 模型是为了验证组合模型的有效性; SVM 模型^[22-23]是目前企业正实际使用的空调达标时间预测模型,通过此模型来还原复现企业历史空调达标时间,验证本文模型预测精度的

提升;NBM^[24]算法适合处理非线性复杂问题,对小样本分类问题也有很好的处理效果,使用此对比模型是用以验证本文模型的优越性.对比模型同样是在五折交叉验证的同等条件下进行预测的,5 种预测模型结果如表 5 所示.

表 5 模型结果
Table 5 The results of the model

模型指标	RMSE/min	MAE/min	折标能耗/kgce	企业效益损失/元
本文模型	6.2	3.2	1.97	3.3
IPSO-DFERF	9.6	6.1	0.47	126
KNN-XGBoost	10.8	7.7	2.14	72.3
SVM	13.58	10.0	2.31	100.33
NBM	13.27	8.87	1.24	113.10

相较于 SVM 模型,本文模型 RMSE 误差值降低了 54.34%,MAE 误差值降低了 68%;每次空调提早开启时,折标能耗降低了 14.71%;每次空调延迟开启时,本文模型的企业效益损失减少 97.03 元;说明本文模型的预测精度优于现用模型.与 IPSO-DFERF 模型相比,本文模型 RMSE 误差值降低了 35.42%,MAE 误差值降低了 47.54%,虽然折标能耗略微增高,但企业效益损失却大大减少;与 KNN-XGBoost 模型相比,本文模型 RMSE 误差值降低了 42.59%,MAE 误差值降低了 58.44%,折标能耗消耗和企业效益损失都更低;这两种模型验证了组合模型的预测效果更好,说明了采取划分空调工作方式策略的重要性.与 NBM 模型相比,本文模型 RMSE 误差值降低了 53.28%,MAE 误差值降低了 63.92%,折标煤能耗略微增加,企业效益损失减少 109.8 元.综合考量可知,本文提出的模型更贴合该企业的实际生产需求.

5 结论

本文以解决中央空调达标时间预测问题为目标展开研究.首先,明确空调达标时间是在空调机组作用下,生产区域从原始环境温湿度到满足工艺要求温湿度环境的时间.通过空调机组对热空气和冷空气不同的工作方式,提出了在划分工作方式的基础上对空调达标时间进行预测的策略.通过分析加温加湿工作方式和降温除湿工作方式下数据的不同特点,采用了不同的算法构建分类预测模型.对于加温加湿工作方式下的均衡非线性小样本数据,采用基于 IPSO-DFERF 算法来构建模型;针对降温除湿工作方式下的样本具有不均衡、非线性和多维数的特点,构建了基于 KNN-XGBoost 算法的预测模型.考虑到空调延迟开启与提前开启对企业运作产生损失不同,提出了一种综合评价方法,从预测精度、能耗和效益损失 3 个角度来评价模型.最终将组合预测模型应用于中央空调达标时间预测,本文模型平均绝对误差为 3.2 min,均方根误差为 6.2 min,相比于现用 SVM 模型,组合模型折标能耗降低了 14.71%,企业效益损耗也大大减少.通过与多种模型对比,验证了本文策略的有效性和模型的实际应用性.

随着企业生产的进行,中央空调系统会继续产生大量数据,模型需要不断补充数据集进行校正,从而进一步提升模型预测精度.后续将进一步研究中央空调蒸汽量预测、中央空调开机台数预测等问题,通过对蒸汽量和开机台数准确预测,减少能源浪费,实现企业节能降耗目的.

[参考文献](References)

- [1] 杨裔. 中央空调节能技术综述[J]. 现代信息技术, 2019, 3(13): 193-194.
- [2] SHENG X, SHANG J F, ZHANG J H, et al. An energy-saving based predictive control strategy on central air-conditioning system[C]//Proceedings of 2009 Transmission & Distribution Conference & Exposition: Asia & Pacific. Seoul, Korea: IEEE, 2009.
- [3] FENG Z X, REN Q C, YU J Q. The optimum start-up of central air-conditioning based on neural network[C]//Proceedings of 2013 3rd International Conference on Computer Science and Information Technology. Chengdu, China: IEEE, 2010.
- [4] 计建惠,王可崇.基于 Elman 网络的空调预冷时间的预测[J]. 微计算机信息, 2008, 24(13): 261-263.
- [5] 杨建军,何利力.基于长短期记忆网络的工业空调启动时间预测[J]. 软件导刊, 2020, 19(6): 48-52.
- [6] 李慧,段培永,罗光林.基于数据的空调系统最优启停时间预测[J]. 山东建筑大学学报, 2012, 27(2): 155-159.
- [7] 朱军,侯振杰,陈树越,等.深度特征重构与权重分配的交通标志识别算法[J]. 小型微型计算机系统, 2019, 40(9): 1932-1939.

-
- [8] 李鸿培,于畅,忽朝俭,等. 2013 工业控制系统及其安全性研究报告[R]. 北京:绿盟科技,2013.
- [9] SU C, JU S G, LIU Y G, et al. Improving random forest and rotation forest for highly imbalanced datasets[J]. *Intelligent Data Analysis*, 2015, 19(6):1409–1432.
- [10] CAO X H, LI R J, GE Y M, et al. Densely connected deep random forest for hyperspectral imagery classification[J]. *International Journal of Remote Sensing*, 2019, 40(9):3606–3622.
- [11] SHAKIBA R, NAJAFIPOUR M R, SELEHI M E. An improved PSO-based path planning algorithm for humanoid soccer playing robots[C]//*Proceedings of the 2013 3rd Joint Conference of AI & Robotics and 5th RoboCup Iran Open International Symposium*. Tehran, Iran:IEEE, 2013.
- [12] QIAN W Y, LI M. Convergence analysis of standard particle swarm optimization algorithm and its improvement[J]. *Soft Computing*, 2018, 22(12):4047–4070.
- [13] HOU G Y, XU Z D, LIU X, et al. Improved particle swarm optimization for selection of shield tunneling parameter values[J]. *Computer Modeling in Engineering and Sciences*, 2019, 118(2):317–337.
- [14] CHANG W D. An improved PSO algorithm for solving nonlinear programming problems with constrained conditions[J]. *International Journal of Modeling, Simulation, and Scientific Computing*, 2021, 12(1):2150001.
- [15] 王志华,刘绍廷,罗齐.基于改进 K-modes 聚类的 KNN 分类算法[J]. *计算机工程与设计*, 2019, 40(8):2228–2234.
- [16] SUN J, LANG J, FUJITA H, et al. Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates[J]. *Information Sciences*, 2018, 425:76–91.
- [17] ZHANG D H, QIAN L Y, MAO B J, et al. A data-driven design for fault detection of wind turbines using random forests and XGBoost[J]. *IEEE Access*, 2018, 6:21020–21031.
- [18] TAHMASSEBI A, WENGERT G J, HELBICH T H, et al. Impact of machine learning with multiparametric magnetic resonance imaging of the breast for early prediction of response to neoadjuvant chemotherapy and survival outcomes in breast cancer patients[J]. *Investigative Radiology*, 2019, 54(2):110–117.
- [19] ZHU M, WU K, ZHOU Y Z, et al. Prediction of cooling moisture content after cut tobacco drying process based on a particle swarm optimization-extreme learning machine algorithm[J]. *Mathematical Biosciences and Engineering*, 2021, 18(3):2496–2507.
- [20] 国家市场监督管理总局, 国家标准化管理委员会. 综合能耗计算通则:GB/T 2589–2020[S]. 北京:中国标准出版社, 2020.
- [21] 简毅文,江亿. 住宅供暖空调能耗计算模式的研究[J]. *暖通空调*, 2005, 35(2):11–14.
- [22] XU L, HOU L, ZHU Z Y, et al. Midterm prediction of electrical energy consumption for crude oil pipelines using a hybrid algorithm of support vector machine and genetic algorithm[J]. *Energy*, 2021, 222(1):119955.
- [23] BESROUR A, KSANTINI R. Incremental subclass support vector machine[J]. *International Journal on Artificial Intelligence Tools*, 2019, 28(7):1950020.
- [24] ALKUBAISI G. The role of ensemble learning in stock market classification model accuracy enhancement based on Naive Bayes Classifiers[J]. *International Journal of Statistics and Probability*, 2020, 9(1):1–36.

[责任编辑:严海琳]

(上接第 45 页)

- [29] YAN Q, ZHANG L, LIU Y, et al. Deep HDR imaging via a non-local network[J]. *IEEE Transactions on Image Processing*, 2020, 29:4308–4322.
- [30] SHI W, CABALLERO J, HUSZÁR F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. USA: IEEE, 2016:1874–1883.
- [31] MARTIN D, FOWLKES C, TAL D, et al. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics[C]//*Proceedings of the Eighth IEEE International Conference on Computer Vision*. ICCV 2001. USA:IEEE, 2002:416–423.

[责任编辑:陈 庆]