doi:10.12068/j.issn.1005－3026.2024.10.001

基于Stacking集成的RF-ET-KDE烧结过程物理指标区间预测模型

康增鑫，陈进朝，王金杨，吴朝霞
（东北大学秦皇岛分校 控制工程学院，河北 秦皇岛 066004）

摘 要：由于烧结过程中存在众多不确定性因素，使得机理分析和点预测结果的可靠性不足．基于此提出随机森林-极限树-核密度估计（random forest-extreme tree-kernel density estimation，RF-ET-KDE）算法对物理指标（粒度、水分）进行区间预测．首先，采用数据预处理和特征选择操作筛选出最适合建模的特征变量．其次，使用基于Stacking的RF-ET算法对指标进行点预测，该算法使得模型有较高的准确性和泛化性．然后，采用KDE算法计算指标的预测误差，得到了一定置信水平下的分布区间和区间预测结果．最后，用所建模型与其余组合模型进行对比．结果表明，RF-ET算法有较高的点预测效果，KDE算法可以很好地量化指标的误差，可以得到较高可靠度的区间预测结果．

关 键 词：烧结过程；随机森林-极限树；核密度估计；物理指标；区间预测
中图分类号：TP 181 **文献标志码：**A **文章编号：**1005-3026(2024)10-1369-10

Interval Prediction Model of RF-ET-KDE Sintering Process Physical Index Based on Stacking Integration

KANG Zeng-xin, *CHEN Jin-chao*, *WANG Jin-yang*, *WU Zhao-xia*
(School of Control Engineering, Northeastern University at Qinhuangdao, Qinhuangdao 066004, China. Corresponding author: WU Zhao-xia, E-mail: ysuwzx@126.com)

Abstract: Due to the many uncertainties in the sintering process, the reliability of mechanism analysis and point prediction results is insufficient. Therefore, a random forest-extreme tree-kernel density estimation (RF-ET-KDE) algorithm is proposed to realize interval predictions for physical indicators, such as particle size and moisture. Firstly, data preprocessing and feature selection operations are adopted to screen out the most suitable feature variables for modeling. Secondly, the RF-ET algorithm based on Stacking is utilized to realize point predictions for the indicators. This algorithm makes the model with higher accuracy and generalization, and then the KDE algorithm is adopted to calculate the prediction error of the indicator. The distribution interval and interval prediction results under a certain confidence level are obtained. Finally, the proposed model is compared with the other combined models. The results show that the RF-ET algorithm has higher point prediction accuracy, and the KDE algorithm can quantify the error of the indicator very well, so that a higher credibility interval prediction result can be obtained.

Key words: sintering process; random forest-extreme tree (RF-ET); kernel density estimation (KDE); physical index; interval prediction

烧结矿是高炉炼铁的主要原料，其质量对环境保护和高炉冶炼尤为重要[1]．烧结过程是工艺流程长、影响因素多、机理复杂的动态系统．烧结矿需要经过配料、加水混合、布料、点火烧结、破碎、成品冷却、筛分、整粒、取样等工序，花费3个多小时才能得到最终的烧结矿，这种烧结过程使得对烧结矿质量的检测具有滞后性，操作员根据实验数据计算出来的结果不能表明当前烧结矿

收稿日期：2023-05-24
基金项目：河北省教育厅科学技术研究项目(BJ2021099)．
作者简介：康增鑫(1999-)，男，河北石家庄人，东北大学硕士研究生；吴朝霞(1969-)，女，浙江嘉兴人，东北大学教授．

质量的情况,这些结果的人为因素较大,且可靠性不足,所以对烧结矿质量进行准确预测很有必要[2-4].

粒度和水分是烧结过程中非常重要的物理指标.粒度和水分的含量对最终生产的烧结矿质量有很大的影响.当前对烧结过程物理指标的研究主要集中于机理分析和数据驱动方法,采用机理分析法分析具有非线性、大滞后等问题的烧结过程难免会有较大的误差,很难建立一个标准的基于机理的模型来评价指标[5].基于经验和知识得到的分析结果很难支撑实际的实验结果,故本文采用数据驱动方法对烧结过程物理指标进行预测.目前对烧结过程物理指标的预测在国内外已有一定的研究.Jiang等[6]利用烧结过程的实时数据和历史数据,建立了烧结过程中主要混合料和水的混合料加水模型,提出了离线深度监督学习和在线自学习非线性自回归模型算法的组合,实验表明该方法可以有效地预测水分.刘月明等[7]结合烧结工艺和实际生产数据,将极限梯度提升树(extreme gradient boosting, XGBoost)算法、因子相关分析和深度神经网络(deep neural network,DNN)算法相结合的大数据技术对烧结矿小于10 mm粒级的含量进行预测,通过和其他传统算法比较,结果表明模型预测效果很好,可以达到精准预测的目的.Ren等[8]基于烧结混合料水分的变化,提出了一种新的核主成分分析-遗传算法优化方法,能够更准确地预测混合料含水率随时间的变化规律,实验数据表明,这种优化方法具有较高的拟合精度和预测精度,在处理烧结过程等具有复杂非线性特征的数据集时,该方法与传统的神经网络建模方法相比具有明显的优势.这些研究表明,利用机理分析和机器学习算法可以对烧结过程中物理指标进行比较准确地分析和预测,但这些结果无法对预测点附近的值进行评估,由此得到的分析和预测结果不能很好地评估烧结过程物理指标的最优化水平.本文采用一种集成和区间预测的方法对烧结矿尺寸小于10 mm的颗粒所占的比例和烧结混合料中水分所占的比例这2个指标进行预测,可以提高所得结果的精度和可靠度.

当前对烧结过程物理指标的研究没有同时考虑集成学习和区间预测从而导致结果的可靠性不足.本文首先采用箱形图和最大信息系数法对烧结原始数据进行预处理,分别选出20个与物理指标强相关的特征变量;其次采用基于Stacking集成的RF-ET算法对指标进行点预测;然后使用KDE算法对物理指标进行区间预测,可以提升预测结果的准确性;最后通过与其他模型对比,评估本文所建模型的预测精度.

1 数据分析与处理

1.1 数据分析

采用中国某钢铁企业3号烧结机为期2年的实际生产数据,通过查询资料对数据进行分析,并结合实际生产过程,可以将全部烧结过程参数分为以下4类:下料量原料参数、烧结混合料参数、烧结机操作参数以及烧结机状态参数,通过对烧结工艺的分析,得出了54个对物理指标有影响的参数,如表1所示.

表1 烧结过程参数

Table 1 Parameters of sintering process

参数类型	参数序号	参数	参数序号	参数
单位时间原料下料量	1	钙石灰粉/(t·h^{-1})	4	钒钛铁精粉/(t·h^{-1})
	2	除尘矿/(t·h^{-1})	5	自返矿/(t·h^{-1})
	3	燃料/(t·h^{-1})	6	高炉返矿/(t·h^{-1})
烧结混合料质量百分数	7	全铁/%	9	氧化钙/%
	8	五氧化二钒/%	10	二氧化硅/%
烧结机操作	11	圆辊速度/(r·h^{-1})	17	一号风门开度/%
	12	机速九辊/(r·h^{-1})	18	二号风门开度/%
	13	烧结机机速/(m·min^{-1})	19	机速板式转速/(r·h^{-1})
	14	点火温度/℃	20	环冷机机速/(m·min^{-1})
	15	煤气流量/(m^3·h^{-1})	21	助燃风流量/(m^3·h^{-1})
	16	风机风量/(m^3·h^{-1})	22	助燃风压力/kPa
烧结机状态	23	南烟道温度/℃	26	北烟道负压/kPa
	24	南烟道负压/kPa	27~40	风箱废弃温度/℃
	25	北烟道温度/℃	41~54	风箱负压/kPa
输出	55	水分的质量分数/%	56	粒度/%

由于烧结生产现场的传感器、测量仪等的采样时间不同,使得各个参数的数据量参差不齐,例如,样本采集周期为1 h,烧结混合料参数采集周期为4 h,水分采集周期为4 h,粒度采集周期为8 h.若不进行操作会产生大量空缺值,会影响后续建模和预测,故对采样周期为4 h和8 h的数据分别进行4倍和8倍复制,这样可以在时间维度上匹配数据,进而获得初步的物理指标数据集.表2列出了烧结原始数据,即后续建模所用数据.

表2　烧结原始数据

Table 2　Original data of sintering

数据类型	单位时间原料下料量/(t·h^{-1})						输出质量百分数/%	
	钙石灰粉	除尘矿	燃料	焦粉	铁粉	钒粉	粒度	水分
均值	30.184	9.768	13.693	5.621	14.792	72.938	24.532	8.295
最小值	11.913	0	0	0	0	0	20.500	7.500
最大值	48.753	24.248	40.821	20.935	67.366	305.117	26.400	9.100

1.2　数据预处理

由于烧结生产是在极其恶劣环境中进行的,传感器检测异常等问题无法避免,由此会产生一些异常数据.这些数据会对后续建模和模型评估产生重大影响,因此对数据进行预处理是进行特征变量选择和建模前必不可少的步骤.

当前对异常值处理的方法有:3σ原则、Z-score法、箱形图等.并不是所有方法均适用于具有非线性特点的烧结数据.3σ原则和Z-score法要求数据服从正态或者近似正态分布,而烧结的数据分布不一定满足其适用条件,故本文采用箱形图[9]法来识别并处理异常值.箱形图如图1所示.箱形图的原理如下:以3条分位线将图形分为4个区域,每个区域占比1/4.3条分位线分别为下四分位线(Q_1)、中位数(median,M_D)、上四分位线(Q_3),上四分位线和下四分位线之间的距离为四分位距(R_{IQ}).$Q_1-1.5R_{IQ}$表示数据中的低异常值的临界点,具体含义是将下四分位线往下偏移1.5倍的四分位距.$Q_3+1.5R_{IQ}$表示数据中高异常值的临界点,具体含义是将上四分位数往上偏移1.5倍的四分位数间距.$Q_1-1.5R_{IQ}$和$Q_3+1.5R_{IQ}$为异常值的边界.数据过大或者过小均为异常值点,这些点均需要进行处理,其中含有一些有用的信息,不能直接删除,故本文采用相邻数据的平均值来代替该异常值.箱形图绘制时需要将参数进行归一化处理,并统一量纲以便形成箱形图.归一化处理可以保证所有数据稳定在一定区间内,且不会影响原始数据在箱形图中的相对位置.归一化算式为

$$\left.\begin{aligned} x^* &= \frac{x_i-\bar{x}}{\sigma},\\ \bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i,\\ \sigma &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}. \end{aligned}\right\} \tag{1}$$

式中:x^*为标准化后参数的值;x_i为初始值;$\bar{x}$为平均值;σ为标准差;n为样本总数.

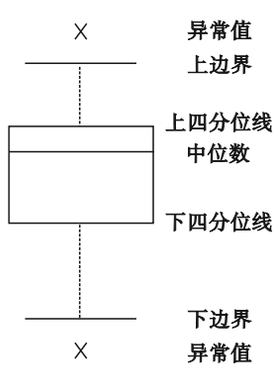

图1　箱形图

Fig. 1　Box diagram

对数据进行预处理,使其在后续进行点预测和区间预测时可以得到可靠的结果.本文共采集到1.3万余条数据,且箱形图横轴数据的顺序均遵循表1中参数序号顺序,纵轴已进行归一化处理,异常值和缺失值用该点前后5个正常点的平均值替代.图2为数据处理结果.

1.3　特征变量选择

特征变量选择的好坏可以直接决定模型的预测精度和稳定性,挑选出了54个影响物理指标的特征变量,若将54个变量全部用于建模,会耗费大量时间,且得到的预测结果并不理想,故需要对特征变量进行筛选.目前常用的特征选择方法有Pearson相关系数法、Spearman相关系数法

等,但这些方法要求数据服从正态分布,而烧结数据的特点为非线性且数据量较大,以上方法的适用性不好.本文采用一种可以对非线性数据进行分析的最大信息系数(maximal information coefficient,MIC)算法,该方法是由Reshef等[10]提出的用于衡量变量间相关性的方法,其计算复杂度较低.最大信息系数M_{IC}的计算公式为

$$M_{IC}=\max_{|x||y|\ll B}\frac{\sum_{x,y}P(x,y)\mathrm{lb}\left(\frac{P(x,y)}{P(x)P(y)}\right)}{\mathrm{lb}\left(\min\left(|x|,|y|\right)\right)}. \quad (2)$$

式中:$|x||y|$分别为x(横轴)方向和y(纵轴)方向划分的区间数;$P(x)$是采样点在x方向上的下落概率;$P(y)$是采样点在y方向上的下落概率;$P(x,y)$是x和y之间的联合概率;B是划分网格的最大数.

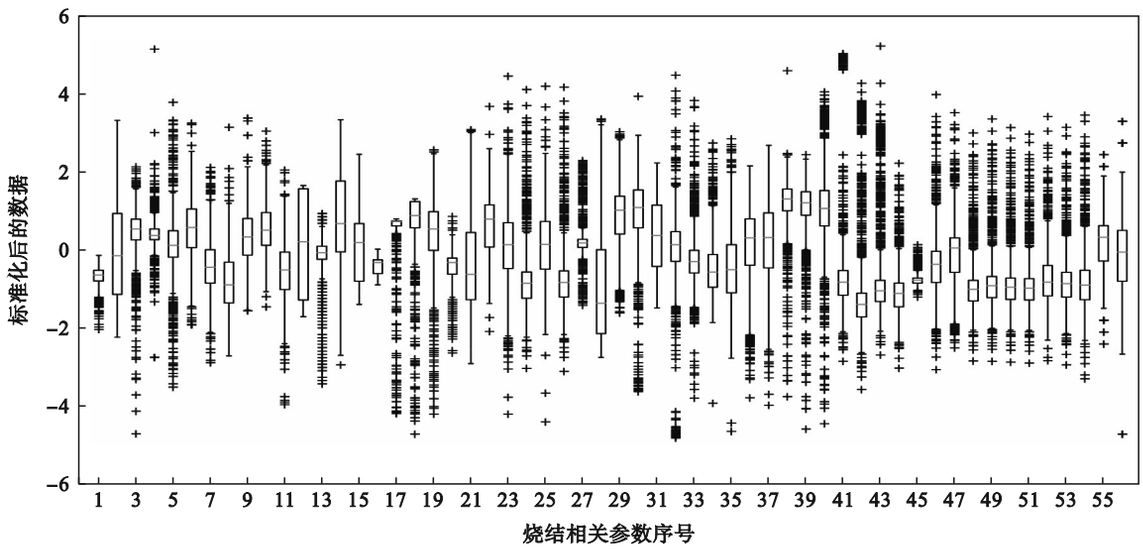

图2 数据处理结果

Fig. 2 Results of data processing

通过经验分析和对M_{IC}的计算,分别确定了20个与水分、粒度指标相关性最强的特征变量,表3列出了M_{IC}计算结果.由表3可见,风机风量、钒钛铁精粉等与水分和粒度这2个物理指标相关性都很强,这些参数都是影响烧结生产的重要因素.当风机风量增加时,烧结机中颗粒运动更加剧烈,促进烧结矿的粒度细化和热风的流动,使得烧结矿中的水分更容易挥发.当添加钒钛铁精粉时,可以改变烧结矿的组成成分,从而影响其烧结特性和物理化学性质.

表3 MIC计算结果

Table 3 Results of MIC

水分				粒度			
工艺参数	M_{IC}	工艺参数	M_{IC}	工艺参数	M_{IC}	工艺参数	M_{IC}
风机风量	0.215 2	1号风箱真空度	0.101 1	风机风量	0.115 9	圆辊速度	0.062 6
11号风箱真空度	0.177 4	5号风箱真空度	0.099 9	钒钛铁精粉	0.105 4	20号风箱废气温度	0.061 9
钒钛铁精粉	0.158 8	高炉返矿	0.095 1	2号风门开度	0.102 8	5号风箱真空度	0.059 8
7号风箱真空度	0.132 0	1号风门开度	0.094 9	自返矿	0.084 1	21号风箱废气温度	0.059 1
钙石灰粉	0.126 3	1号风箱废气温度	0.093 2	1号风门开度	0.083 0	3号风箱废气温度	0.058 6
2号风门开度	0.116 5	烧结用白煤	0.091 6	7号风箱废气温度	0.071 6	除尘矿	0.058 1
20号风箱废气温度	0.113 6	9号风箱真空度	0.087 4	烧结用白煤	0.069 4	22号风箱废气温度	0.056 4
自返矿	0.106 0	9号风箱废气温度	0.086 9	钙石灰粉	0.065 1	1号风箱废气温度	0.055 7
7号风箱废气温度	0.104 9	11号风箱废气温度	0.085 3	13号风箱真空度	0.063 7	南烟道温度	0.055 2
2号风箱真空度	0.101 8	5号风箱废气温度	0.085 2	北烟道温度	0.063 1	高炉返矿	0.055 0

2　物理指标区间预测模型

2.1　RF算法和ET算法

随机森林算法[11]是由Breiman[12]将Bagging集成学习理论和随机子空间方法相结合而提出的一种机器学习算法,该算法通过组合多棵树模型对数据进行训练和预测,可以用于分类和回归任务.其模型流程如图3所示.

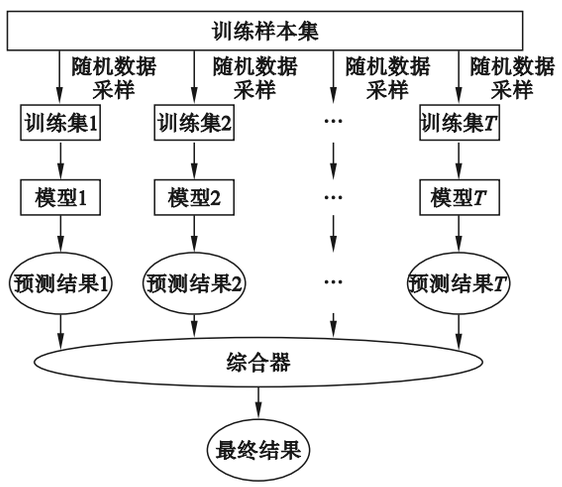

图3　RF算法流程图

Fig. 3　Flowchart of RF algorithm

该算法的工作原理为

1)从训练样本集中随机有放回(即自助采样)的抽取n个样本作为决策树的训练样本;

2)对于每棵决策树,从全部特征中随机选择若干个特征子集,然后从中找到最佳分割点,迭代至指定数目的决策树,进而可以在特征空间的不同子空间上训练各个决策树;

3)通过随机抽取的T个训练集去训练决策树模型,得到多个输出结果,然后将这些结果进行平均操作,得到最终的预测结果.

ET算法是一种基于决策树的集成学习模型,其原理与RF算法类似,均通过多棵决策树组合来达到更好的预测效果,不同之处在于ET模型在节点划分时对每次候选特征的阈值进行随机抽样,随机性更强.ET算法在正常过程中使用的是整个学习样本,并且该算法不需要对特征进行选择,计算复杂度降低了许多,进而也提升了模型的运行效率[13].

2.2　基于Stacking集成学习算法

Stacking是一种集成学习的方法,它通过将不同的学习器组合起来形成一个强大的预测模型,各学习器不仅要有较好的预测精度还要存在一定差异,这样才能使各学习器优势互补发挥最大的预测性能[14].本文选取RF和ET算法叠加对物理指标做点预测.RF算法通过随机抽样和选取特征可以减少过拟合的风险,对异常值和缺失值问题有较好的鲁棒性;ET算法的训练速度较快,且不需要特征选择,可扩展性强.这2种算法都可以对非线性数据进行回归预测,将这2种算法进行集成可以将各自优势互补发挥树模型最大的优势,进而实现对物理指标精准预测的目的.本文选用的模型参数为弱学习器的个数(n_estimators:200),用于控制随机数生成器的参数(random_state:42),其余参数均为默认值.

Stacking算法的学习框架如图4所示.

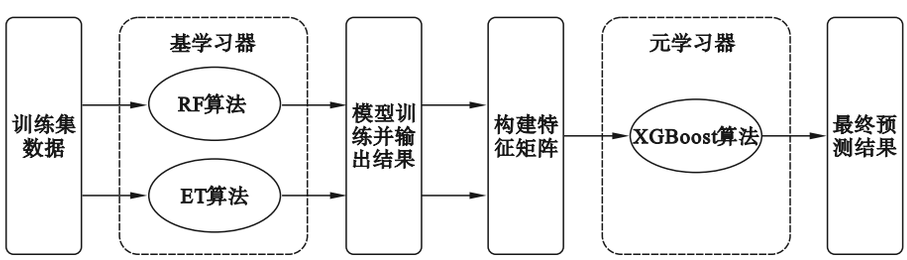

图4　Stacking集成学习框架

Fig. 4　Framework of Stacking integration learning

Stacking算法的工作流程如下:

1)将数据按照一定比例划分出来训练集,并使用RF和ET算法学习器对训练集的数据进行拟合训练.

2)将训练好的基学习器进行预测,并将这些预测结果作为新的特征加入到数据集中.

3)将上步得到的预测结果构建特征矩阵,并将其作为元学习器的输入,使用训练集对元学习器进行训练,本文使用的元学习器为XGBoost算法.

4)使用训练好的元学习器进行预测,然后就可以得到最终预测结果.

2.3　KDE算法

基于烧结工艺大滞后、非线性等特性,点预测得到的结果难以准确表明未来某一时刻的真

实值,为了使操作工人有更优的决策空间,本文采用KDE算法来对烧结过程物理指标作区间预测,该方法是一种非参数的概率密度估计方法,其区间预测结果由数据本身的分布决定,比其他参数估计算法适用性更广泛.KDE算法的工作流程为首先根据预测误差计算出物理指标的概率密度函数(probability density function,PDF),然后计算一定置信水平下的置信区间,由该区间对应的上分位点、下分位点和点预测值可以得到最终的区间预测结果[15].设概率密度函数为$\hat{f}(t)$,则KDE算法的计算公式可以表示为

$$\hat{f}(t)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{t-t_i}{h}\right). \tag{3}$$

式中:h为核函数的带宽;t为预测误差;t_i为第i个误差样本;$K(\cdot)$为非负核函数.

KDE算法的参数有核函数和带宽,参数选取的好坏直接决定了区间预测结果的优劣.当前常用的核函数有高斯核函数、均匀核函数、指数核函数等.由于高斯核函数曲线比较平滑,可以应用于众多场合[16].本文采用高斯核函数,

$$K(u)=(2\pi)^{-\frac{1}{2}}\mathrm{e}^{-\frac{u^2}{2}}. \tag{4}$$

式中,u为自变量.

KDE算法核函数的带宽h的选择更加重要,它会直接影响到曲线的拟合程度,h的选择取决于函数的平滑度和精度,本文采用经验公式来求取h,

$$h=1.06\varepsilon n^{-\frac{1}{5}}. \tag{5}$$

式中,ε是由$\min[\sigma,0.75R_{\mathrm{IQ}}]$得到的.

通过式(3)可以求出物理指标对应的PDF,在一定置信水平$(1-\alpha)$下,可以分别求出区间上分位点$F_{\frac{\alpha}{2}}$和下分位点$F_{\left(1-\frac{\alpha}{2}\right)}$,进而可以得到物理指标的区间预测结果,$\tau=[U_\alpha,L_\alpha]$,$U_\alpha$和$L_\alpha$分别代表区间估计结果的上界和下界,$\alpha$为犯第1类错误的概率,如式(6)所示.

$$\left.\begin{aligned}U_\alpha&=\tilde{p}+F_{\frac{\alpha}{2}},\\L_\alpha&=\tilde{p}+F_{\left(1-\frac{\alpha}{2}\right)}.\end{aligned}\right\} \tag{6}$$

式中,$\tilde{p}$为物理指标的预测值.

2.4 RF-ET-KDE区间预测模型

首先用RF-ET集成模型对物理指标进行点预测,进而使用KDE算法作区间预测,图5为区间预测模型流程图.其基本步骤如下:

1)从某烧结工厂3号烧结机的Oracle和SQL Server数据库中提取出本文所需的全部生产数据,并将其进行初步整理;

2)用箱形图对数据异常值进行预处理,然后用MIC算法计算出与物理指标相关性最大的特征变量,由此可以得到用于建模的物理指标数据集;

3)将物理指标数据集按照数据数量为8∶2随机分为训练集和测试集,然后送至基学习器和元学习器中,基于Stacking集成学习算法进行训练并输出点预测结果;

4)用点预测结果与测试集进行差分可以得到计算误差序列,将其送至KDE算法中求取预测误差的PDF,可以计算出一定置信水平下的区间上分位点和下分位点,从而获得相应预测误差的置信区间值;

5)由物理指标的点预测结果和第4)步所求得的预测误差的置信区间值可以最终求得物理指标的区间预测结果.

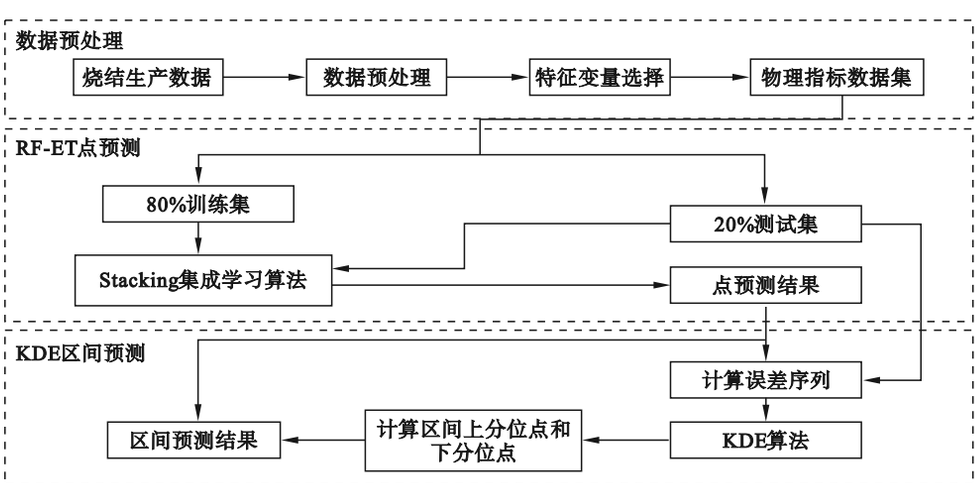

图5 区间预测模型流程图

Fig. 5 Flowchart of interval prediction model

2.5　模型评估指标

模型预测能力的好坏需要通过评估指标来衡量,使用点预测和区间预测评估指标分别对烧结过程物理指标的RF-ET点预测模型和KDE区间预测模型进行评价.

选取的点预测评估指标为平均绝对误差(M_{AE})、均方误差(M_{SE})和均方根误差(R_{MSE}),这3个指标的值越小,说明模型的预测效果越好[17].指标定义式为

$$M_{AE}=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\hat{y}_i\right|, \tag{7}$$

$$M_{SE}=\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i-y_i\right)^2, \tag{8}$$

$$R_{MSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}. \tag{9}$$

式中:$\hat{y}_i$为第i个样本的预测值;y_i为第i个样本的实际值.

本文选取的区间预测评估指标为预测区间覆盖率(P_{ICP})、预测区间归一化平均宽度(W_{PINA})和F值(F)[18].

P_{ICP}表示的是物理指标实际值落在预测区间内的百分比,该指标应当和KDE对应的置信水平相对应.P_{ICP}的定义如下:

$$P_{ICP}=\frac{1}{n}\sum_{i=1}^{n}c_i. \tag{10}$$

式中:当$y_i\in[L_i,U_i]$时,$c_i=1$,否则$c_i=0$.L_i和U_i分别代表第i个区间的下界和上界.

W_{PINA}是用来衡量预测区间的平均宽度,该指标越小,说明区间预测结果越理想,定义如下:

$$W_{PINA}=\frac{1}{nR}\sum_{i=1}^{n}\left(U_i-L_i\right). \tag{11}$$

式中:R是物理指标最大与最小值的差值.

区间预测结果的理想状态应该是有对真实值较高的覆盖率和较小的区间宽度,显然仅用上面2个相互冲突的指标不能满足上述要求,P_{ICP}值越高,宽度会越大,故本文使用一种以覆盖率和宽度为基准的综合评价指标F值来评估区间预测结果的优劣,该值越大,说明区间预测的效果越好[19].本文采用F来最终评价区间预测的质量,定义为

$$F=\frac{2\times P_{ICP}\times\dfrac{1}{W_{PINA}}}{P_{ICP}+\dfrac{1}{W_{PINA}}}. \tag{12}$$

3　结果与分析

3.1　点预测结果分析与对比

在点预测阶段,使用基于Stacking集成的RF-ET算法对烧结过程物理指标进行点预测,由此可以避免单一模型预测精度不足的缺点,且可以将树模型的预测能力发挥到极致.图6和图7分别为粒度和水分叠加模型的预测结果.本文使用的软件为python3.7, scikitlearn1.0.2,硬件为AMD Ryzen 75800H with Radeon Graphics, GeForce RTX 3060 Laptop和16.0 GB RAM.

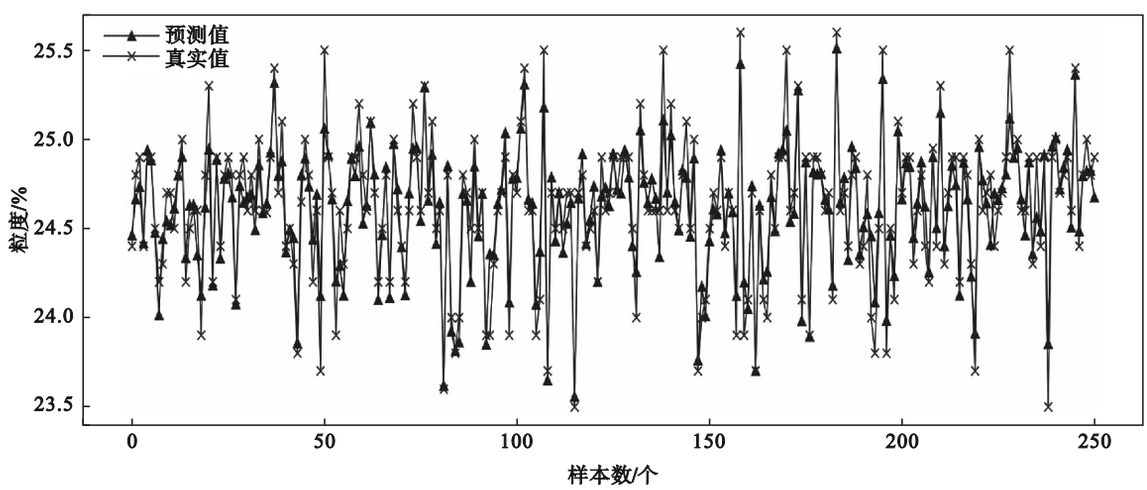

图6　粒度叠加模型预测结果

Fig. 6　Prediction results of particle size Stacking model

为了更好展现基于Stacking集成学习模型的预测优势,将本文所建模型与未集成的光梯度提升机(light gradient boosting machine, LGBM)算法、RF算法和ET算法进行对比,采用2.5节中的点预测指标对各个模型进行性能评估,表4和表5分别为粒度和水分的模型对比结果.由表4和

表5可知,基于Stacking叠加的RF-ET算法预测误差结果均大幅小于其他模型,本文所建模型有较好的点预测能力.

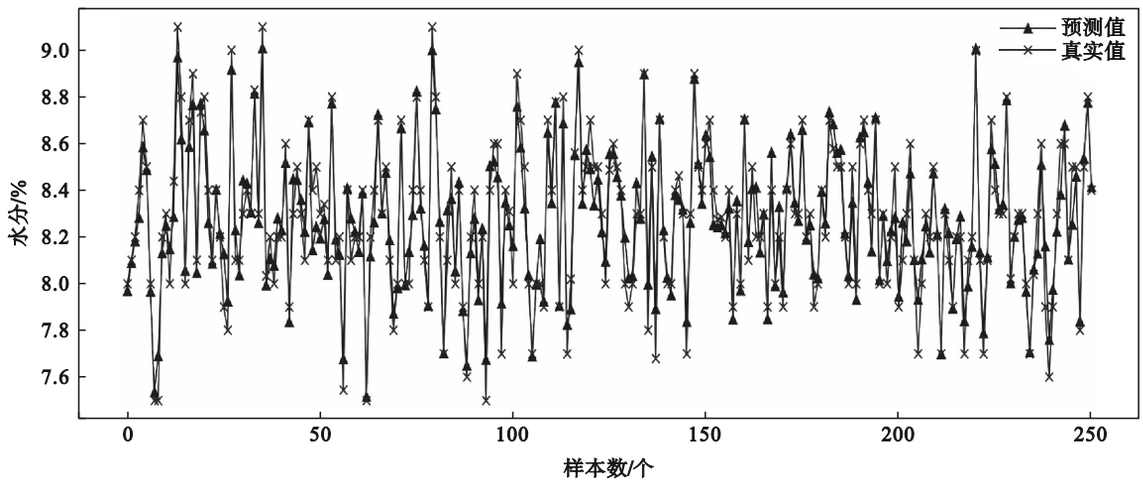

图7 水分叠加模型预测结果

Fig. 7 Prediction results of moisture Stacking model

表4 粒度模型对比结果

Table 4 Comparison results of particle size model

算法	M_{AE}	M_{SE}	R_{MSE}
LGBM	0.233 02	0.094 58	0.307 54
RF	0.222 21	0.090 15	0.300 25
ET	0.198 21	0.072 47	0.269 21
RF-ET	0.102 12	0.019 82	0.140 79

表5 水分模型对比结果

Table 5 Comparison results of moisture model

算法	M_{AE}	M_{SE}	R_{MSE}
LGBM	0.154 35	0.040 07	0.200 19
RF	0.144 80	0.038 52	0.196 27
ET	0.136 11	0.034 01	0.184 41
RF-ET	0.078 52	0.011 23	0.105 95

LGBM,RF和ET算法均为单一算法,LGBM算法对超参数比较敏感,该算法容易出现过拟合问题;RF算法的计算时间和空间开销较大,复杂度较高;ET算法对输入数据尺度和范围比较敏感,且随机化权重容易导致过拟合.将RF-ET算法用Stacking集成学习算法叠加后可以充分利用2个树模型的优点,达到优势互补的效果.

3.2 区间预测结果分析

集成学习算法训练可以得到点预测结果,将其与测试集数据进行差分可以得到物理指标的误差序列,然后采用KDE算法可以计算出对应的PDF,为了有效地看到PDF的分布情况,本文采用频率直方图来展示其拟合效果,频率直方图显示出数据的分布情况,进而分析数据的分布特征和规律.为了突出KDE算法的优势,将KDE与正态分布密度估计(normal density estimation,NDE)共同在频率直方图上进行对比,水分和粒度的PDF对比结果如图8所示,从图中可以看出KDE相较于NDE有更好的拟合效果,用KDE算法可以得到更加可靠的区间估计结果.

用KDE算法求出物理指标的PDF后,选取一定的置信度,可以求出区间上分位点和下分位点,从而得到物理指标区间预测结果.本文选取95%,90%和85%共3个置信水平观测物理指标的预测波动范围,粒度和水分的区间预测结果分别如图9和图10所示.由图可知,随着置信水平的递减,区间预测宽度逐渐减小,覆盖率随之降低,有某些实际值落在区间外,区间预测结果的可靠性降低,在95%置信水平时,区间覆盖效果最好,基本可以覆盖全部值.

为了展现出本文所建模型对烧结过程物理指标的区间预测能力,利用区间预测评估指标对3.1节中的所有模型进行评价,粒度的区间预测对比结果如表6所示,水分的区间预测对比结果如表7所示.由表6和表7可知,本文所建模型的预测效果均好于其他组合模型,其综合评价指标F值最好;该模型在90%置信水平时对粒度的区间预测能力最强,其F值为1.497 02;在95%置信水平时对水分的区间预测能力最强,其F值为1.563 78.由此可见RF-ET-KDE模型有强大的区间预测能力.

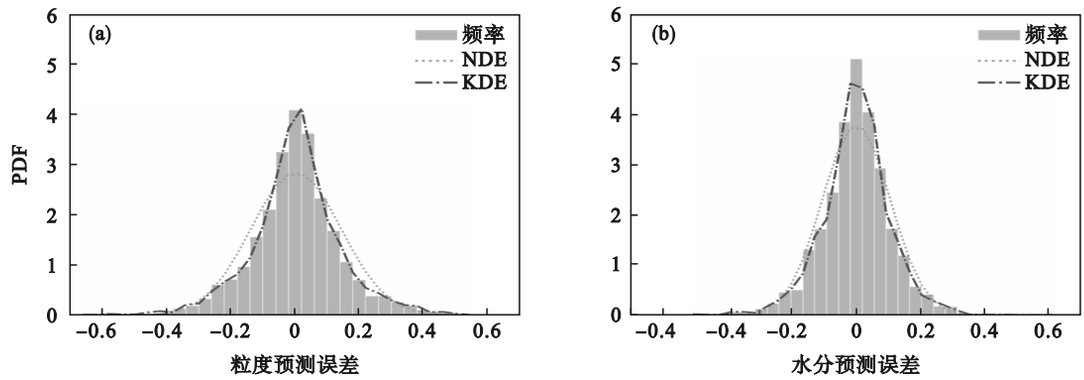

图8　粒度和水分PDF对比

Fig. 8　PDF comparison of particle size and moisture

(a)—粒度PDF;(b)—水分PDF.

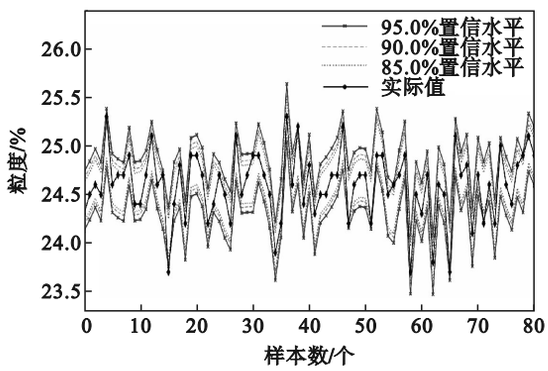

图9　粒度区间预测结果

Fig. 9　Interval prediction results of particle size

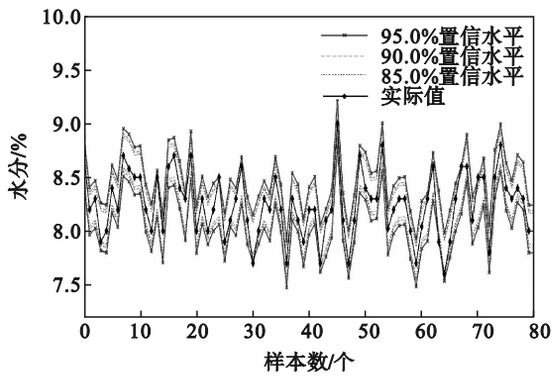

图10　水分区间预测结果

Fig. 10　Interval prediction results of moisture

表6　粒度区间预测对比结果

Table 6　Interval prediction comparison results of particle size

模型	置信水平	P_{ICP}	W_{PINA}	F
LGBM-KDE	95%	0.950 66	0.602 58	1.208 84
	90%	0.900 57	0.490 30	1.249 45
	85%	0.851 23	0.414 60	1.258 36
RF-KDE	95%	0.950 28	0.611 95	1.201 73
	90%	0.900 19	0.487 00	1.251 66
	85%	0.850 47	0.397 54	1.271 17
ET-KDE	95%	0.950 28	0.580 73	1.224 70
	90%	0.899 43	0.453 16	1.277 98
	85%	0.850 47	0.376 63	1.288 29
RF-ET-KDE	95%	0.951 80	0.288 44	1.493 56
	90%	0.902 85	0.228 38	1.497 02
	85%	0.851 23	0.188 96	1.466 57

表7　水分区间预测对比结果

Table 7　Interval prediction comparison results of moisture

模型	置信水平	P_{ICP}	W_{PINA}	F
LGBM-KDE	95%	0.951 42	0.503 75	1.286 33
	90%	0.900 57	0.414 08	1.311 91
	85%	0.850 47	0.351 91	1.309 14
RF-KDE	95%	0.950 66	0.494 05	1.293 70
	90%	0.899 43	0.401 61	1.321 51
	85%	0.850 09	0.340 49	1.318 54
ET-KDE	95%	0.950 28	0.484 68	1.301 24
	90%	0.900 19	0.380 80	1.340 77
	85%	0.85123	0.319 98	1.338 02
RF-ET-KDE	95%	0.952 56	0.229 15	1.563 78
	90%	0.903 23	0.182 32	1.551 03
	85%	0.858 82	0.153 17	1.517 96

由于区间预测结果提供了预测不确定性的度量,且给出了区间置信度、区间宽度等辅助信息,操作人员在进行决策时可以依据区间预测结果来评估风险和不确定性,进而更好地控制烧结过程,提高烧结产品的质量和生产效率.

4　结　　语

针对当前对烧结过程物理指标预测精度不足且没有考虑区间预测的问题,采用基于

Stacking集成的RF-ET-KDE组合算法对物理指标进行点预测和区间预测.RF-ET算法弥补了单一模型预测精度不足的缺点,充分发挥树模型的预测能力,KDE算法对预测误差进行量化,进而可以得到有效可靠的区间预测结果.通过与单一模型对比可得,本文所建算法的点预测能力最强,区间预测效果最好,其综合指标F值高于其他组合模型,该算法在置信水平90%时对粒度有较好的区间预测效果,在95%时对水分有较好的区间预测效果.综上,本文提出的集成学习算法区间预测模型为烧结矿的生产过程提供了更好的评判方法,为烧结操作人员提供了更大的决策空间.可以通过点预测和区间预测结果来推断物理指标的走势和波动范围,进而调整进料量或相关操作参数,生产出高质量的烧结矿.

参考文献:

[1] Li Y F, Zhang Q W, Zhu Y, et al. A model study on raw material chemical composition to predict sinter quality based on GA-RNN [J]. *Computational Intelligence and Neuroscience*, 2022, 2022: 3343427.

[2] Liu S, Lyu Q, Liu X J, et al. Synthetically predicting the quality index of sinter using machine learning model [J]. *Ironmaking & Steelmaking*, 2020, 47(7): 828-836.

[3] Xia G L, Wu Z X, Liu M Y, et al. Prediction interval estimation of sinter drum index based on light gradient boosting machine and kernel density estimation [J]. *Ironmaking & Steelmaking*, 2023, 50(8): 909-920.

[4] Liu S, Liu X J, Lyu Q, et al. Comprehensive system based on a DNN and LSTM for predicting sinter composition [J]. *Applied Soft Computing*, 2020, 95: 106574.

[5] Yang C, Yang C J, Li J F, et al. Forecasting of iron ore sintering quality index: a latent variable method with deep inner structure [J]. *Computers in Industry*, 2022, 141: 103713.

[6] Jiang Y S, Yang N, Yao Q Q, et al. Real-time moisture control in sintering process using offline-online NARX neural networks [J]. *Neurocomputing*, 2020, 396: 209-215.

[7] 刘月明,刘小杰,吕庆,等.基于烧结大数据预测小于10 mm烧结矿含量模型[J].中国冶金,2019,29(11):31-38.
(Liu Yue-ming, Liu Xiao-jie, Lyu Qing, et al. Prediction model of sinter content less than 10 mm based on sintering big data [J]. *China Metallurgy*, 2019, 29(11): 31-38.)

[8] Ren Y Q, Huang C Q, Jiang Y S, et al. Neural network prediction model for sinter mixture water content based on KPCA-GA optimization [J]. *Metals*, 2022, 12(8): 1287.

[9] Li D C, Huang W T, Chen C C, et al. Employing box plots to build high-dimensional manufacturing models for new products in TFT-LCD plants [J]. *Neurocomputing*, 2014, 142(sup1): 73-85.

[10] Reshef D N, Reshef Y A, Finucane H K, et al. Detecting novel associations in large data sets [J]. *Science*, 2011, 334(6062): 1518-1524.

[11] 丁敬国,郭锦华.基于主成分分析协同随机森林算法的热连轧带钢宽度预测[J].东北大学学报(自然科学版),2021,42(9):1268-1274,1289.
(Ding Jing-guo, Guo Jin-hua. Prediction of rough rolling width based on principal component analysis collaborated with random forest algorithm [J]. *Journal of Northeastern University (Natural Science)*, 2021, 42(9): 1268-1274, 1289.)

[12] Breiman L. Bagging predictors [J]. *Machine Learning*, 1996, 24(2): 123-140.

[13] 朱子龙,张立臣.基于堆叠极限树集成算法的信息物理系统入侵检测方法[J].计算机应用与软件,2021,38(11):314-321.
(Zhu Zi-long, Zhang Li-chen. Intrusion detection method of cyber-physical system based on stacking extra tree integration algorithm [J]. *Computer Applications and Software*, 2021, 38(11): 314-321.)

[14] 李泉伦,陈争光,焦峰.基于Stacking集成学习的近红外光谱油页岩含油率预测[J].光谱学与光谱分析,2023,43(4):1030-1036.
(Li Quan-lun, Chen Zheng-guang, Jiao Feng. Prediction of oil content in oil shale by near-infrared spectroscopy based on stacking ensemble learning [J]. *Spectroscopy and Spectral Analysis*, 2023, 43(4): 1030-1036.)

[15] Zhang L, Lu S Y, Ding Y F, et al. Probability prediction of short-term user-level load based on random forest and kernel density estimation [J]. *Energy Reports*, 2022, 8(sup5): 1130-1138.

[16] Zhou B W, Ma X J, Luo Y H, et al. Wind power prediction based on LSTM networks and nonparametric kernel density estimation [J]. *IEEE Access*, 2019, 7: 165279-165292.

[17] Al Fuhaid A F, Alanazi H. Prediction of chloride diffusion coefficient in concrete modified with supplementary cementitious materials using machine learning algorithms [J]. *Materials*, 2023, 16(3): 1277.

[18] Peng G Z, Cheng Y L, Wang H W, et al. Industrial IoT-enabled prediction interval estimation of mechanical performances for hot-rolling steel [J]. *IEEE Transactions on Instrumentation and Measurement*, 2022, 71: 3508010.

[19] Du B G, Huang S, Guo J, et al. Interval forecasting for urban water demand using PSO optimized KDE distribution and LSTM neural networks [J]. *Applied Soft Computing*, 2022, 122: 108875.