·应用与服役·

基于MI和XGBoost算法电渣重熔终点磷含量预报模型

刘玉潇[1]，董艳伍[1,2,3]，姜周华[1,2,3]，陈　玺[1]
（1 东北大学冶金学院，沈阳 110819；2 东北大学轧制及自动化国家重点实验室，沈阳 110819；
3 东北大学多金属共生矿生态化冶金教育部重点实验室，沈阳 110819）

摘　要：研究针对电渣重熔流程提出了一种基于互信息法（MI）和XGBoost的电渣重熔终点磷含量预报模型，利用互信息法对影响终点磷含量的因素进行特征选择与特征评估，特征选择后的数据集作为模型的输入变量。建立MI-XGBoost模型对生产数据进行训练及验证，利用网格搜索交叉验证对模型结构调整和超参数优化，并与MI-RR、MI-RF、MI-GBDT和MI-KNN模型进行横向对比，结果表明，MI-XGBoost模型具有最高的预测精度，MI和GridSearchCV的加入提高了模型预测性能和拟合能力。通过对于测试集的验证，MI-XGBoost模型的R^2、平均绝对误差、解释方差分数和最大误差的数值分别为0.889 4、0.000 4、0.897 2和0.004 1，均优于MI-RR、MI-RF、MI-GBDT和MI-KNN模型。MI-XGBoost模型实现了终点磷含量的有效预测，为电渣重熔流程终点控制和判断提供了很好的参考，为实现电渣重熔过程智能化提供了一个新思路。

关键词：电渣重熔；互信息法；XGBoost算法；磷含量；机器学习
DOI:10.20057/j.1003-8620.2024-00096　　**中图分类号：**TF142

Prediction Model of Phosphorus Content at the End Point of Electroslag Remelting Based on MI and XGBoost Algorithms

Liu Yuxiao[1], Dong Yanwu[1,2,3], Jiang Zhouhua[1,2,3], Chen Xi[1]
(1 School of Metallurgy, Northeastern University, Shenyang 110819, China; 2 State Key Laboratory of Rolling and Automation, Northeastern University, Shenyang 110819, China; 3 Key Laboratory of Ecological Metallurgy of Multimetallic Mineral, Northeastern University, Minist Educ, Shenyang 110819, China)

Abstract: This study proposes a phosphorus content prediction model for the endpoint of electroslag remelting (ESR) refining process based on Mutual Information (MI) method and XGBoost. The MI method is utilized for feature selection and assessment of factors affecting the endpoint phosphorus content. The dataset after feature selection serves as the input variables for the model. The MI-XGBoost model is trained and validated using production data. Grid search cross-validation is employed for model structure adjustment and hyperparameter optimization. It is compared horizontally with MI-RR, MI-RF, MI-GBDT, and MI-KNN models. The results demonstrate that the MI-XGBoost model exhibits the highest prediction accuracy. The incorporation of MI and GridSearchCV enhances the model's predictive performance and fitting ability. Validation of the test set shows that the MI-XGBoost model achieves R^2, Mean Absolute Error, Explained Variance Score, and Maximum Error values of 0.889 4, 0.000 4, 0.897 2, and 0.004 1, respectively, all superior to MI-RR, MI-RF, MI-GBDT, and MI-KNN models. The MI-XGBoost model effectively predicts the endpoint phosphorus content, providing valuable reference for endpoint control and determination in the ESR refining process. It presents a new perspective for realizing the intelligence of the ESR refining process.

Key Words: Electroslag Remelting; Mutual Information Method; XGBoost Algorithm; Phosphorus Content; Machine Learning

电渣重熔技术作为冶炼优质钢锭的一种手段，以其优良的冶金反应条件及特殊的熔炼结晶方式有着其它生产工艺所不能替代的优越性[1]。电渣重熔既可以改善金属纯净度又可以对钢锭的凝固组织进行有效控制，是一种可以将精炼和铸造这两个过程结合起来同时完成的工艺[2]。近年来有众多学者研究机器学习算法与冶金领域中的应用。刘晓航等[3]建立水模型，利用BP神经网络算法拟合实验数据生成模型，对精炼过程中渣眼的演化行为进行预测。Liu等[4]提出了一种基于XGBoost的氢含量

基金项目：国家自然科学基金(No. 52174303)、国家自然科学基金(No. 51874084)、中央高校基本科研业务费(No. 2125026)
作者简介：刘玉潇（1998—），男，博士；　**E-mail:**liu2428332@163.com；　**收稿日期：**2024-04-17
通信作者：董艳伍（1978—），男，博士生导师，教授；　**E-mail:**dongyw@smm.neu.edu.cn

预测模型。在电渣重熔流程中对于磷元素的控制是十分必要的。在钢中磷易偏析，磷含量过高会对钢材的性能产生负面影响，将影响钢的韧性、延展性等力学性能。目前，对于电渣重熔终点磷含量的预测一般有两种方法，第一种是使用基于历史生产数据的简单经验公式。简单经验公式的准确率很低，主要还是依靠人工经验。另一种方法是使用fluent、procast等数值模拟软件，依据冶金机理对终点磷含量进行预测。该方法计算过程耗时过长，难以满足快速炼钢的需求。由于电渣过程内部存在复杂的多元多相耦合反应且过程变量类型混杂，维数高，规模大，具有多变量、强耦合、非线性和大滞后的特点。工艺过程中的各流程都是“黑箱”作业，且电渣炉的数量相对较少，导致生产数据比较匮乏。目前尚未查阅到有关于将机器学习与电渣重熔流程结合预测终点磷含量的相关文献或报道，将机器学习应用于电渣重熔流程可以很好的对上述问题进行解决。

综上所述，研究针对电渣重熔流程提出了一种基于互信息法（MI）和XGBoost的电渣重熔终点磷含量预报模型，使用python作为编程语言，pycharm作为建模平台，用从现场实地采集的电渣重熔流程数据，运用pandas库、scipy库和numpy库进行数据分析与数据清洗工作，采用互信息法度量影响终点磷含量的因素特征与目标变量之间的非线性相关性，选择对目标变量贡献最大的特征子集，将电极尺寸、渣量、锭重、氟化钙、氧化铝、氧化钙、氧化镁、硅铁粉、铝粉、碳粉、结晶器尺寸、熔炼时间、w[Si]、w[Mn]、w[S]、w[Cr]、w[Ni]、w[Cu]、w[Mo]、w[Ti]、w[Al]作为改进预测模型的输入变量。建立改进预测模型对钢厂实地取得的数据进行训练及验证，并利用网格搜索交叉验证对改进模型进行超参数优化，后与MI-RR、MI-RF、MI-GBDT和MI-KNN模型进行横向对比，结果表明，MI-XGBoost模型具有最高的预测精度，MI的加入提高了模型回归预测性能。通过对于测试集的验证，MI-XGBoost模型的R^2、平均绝对误差、解释方差分数和最大误差的数值分别为0.889 4、0.000 4、0.897 2和0.004 1，均优于MI-RR、MI-RF、MI-GBDT和MI-KNN模型。MI-XGBoost模型实现了终点磷含量的有效预测，为电渣重熔流程终点控制和判断提供了很好的参考，为实现电渣过程智能化提供了一个新思路。

1 互信息法和机器学习算法

1.1 互信息法

互信息法（Mutual Information method）是一种特征选择和特征评估方法，通过衡量特征与目标变量之间的关联性来选择最相关的特征[5-7]。对于多变量应用场景而言，由于多变量之间的相互依赖程度不同，数据集中包含大量的特征会导致特征空间的维度很高，不进行特征选择和特征评估会导致问题分析时相对复杂且增加运算时间，还容易使模型过拟合，从而降低模型的泛化能力。互信息法通过计算特征和目标变量之间的互信息值来评估它们之间的相关性。通过计算特征与目标变量之间的互信息值，可以衡量特征对目标变量的预测能力，从而筛选出最具有代表性和预测能力的特征，剔除那些对目标变量影响较小的特征。这有助于提高模型的预测性能，并且可以减少特征空间的维度，降低模型复杂度。该方法的原理和具体流程如下：

假设存在两个变量X、Y，在信息论中，熵是对不确定性的测量，用H表示。变量X的熵和变量Y的熵如式（1），式（2）所示。

$$H(X)=-\sum_X p(X)\log p(X) \tag{1}$$

$$H(Y)=-\sum_X p(Y)\log p(Y) \tag{2}$$

式中，$p(X)$为变量X的边缘概率分布函数，$p(Y)$为变量Y的边缘概率分布函数。

根据熵的连锁规则，联合熵$H(X,Y)$如式（3）所示。

$$H(X,Y)=H(X)+H(Y|X) \tag{3}$$

式中，$H(Y|X)$为条件熵。

互信息$I(X;Y)$是联合分布$p(X,Y)$与边缘分布$p(X)$，$p(Y)$的相对熵，如式（4）所示。

$$I(X;Y)=\sum_{X,Y} p(X,Y)\log\frac{p(X,Y)}{p(X)p(Y)} \tag{4}$$

1.2 XGBOOST算法

XGBoost（Xtreme Gradient Boosting）基本算法是GBDT算法模式，属Boosting迭代类、树类方法的范畴[8-10]。XGBoost算法思路是通过特征分析来生长一棵树，并持续地增加一棵树，而每增加一种树，实际上就要去拟合上次估计的残差并获取新函数，通过逐次迭代来改善模型特性。如果训练完成得到了K棵树叶，就需要预估下某个样品的得分了。它将会根据这个样品的特点，从每种树叶中都会落到

相应的一些树叶节点,而每种树叶节点也就相应着一种得分,所以最终结果只需把每种树叶相应的得分加起来,便是该样本的预测值[11-12]。

XGBoost算法采取了分步的前向加法模式,是在一次迭代中产生弱学习器后不再要求重新计算某个关系,模型格式如式(5)所示。

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i), f_k \subset F \tag{5}$$

式中,K表示树的数目,f_k表示函数空间F中的一个函数,代表树这种抽象结构。F表示的即为最终预测结果,$\widehat{y_i}$是样本x_i的预测值,然后对目标函数$obj(\theta)$进行定义,如式(6)所示。

$$obj(\theta) = \sum_{i}^{n} l\left(y_i, \widehat{y_i}\right) + \sum_{k=1}^{K} \Omega\left(f_k\right) \tag{6}$$

式中,l为实验的损失函数,Ω为惩罚项,y_i是第i个样本,n代表样本数量。它在形式上如式(7)所示.

$$\Omega\left(f_t\right) = \gamma T + \frac{1}{2}\tau \sum_{j=1}^{T} \omega_j^2 \tag{7}$$

式中,T为叶结点数,ω_j为j叶子结点权重,γ和τ为预先设计超参数,f_t是第t次迭代的数学模型。当引入一个正则化项时,计算通常会选取简化且性能优异的模式,损失函数中最右端的正则化项是用于在每个迭代中控制最弱学习器$f_i(x)$过拟合的能力,而不涉及最后一个模块的集合。

$$\widehat{y_i}^{(0)} = 0 \tag{8}$$

$$\widehat{y_i}^{(1)} = f_1(x_i) = \widehat{y_i}^{(0)} + f_1(x_i) \tag{9}$$

$$\widehat{y_i}^{(2)} = f_1(x_i) + f_2(x_i) = \widehat{y_i}^{(1)} + f_2(x_i) \tag{10}$$

…

$$\widehat{y_i}^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \widehat{y_i}^{(t-1)} + f_t(x_i) \tag{11}$$

式中,$\widehat{y_i}^{(t-1)}$为模型第t-1次迭代后样本i的预测值。如式(8)至式(11)所示,XGBoost系统的一次迭代就会形成一个新的决策树,决策树根据与实际值的残差来建立,后就会建立的一个决策树就会预测第一次的结果。至此,可以得到第t棵树模型的预测结果,在数值上等于前面t-1棵树的预测结果,加上第t棵树的表现。那么对于t棵树我们的目标函数如式(12)所示。

$$obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i}^{(t-1)} + f_t(x_i)\right) + \Omega\left(f_t\right) + constant \tag{12}$$

用泰勒公式来近似上述目标,如式(13)至式(15)所示。

$$f(x + \Delta x) \cong f(x) + f'(x)\Delta x + \frac{1}{2} f''(x)\Delta x^2 \tag{13}$$

$$g_i = \partial \widehat{y_i}^{(t-1)} l\left(y_i, \widehat{y_i}^{(t-1)}\right), h_i = \partial^2 \widehat{y_i}^{(t-1)} l\left(y_i, \widehat{y_i}^{(t-1)}\right) \tag{14}$$

$$obj^{(t)} \cong \sum_{i=1}^{n} \left[l\left(y_i, \widehat{y_i}^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega\left(f_t\right) \tag{15}$$

式中,g_i为样本i的一阶导数;h_i表示样本i的二阶导数。在求到模型第t棵树时,前面t-1棵树的结果和模型结构是固定好的,故可以看到目标函数就被转换成了关于g_i和h_i的函数,在对于结果进行运算的过程中,结果的准确与否单单是样本的误差函数决定的,在求最终结果的过程中,回归过程都是相似的,在到达对于模型结构参数的选取值时便结束此过程。

1.3　其他机器学习算法

岭回归(ridge regression)属于线性回归算法,线性回归算法由于其局限性,引入了岭回归的概念,将线性回归标准方程中加入岭系数得到岭回归算法[13]。它通过添加L2惩罚项来完成特征选择和模型复杂度控制,避免过拟合问题。其计算公式如式(16)所示。

$$w = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \vartheta \boldsymbol{I}_n\right)^{-1} \boldsymbol{X}^{\mathrm{T}} y \tag{16}$$

式中,w和$\boldsymbol{X}$表示最小二乘估计和矩阵,ϑ和y为向量和岭系数,$\boldsymbol{I}_n$为单位矩阵。

随机森林(Random Forest)一种基于决策树的学习方法。它通过构建多个决策树且每个决策树的训练样本和特征都是随机选择的,最后使这些分类器进行投票来得到结果[14]。随机森林利用bagging和随机特征选择来减轻过拟合,它集成了多个分类器以提高效果和预测稳定性。

K-近邻算法(KNN)是一种有监督的学习算法,用于对连续值进行预测[15]。KNN的基本思想是对于任意n维输入向量,分别对应于特征空间中的一个点,输出为该特征向量所对应的类别标签或预测值。

梯度提升决策树(GBDT)是一种集成学习算法,基本思想是在每次梯度提升算法的迭代中,每个弱学习器的目标是拟合先前累加模型的损失函数的负梯度,使加上该弱学习器后的累积模型损失往负梯度的方向减少[16]。计算过程如式(17)至式(20)所示。

$$f_0(x) = argmin \sum_{i=1}^{n} L\left(y_i, f(x)\right) \tag{17}$$

$$f_t(x) = f_{t-1}(x) + \Delta f_t(x) \tag{18}$$

$$\Delta f_t(x) = \rho_t h_t \tag{19}$$

$$F(x) = \sum_{t=0}^{T} f_t(x) \tag{20}$$

其中，$f_t(x)$为第 t 次迭代之后的函数，$L(y_i, f(x))$为损失函数，$\Delta f_t(x)$为每次迭代后的 boost 的值，ρ_t为最佳梯度下降步长，h_t为 base-learner 函数，迭代之后，$F(x)$为总 boost 值。

2 电渣重熔磷元素机理分析

电渣重熔流程是将自耗电极插入渣池中，当自耗电极、渣池、金属熔池、钢锭、底水箱通过短网和变压器形成供电回路，在通电过程中，变压器输出将会产生电流流经熔渣，渣池具有一定的电阻，便会放出大量的焦耳热。电极端头熔化并滴落，穿过渣池并逐渐形成金属熔池。通过结晶器的强制水冷效用下，金属铸锭从底部到顶部慢慢凝固[17]。

磷在钢中以磷化铁(Fe_3P、Fe_2P)、磷化镍(Ni_3P)和其他元素的磷化物形式存在。脱磷反应式如式(21)所示。

$$2[P] + 5(FeO) + 4(CaO) = (4CaO \cdot P_2O_5) + 5[Fe] \tag{21}$$

平衡常数如式(22)所示。

$$K = \frac{a_{(4CaO \cdot P_2O_5)} \cdot a_{[Fe]}^5}{a_{[P]}^2 \cdot a_{(FeO)}^5 \cdot a_{(CaO)}^4} \tag{22}$$

磷的传质包括以下步骤：(1)PO_4^{3-}从渣相传质到渣金界面；(2)Fe 与 PO_4^{3-}反应非常迅速，在界面处形成[P]、(FeO)和(O^{2-})；(3)[P]和[O]从界面到金属的传质，以及(O^{2-})从界面到渣相的传质。

对于电渣锭中磷含量的影响因素，有不少学者做了相关研究。Li 等[18]研究了 G20CrNi2Mo 轴承钢电渣重熔(ESR)过程中磷转移的动力学行为，开发了基于薄膜和渗透理论的动力学模型，以阐明磷从金属薄膜到液滴和金属池的变化；Gao 等[19]研究了外磁场和不同电气参数对电渣重熔过程的影响，当外磁场与重熔电流相互作用产生的电磁力较小时，磷偏析得到改善。然而，过大的电磁力加剧了碳和磷的偏析；Zhong 等[20]研究横向磁场对电渣重熔 GCr15 轴承钢的影响，发现横向磁场的存在有利于磷元素的去除，从而显著提高 GCr15 轴承钢的力学性能。

3 电渣重熔终点磷含量预测模型建立

3.1 模型输入变量特征选择

3.1.1 原始数据分析

研究所采用的数据为国内某钢厂 G20Cr2Ni4A 轴承钢的实际生产数据，初始数据集包含 23 个输入变量，单一输出变量，其输入变量分别为：电极尺寸、渣量、锭重、氟化钙、氧化铝、氧化钙、氧化镁、硅铁粉、铝粉、碳粉、结晶器尺寸、熔炼时间及电极成分(w[C]、w[Si]、w[Mn]、w[S]、w[Cr]、w[Ni]、w[Cu]、w[Mo]、w[W]、w[Ti]、w[Al])，其中单一输出变量为终点磷含量。

在数据采集完毕后，将对从现场所采集到的原始样本数据库采用数据清洗操作，在电渣重熔过程的实际生产中，由于人为或环境等因素，可能会导致采集的原始数据的某些炉次中存在缺失值、异常值和重复值，这些数据不能作为样本数据，需要预先进行处理[21]。本研究采用 python 编程语言进行编程，在 pycharm 平台进行建模，在进行数据预处理过程，将采用 pandas 库和 numpy 库这两个数据处理库进行数据分析与预处理工作。初始数据集有 1000 余组数据，由于重复储存、关键数据丢失等问题，选择了 714 组有效生产数据，利用以上数据为基础数据进行本研究建模。

3.1.2 输入变量输出变量相关性分析

在数据清洗之后要进行特征选择，在此模型进行特征选择时使用互信息法对各变量之间进行互信息值计算。使用 pycharm 平台进行编程，由互信息法将互信息值较小的输入变量进行剔除，最终选择电极尺寸、渣量、锭重、氟化钙、氧化铝、氧化钙、氧化镁、硅铁粉、铝粉、碳粉、结晶器尺寸、熔炼时间、w[Si]、w[Mn]、w[S]、w[Cr]、w[Ni]、w[Cu]、w[Mo]、w[Ti]、w[Al]作为输入变量。各特征变量的互信息得分如图 1 所示。

在完成特征选择之后，由于本研究的原始数据输入变量的数值差异比较大，故再对数据结果加以标准化的处理。归一化的处理，式(23)适用于数据

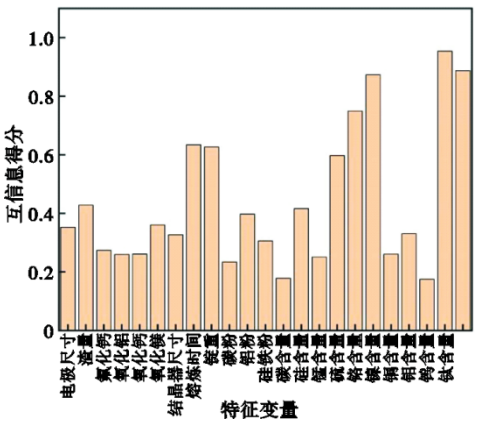

图 1 特征变量互信息得分图

Fig. 1 Mutual information score plot of characteristic variables

归一化。

$$x_{normalized} = \frac{x_{input} - x_{min}}{x_{max} - x_{min}} \tag{23}$$

式中,$x_{normalized}$为归一化数据;x_{input}为输入数据;x_{min}为数据样本中变量的最小值;x_{max}为数据样本中变量的最大值。

3.2　机器学习终点磷含量模型

研究采用岭回归(RR)、随机森林(RF)、XGBoost算法、梯度提升决策树(GBDT)和K-近邻回归(KNN)五种机器学习算法对终点磷含量进行预测,使用Python语言对五种预测模型进行编程,采用未经过特征选择的原始数据集作为数据样本,将数据样本总数进行80%训练集及20%验证集的分隔,然后使用随机选择的数据总集中的训练集利用算法进行模型训练。

在训练完成后,将已建立并训练完成的模型得到的最终输出预测值进行输出,并将随机选取的验证集中的最终预测值与原始数据样本中的最终实际值进行输出并对于最终输出的预测值与原始数据样本中的最终实际值的散点图进行输出,来对已建立并训练完成的模型的预测最终效果进行判断。在预测值输出后,绘制模型误差统计图,此模型的评价标准是平均绝对误差(*MAE*)、R^2、解释方差分数(*EVS*)和最大误差(*ME*)[22]。以上评价指标的公式如式(24)至式(27)所示。

$$R^2 = 1 - \frac{\sum\left(y_i - \widehat{y_i}\right)^2}{\sum\left(y_i - \overline{y_i}\right)^2} \tag{24}$$

$$MAE = \frac{1}{n \cdot \sum\left|y_i - \widehat{y_i}\right|} \tag{25}$$

$$EVS = 1 - \frac{Var\left(y_i - \widehat{y_i}\right)}{Var\left(y_i\right)} \tag{26}$$

$$ME = \max\left(\left|y_i - \widehat{y_i}\right|\right) \tag{27}$$

式中,y_i为实际值;$\widehat{y_i}$为预测值;$\overline{y_i}$为实际值的均值;$Var()$表示方差。

本研究采用岭回归(RR)、随机森林(RF)、XGBoost算法、梯度提升决策树(GBDT)和K-近邻回归(KNN)五种机器学习算法对终点磷含量进行预测,表1为各个模型所采用的超参数。

采用原始数据集对五个模型训练后,各模型预测值与实际值的对比如图2所示。为了能更清晰地进一步对预测效果进行定量判断,将利用验证集中预测值与真实值的差值进行绘制,如图3所示。

通过真实数据对各模型进行训练和验证后,从图3中看到验证集中各炉的预测误差。XGBoost模型最大的误差为4.2×10^{-5},最小的误差为0;RF模型最大的误差为4.9×10^{-5},最小的误差为0;RR模型最大的误差为8.1×10^{-5},最小的误差为1.7×10^{-5};GBDT模型最大的误差为4.9×10^{-5},最小的误差为3×10^{-6};KNN模型最大的误差为6×10^{-5},最小的误差为0。计算各模型的R^2、平均绝对误差(MAE)、解释方差分数(EVS)和最大误差作(ME)为评价指标,见表2。

从表2可以得到,在评价指标为R^2、平均绝对误差、解释方差分数和最大误差中,本节所建立的五个磷含量模型预测能力最优的是XGBoost模型。XGBoost模型与随机森林平均绝对误差模型数值相同,优于岭回归、梯度提升决策树和K-近邻模型。本节五种模型从优到劣排序为XGBoost模型、梯度

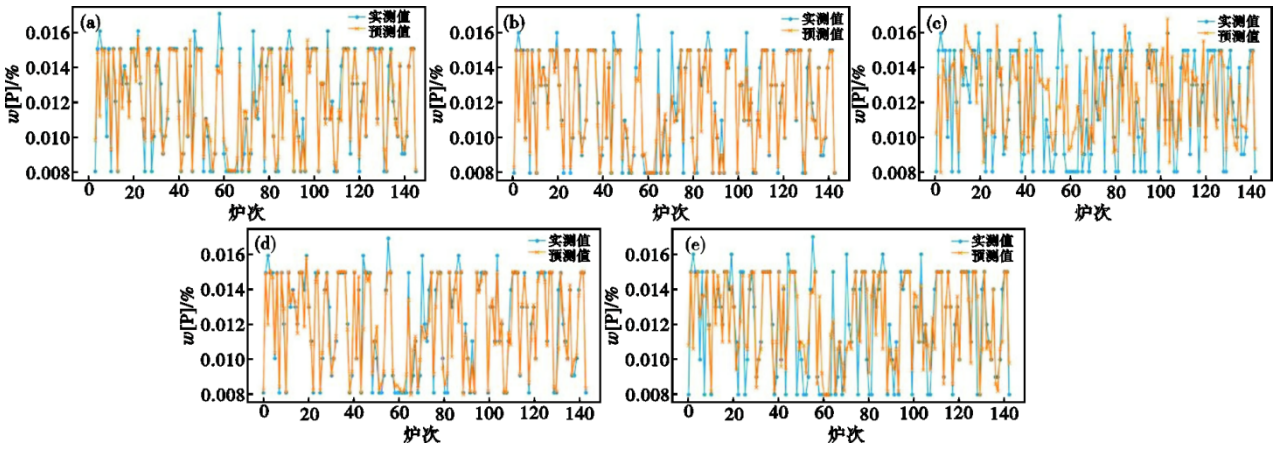

图2　磷含量模型预测效果图:(a)XGBoost,(b)RF,(c)RR,(d)GBDT,(e)KNN

Fig. 2　Effect of phosphorus content model prediction : (a)XGBoost, (b) RF, (c) RR, (d) GBDT, (e) KNN

表 1 不同机器学习模型中的超参数

Table 1 Hyperparameters in different machine learning models

模型	超参数	数值
岭回归模型	Max_iter	None
	solver	lsqr
随机森林模型	n_estimators	100
	max_depth	auto
	min_samples_split	2
XGBoost 模型	base_score	0.5
	max_depth	6
	n_jobs	12
	booster	gbtree
	n_estimators	100
	learning_rate	0.1
	subsample	0.1
	colsample_bytree	0.1
梯度提升决策树模型	learning_rate	0.1
	n_estimators	100
	max_depth	3
	min_samples_split	5
K-近邻回归模型	n_neighbors	5
	weights	uniform
	algorithm	auto

提升决策树模型、随机森林模型、K-近邻模型和岭回归模型，表明 XGBoost 模型对数据的解释能力最强。由表 2 可知，XGBoost 模型的解释方差分数均优于其他模型，表明其对数据方差的解释能力均强于其他模型。

不同机器学习模型的预测性能取决于算法的特性。从表 2 中可知，岭回归模型的预测精度最低。由于岭回归模型属于线性回归模型的拓展，电渣重熔过程影响磷含量的因素较多，故岭回归模型无法

表 2 磷含量模型评价指标对比表

Table 2 Comparison of indicators for evaluation of phosphorus content models

指标	XGBoost	RF	RR	GBDT	KNN
R^2	0.879 6	0.855 9	0.345 6	0.861 7	0.588
MAE	0.000 5	0.000 5	0.001 7	0.000 6	0.000 8
EVS	0.885 4	0.871 1	0.350 1	0.882 4	0.853 2
ME	0.004 2	0.004 9	0.008 1	0.004 9	0.006 0

进行精准的预测。K-近邻算法属于非参数学习算法，计算复杂度高，对异常值敏感，相较于岭回归模型对非线性问题的预测能力提升。从结果可以得到，K-近邻模型预测精度仍低于 XGBoost 模型、梯度提升决策树模型和随机森林模型。XGBoost 模型、梯度提升决策树模型和随机森林模型均属于集成算法，将多个学习器结合，从而提高预测性能。结果分析得到，XGBoost 模型的预测精度高于梯度提升决策树模型和随机森林模型。

3.3 MI 改进模型的建立

MI 改进模型将由 3.2 节所建立的五种模型的基础上进行改进，一方面是将 3.1.3 节互信息法进行特征选择后的数据集作为数据样本，另一方面是利用网格搜索交叉验证（GridSearchCV）对各个算法模型进行超参数优化。

本模型采用岭回归（RR）、随机森林（RF）、XGBoost 算法、梯度提升决策树（GBDT）和 K-近邻回归（KNN）回归算法对磷含量进行预测，采用互信息法进行特征选择后的数据样本进行训练，相较于 3.2 节中的 5 种模型，使用网格搜索交叉验证（GridSearchCV）对算法模型进行超参数优化。通过对超参数的优化，充分考虑模型的准确性和鲁棒性，将

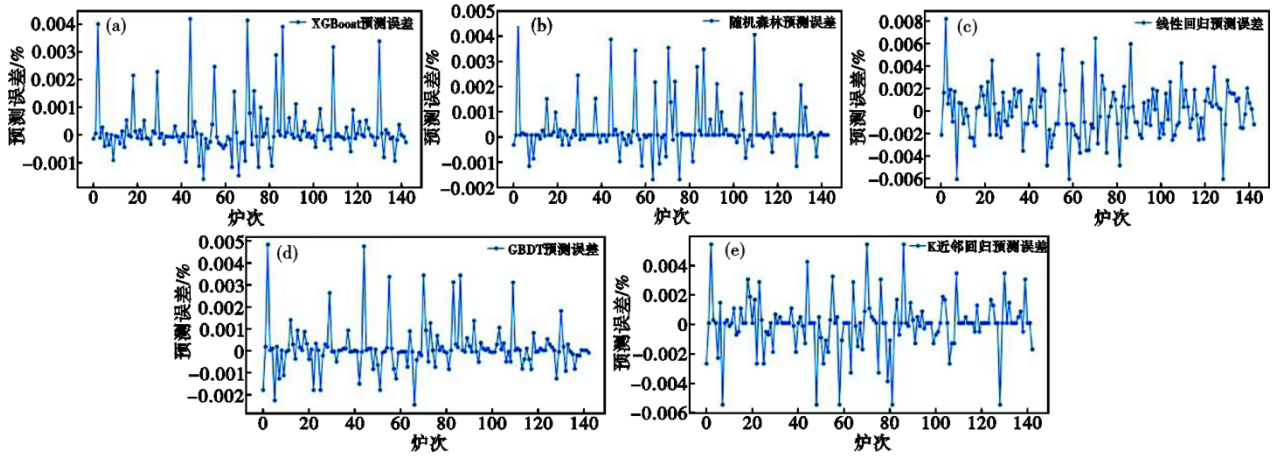

图 3 磷含量模型误差图：(a)XGBoost，(b)RF，(c)RR，(d)GBDT，(e)KNN

Fig. 3 Error diagram of phosphorus content model: (a) XGBoost, (b) RF, (c) RR, (d) GBDT, (e) KNN

R^2、MAE、EVS和ME最优的模型作为机器学习算法的最佳模型。改进后各模型的最优超参数和相关优化范围见表3。

对训练完成的模型进行最终预测,并将预测值与随机选取的验证集中的实际值进行比较。生成预测值与实际值的散点图,评估模型预测效果。MI-RR、MI-RF、MI-XGBoost、MI-GBDT和MI-KNN磷含量预测模型的预测效果如图4所示。

采用互信息法处理后的数据集数据对五个模型训练后,图4表示各模型预测值与实际值的对比,其横轴为随机选取验证集的炉数,竖轴为磷含量的数值分布,为了能更清晰地进一步对预测效果进行定量判断,将利用验证集中预测值与真实值的差值进行绘制,各模型的磷含量模型预测误差统计如图5所示。

从图5可以看出,验证集每炉次预测值与实际值的差值,即预测误差。MI-XGBoost模型最大的误差为4.1×10^{-5},最小的误差为0;MI-RF模型最大的误差为4.9×10^{-5},最小的误差为0;MI-RR模型最大的误差为8.1×10^{-5},最小的误差为1.7×10^{-5};MI-GBDT模型最大的误差为4.3×10^{-5},最小的误差为2×10^{-6};MI-KNN模型最大的误差为6×10^{-5},最小的误差为0。计算各改进模型的R^2、平均绝对误差(MAE)、解释方差分数(EVS)和最大误差(ME)作为评价指标,见表4。

从表4可得,XGBoost模型、随机森林模型、岭回归模型、梯度提升决策树模型和K-近邻模型在结合互信息法与网格搜索交叉验证后,模型性能均有不同幅度的提升。K-近邻模型的R^2提升最为明显,R^2数值从0.588 0提升至0.845 7,梯度提升决策树模型的R^2数值从0.861 7提升至0.882 2,其余模型R^2也均有相应提升,表明改进模型对数据的解释能力变

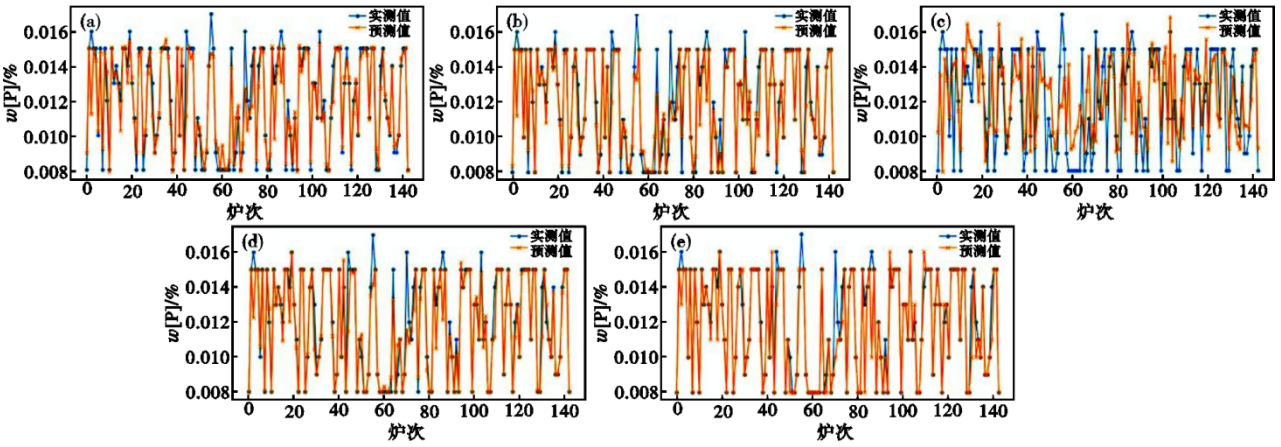

图4　改进磷含量模型预测效果图:(a)MI-XGBoost,(b) MI-RF,(c) MI-RR,(d) MI-GBDT,(e) MI-KNN

Fig. 4　Predictive effectiveness of improved phosphorus content models: (a) MI-XGBoost, (b) MI-RF, (c) MI-RR, (d) MI-GBDT, (e) MI-KNN

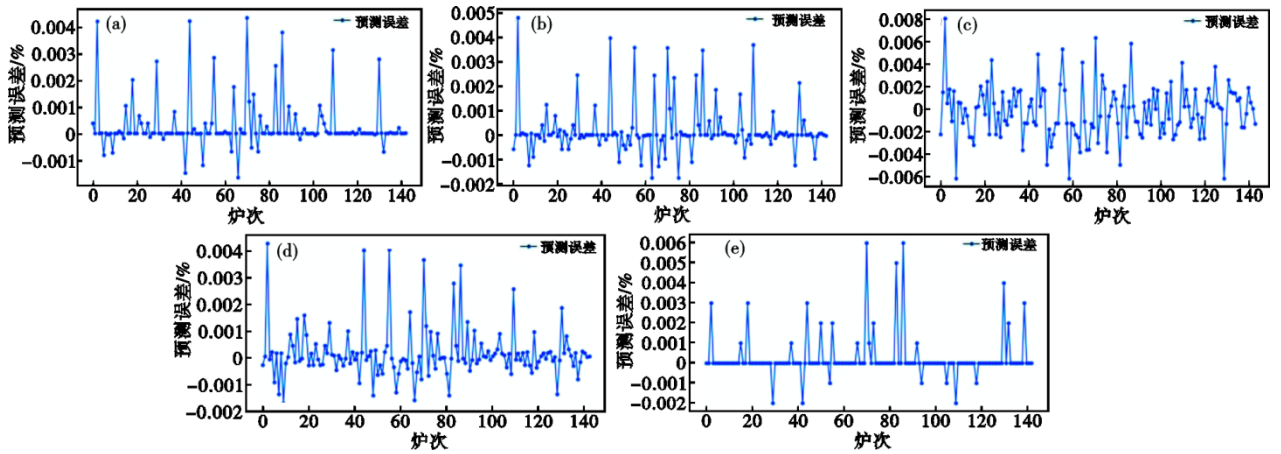

图5　改进磷含量模型预测误差图:(a)MI-XGBoost,(b) MI-RF,(c) MI-RR,(d) MI-GBDT,(e) MI-KNN

Fig. 5　Improving phosphorus content model prediction errors: (a) MI-XGBoost, (b) MI-RF, (c) MI-RR, (d) MI-GBDT, (e) MI-KNN

表 3 改进机器学习算法中的超参数

Table 3 Improving hyperparameters in machine learning algorithms

模型	超参数	寻优范围	最优值
MI-RR	Max_iter	(100,10 000)	1 200
	solver	[svd, lsqr]	lsqr
MI-RF	n_estimators	(100,1 000)	113
	max_depth	(3,10)	9
	min_samples_split	(2,10)	2
MI-XGBoost	base_score	(0,1)	0.5
	max_depth	(3,10)	8
	n_jobs	(1,20)	12
	booster	[gbtree, gblinear]	gbtree
	n_estimators	(100,1 000)	790
	learning_rate	(0.001,0.1)	0.02
	subsample	(0.1,1.0)	0.98
	colsample_bytree	(0.1,1.0)	0.68
MI-GBDT	learning_rate	(0.001,0.1)	0.07
	n_estimators	(100,1 000)	959
	max_depth	(3,10)	3
	min_samples_split	(2,20)	5
MI-KNN	n_neighbors	(1,10)	1
	weights	['uniform', 'distance']	uniform
	algorithm	['ball_tree', 'kd_tree']	kd_tree

表 4 改进磷含量模型评价指标对比图

Table 4 Comparison of evaluation indicators of the improved phosphorus content model

指标	MI-XGBoost	MI-RF	MI-RR	MI-GBDT	MI-KNN
R^2	0.889 4	0.869 8	0.345 7	0.882 2	0.845 7
MAE	0.000 4	0.000 5	0.001 8	0.000 5	0.000 4
EVS	0.897 2	0.874 2	0.350 1	0.892 4	0.853 2
ME	0.004 1	0.004 9	0.008 1	0.004 3	0.006 0

强。K-近邻模型的平均绝对误差提升最为明显,平均绝对误差数值从0.000 8减少至0.000 4,其余模型在优化后也均有不同程度降幅。平均绝对误差数值越小,表示模型对样本的预测越准确。在优化后各模型的解释方差分数和最大误差较3.2节模型同样有不同程度优化。

从表4可得,相比于MI-RR、MI-KNN模型,集成学习算法的模型预测精度较高。将集成模型XGBoost模型、随机森林模型和梯度提升决策树模型在结合互信息法与网格搜索交叉验证后,MI-XGBoost模型的R^2、平均绝对误差、解释方差分数和最大误差均优于MI-RF模型和MI-GBDT。故MI-XGBoost模型为本研究最优模型,互信息法与网格搜索交叉验证的加入提高了模型预测性能和拟合能力,可以为电渣重熔生产提供很好的参考。

4 结论

根据电渣重熔流程的实际控制要求,提出了一种MI-XGBoost的电渣重熔终点磷含量预测模型。首先,将采集到的数据进行数据清洗及数据分析,利用互信息法对影响终点磷含量的因素进行特征选择与特征评估。然后将特征选择后的数据集作为模型的输入变量,终点磷含量为单一输出变量,利用网格搜索交叉验证进行超参数优化,建立了MI-XGBoost的电渣重熔终点磷含量预测模型,从研究中得出结论。

(1)基于现场取得的实验数据作为初始数据集,采用未经过特征选择的原始数据集作为数据样本,分别建立XGBoost模型、梯度提升决策树模型、随机森林模型、K-近邻模型和岭回归模型,仿真实验结果表明,5种模型对个别炉次预测误差较大,虽然XGBoost的预测精度高于其余四个模型,但仍不能满足实际生产需求。

(2)将互信息法与XGBoost模型、梯度提升决策树模型、随机森林模型、K-近邻模型和岭回归模型结合,并利用网格搜索交叉验证对改进模型进行超参数优化,建立MI-RR、MI-RF、MI-XGBoost、MI-GBDT和MI-KNN模型。仿真实验结果表明,在结合互信息法与网格搜索交叉验证后,模型性能均有不同幅度的提升,互信息法与网格搜索交叉验证的加入提高了模型预测性能和拟合能力。MI-XGBoost模型的R^2、平均绝对误差、解释方差分数和最大误差的数值分别为0.889 4、0.000 4、0.897 2和0.004 1,均优于MI-RR、MI-RF、MI-GBDT和MI-KNN模型。因此,MI-XGBoost模型可以实现电渣重熔终点磷含量的有效预测,为电渣重熔流程终点控制和判断提供了很好的参考。

参考文献

[1] Shi C B, Wang S J, Li J, et al. Non-metallic inclusions in electroslag remelting: A review[J]. Journal of Iron and Steel Research International, 2021, 28(12): 1483-1503.

[2] Ahmadi S, Arabi H, Shokuhfar A, et al. Evaluation of the electroslag remelting process in medical grade of 316LC stainless steel [J]. Journal of Materials Science & Technology, 2009, 25(5):

592-596.

[3] 刘晓航, 王胙杰, 刘　畅, 等. 基于BP神经网络的钢包渣眼演化行为的预测[J]. 钢铁研究学报, 2024, 36(4): 469-480.

[4] Liu Y X, Dong Y W, Jiang Z H, et al. XGBoost-based model for predicting hydrogen content in electroslag remelting[J]. Journal of Iron and Steel Research International, 2023, 30(5): 887-896.

[5] 孙　辉, 杨　帆, 高正男, 等. 考虑特征重要性值波动的MI-BILSTM短期负荷预测[J]. 电力系统自动化, 2022, 46(8): 95-103.

[6] 谢万鹏, 刘　欢, 吴银花, 等. 基于改进SIFT和互信息法的单色和彩色视频高精度配准[J]. 液晶与显示, 2023, 38(12): 1689-1697.

[7] 李占山, 杨云凯, 张家晨. 基于熵权法的过滤式特征选择算法[J]. 东北大学学报(自然科学版), 2022, 43(7): 921-929.

[8] Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms[J]. Artificial Intelligence Review, 2021, 54(3): 1937-1967.

[9] 江　丽, 张智谟, 王琦玮, 等. 基于不同机器学习模型的石油测井数据岩性分类对比研究[J]. 物探与化探, 2024, 48(2): 489-497.

[10] 宋家威, 郇宝乾, 秦　涛, 等. 基于IGWO-CatBoost模型的岩石爆破块度预测[J]. 爆破器材, 2024, 53(2): 56-64.

[11] Sheridan R P, Wang W M, Liaw A, et al. Extreme gradient boosting as a method for quantitative structure-activity relationships[J]. Journal of Chemical Information and Modeling, 2016, 56(12): 2353-2360.

[12] Montero-Manso P, Athanasopoulos G, Hyndman R J, et al. FFORMA: Feature-based forecast model averaging[J]. International Journal of Forecasting, 2020, 36(1): 86-92.

[13] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent[J]. Journal of Statistical Software, 2010, 33(1): 1-22.

[14] 梁佳佳, 何晓霞, 肖浩逸. 基于CS-DBN的锂电池剩余寿命预测[J]. 太阳能学报, 2024, 45(3): 251-259.

[15] 和　征, 李忠鹏, 杨小红. 基于数字孪生与k-近邻算法的车间设备运行状态预测研究[J]. 制造技术与机床, 2024(3): 193-199.

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

[17] 彭雷朕, 姜周华, 沈国劬, 等. 电渣重熔9CrMoCoB钢过程电极表面氧化行为及脱氧制度[J]. 特殊钢, 2023, 44(6): 70-77.

[18] Li S J, Cheng G G, Huang Y, et al. Kinetics of phosphorus transfer during industrial electroslag remelting of G20CrNi2Mo bearing steel[J]. Metals, 2019, 9(4): 467.

[19] Gao G, Zhu C L, Shi X F, et al. Effect of magnetic field on elements segregation in electroslag ingot[J]. Journal of Iron and Steel Research International, 2022, 29(3): 434-444.

[20] Zhong Y B, Qiang L, Fang Y P, et al. Effect of transverse static magnetic field on microstructure and properties of GCr15 bearing steel in electroslag continuous casting process[J]. Materials Science and Engineering: A, 2016, 660: 118-126.

[21] 李练兵, 肖亚泽, 张　萍, 等. 基于CWT-RES34的风电机组叶片裂纹状态评估[J]. 噪声与振动控制, 2024, 44(2): 143-148+293.

[22] 谭建荣, 高铭宇, 徐敬华, 等. 数智化正向设计方法及其在制造装备与过程中的应用[J]. 机械工程学报, 2023, 59(19): 111-125.