基于重要性度量的脱硫剂加入量预测特征选择方法

赵海杰[1,2], 但斌斌[1,2*], 刘　洋[3], 任泽宇[4], 都李平[1,2], 周　纯[1,2]

(1. 武汉科技大学冶金装备及其控制教育部重点实验室, 湖北 武汉 430081; 2. 武汉科技大学机械传动与制造工程湖北省重点实验室, 湖北 武汉 430081; 3. 宝钢股份中央研究院 (武钢有限技术中心), 湖北 武汉 430080; 4. 湖南理工学院物理与电子科学学院, 湖南 岳阳 414006)

摘　要: 针对铁水 KR 脱硫生产工序中参数维度高、特征冗余性强以及目标变量与特征间相关性较弱的问题, 提出了一种基于重要性度量的集成式特征选择方法 IMFS(Feature selection based on importance measure)。在过滤式预筛选阶段, 通过最大互信息系数(MIC)度量各参数与目标变量的关联性以及各参数之间冗余性, 并根据最大相关、最小冗余准则缩小候选参数规模; 在嵌入式精选阶段, 引入 LightGBM 算法作为量化信息贡献度与数据敏感度的依托模型, 采用熵权法对双重度量结果进行赋权融合; 最后, 根据特征重要性系数, 结合 GBT 序列向前搜索策略优化特征子集。试验结果表明, IMFS 相较于其他方法, 在消除冗余特征和提升预测准确性方面具有显著优势, 并且能够有效平衡特征数量与预测精度。

关键词: 脱硫剂加入量; 特征选择; 重要性系数; 双重度量; 搜索策略

中图分类号: TP391.4　　**文献标志码:** A　　**文章编号:** 1004−7638(2025)05−0046−08

DOI: 10.7513/j.issn.1004-7638.2025.05.005　　**开放科学 (资源服务) 标识码 (OSID):**

听语音 聊科研 与作者互动

A feature selection method for desulfurizer addition prediction based on importance measure

ZHAO Haijie[1,2], DAN Binbin[1,2*], LIU Yang[3], REN Zeyu[4], DU Liping[1,2], ZHOU Chun[1,2]

(1. Key Laboratory of Metallurgical Equipment and Control Technology, Wuhan University of Science and Technology, Ministry of Education, Wuhan 430081, Hubei, China; 2. Hubei Key Laboratory of Mechanical Transmission and Manufacturing Engineering, Wuhan University of Science and Technology, Wuhan 430081, Hubei, China; 3. Baosteel Central Research Institute (Wuhan Iron and Steel Limited Technology Centre), Wuhan 430080, Hubei, China; 4. School of Physics and Electronic Science, Hunan Institute of Science and Technology, Yueyang 414006, Hubei, China)

Abstract: Aiming at the problems of high parameter dimension, strong feature redundancy and weak correlation between target variables and features in hot metal KR desulfurization production process, an integrated feature selection method IMFS (Feature selection based on importance measure) based on importance measure is proposed. In the filtering pre-screening stage, the maximal mutual information coefficient (MIC) is used to measure the correlation between each parameter and the target variable, as well as the redundancy among each parameter, and the scale of candidate parameters is reduced according to

收稿日期: 2025−07−22; **修回日期:** 2025−09−05; **接受日期:** 2025−09−05

基金项目: 国家自然科学基金项目(51475340); 湖北省重点研发计划项目(2022BAA059); 湖北省中央引导地方科技发展专项(2020ZYYD022)。

作者简介: 赵海杰, 1999 年出生, 男, 浙江诸暨人, 硕士研究生, 研究方向: 工艺参数分析与预测, E-mail: 15257541674@163.com;　*通讯作者: 但斌斌, 1970 年出生, 男, 湖北鄂州人, 教授, 研究方向: 炼钢铁水预处理工艺与设备优化, E-mail: danbinbin@163.com。

the maximum relevance and minimal redundancy criteria. In the embedded selection stage, the LightGBM algorithm is introduced as the supporting model for quantifying information contribution and data sensitivity, and the entropy weight method is used to weight and fuse the dual measurement results. Finally, according to the feature importance coefficient, the feature subset is optimized by combining the GBT sequential forward search strategy. The experimental results show that compared with other methods, IMFS has significant advantages in eliminating redundant features and improving prediction accuracy, and can effectively balance the number of features and prediction accuracy.

Key words: desulfurizer addition, feature selection, importance coefficient, double metric, search strategy

0　引言

在钢铁工业中, 铁水硫含量的精准控制是生产高质量钢材的关键。当硫含量较高时, 对钢材的力学性能以及成形特性具有一定影响[1-2]。铁水KR(Kambara Reactor)脱硫工艺作为钢铁冶炼中的重要环节, 其脱硫剂加入量的精确预测不仅涉及生产成本的控制, 也关乎产品质量的稳定性[3-4]。

目前, 基于历史数据的脱硫剂加入量预测中, 其核心挑战在于所涉及的工艺参数繁多且耦合复杂, 根据冶金机理直接建模往往难以取得理想的预测效果。因此, 如何从众多工艺参数中筛选出与脱硫剂加入量预测高度相关的特征, 是建立精准预测模型的首要任务。

现有研究主要分为基于统计度量和基于机器学习的两类特征选择方法[5]。前者因其计算效率高、实现简便, 已广泛应用于工业预测建模中。如王宁等[6] 基于互信息(MI)的特征选择方法, 通过衡量特征与目标变量间的依赖关系, 使能耗预测精度较传统方法提升 3.16%。严旭梅等[7] 采用皮尔逊相关系数筛选关键特征, 为 AdaBoost 模型预测脱硫率提供重要依据。然而, 相关方法常面临冗余特征难以剔除、全局依赖性难以捕捉等问题。

为克服上述局限, 基于机器学习的特征选择方法通过模型自身机制评估特征重要性 FI(Feature Importance), 展现出更强的适应性。例如, 方一飞等[8] 利用梯度提升决策树, 鉴别出对目标预测具有显著影响的特征集。徐猛等[9] 基于随机森林的袋外误差作为筛选评价标准, 优化特征维度。虽此类方法在特定场景下表现优异, 但仍存在不可忽视的缺陷: ①特征评估依赖基学习器的性能, 易受模型偏差的影响; ②单维度的选择机制难以表征复杂的特征交互, 导致结果的稳定性和泛化性不足。

鉴于此, 提出一种基于重要性度量的集成式特征选择方法 IMFS(Feature selection based on importance measure)。首先设计基于互信息的过滤式参数预筛选方法, 依据最大相关与最小冗余特性剔除部分不相关特征, 降低数据维度与运算成本; 在此基础上, 构建嵌入式关键参数识别模型, 利用 LightGBM 算法双重评估候选特征, 并引入 GBT 序列向前搜索策略, 动态识别关键工艺参数。

1　理论基础

1.1　最大互信息系数 MIC

最大互信息系数 MIC(Maximal Information Coefficient)通过度量两个随机变量之间的联合概率密度衡量其关联性, 能够有效捕捉包括线性、非线性及周期性等多类关系[10]。

MIC 的计算依赖于互信息和网格划分。对于给定的两个变量X和Y, 分别代表特征参数和目标变量, 其互信息$I(X:Y)$的计算如式 (1) 所示。

$$I(X:Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{1}$$

式中, $p(x,\ y)$为X和Y的联合概率密度。

给定一个有限且有序的集合$D=\{(x_i,y_i),\ i=1,2,\ldots,n\}$, 将其对应的散点图划分为$x\times y$的网格, 并计算每个网格中的互信息。选取不同划分方式下互信息的最大值, 求得最大互信息系数。其计算如式 (2) 所示。

$$\mathrm{MIC}(x,y)=\max_{|X||Y|<B}\frac{\max(I(X,Y))}{\log_2(\min(|X|,|Y|))} \tag{2}$$

式中, B为网格划分$x\times y$的上限值, 通常取$B(n)=n^{0.6}$。

1.2　LightGBM 模型

轻量级梯度提升机(Light Gradient Boosting Machine, LightGBM)是基于梯度提升决策树(Gradient Boosting Decision Tree, GBDT)的一种改进型迭代提升树系统[11]。通过引入独立特征合并算法和单侧采样算法优化, LightGBM 提升了传统 GBDT 的计算效率, 解决其在大规模数据处理中的计算瓶颈。

在模型中, LightGBM 提供了两种特征度量标准: ①Split, 即每个特征在所有决策树中被用作分裂

点的次数；②Gain，即特征作为分裂点所带来的信息增益。分裂点的增益(Gain)定义如式 (3) 所示。

$$\text{Gain} = \sum_{i=1}^{\text{left}} w_i + \sum_{j=1}^{\text{right}} w_j - \sum_{k=1}^{\text{all}} w_k \tag{3}$$

2 基于重要性度量的集成式特征选择方法 IMFS

面向 KR 脱硫工艺参数的集成式特征选择方法 IMFS 如图 1 所示。首先，对原始 KR 脱硫监测参数进行数据预处理，解决缺失值、异常值及量纲差异问题；继而，设计基于互信息的过滤式参数预筛选阶段，旨在获取与脱硫剂加入量高度相关的参数，同时计算各参数之间的关联性，并结合最大相关最小冗余特性对 KR 脱硫参数进行过滤式预筛选；随后，建立嵌入式特征精选方式，引入 LightGBM 算法对候选特征的信息贡献度与数据敏感度进行双重考量，通过熵权法融合二者结果，形成特征重要性系数 I_i。最后，利用 GBT 序列向前搜索策略，筛选能有效提升预测性能的特征组合，输出最优特征子集。

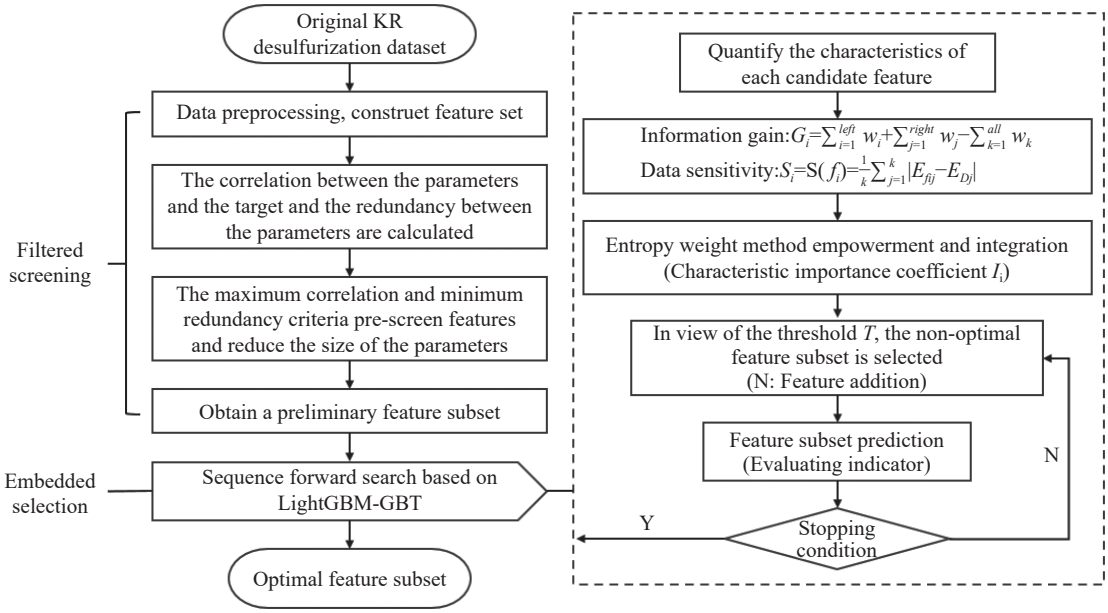


图 1 基于重要性度量的集成式特征选择方法 IMFS 框架

Fig. 1 An integrated feature selection method IMFS framework based on importance measure

2.1 过滤式 MIC 特征相关性分析

鉴于部分特征可能与因变量无关，采用式 (2) 评估特征间冗余性及其与目标变量的相关性，从而筛选有效特征集D。假定特征集F含n条样本和m个特征，任意两特征f_i和f_j的相关系数记作$\text{MIC}(f_i, f_j)$。当$\text{MIC}(f_i, f_j) > 0.8$时，表明特征间存在显著冗余性[12]。据此制定筛选规则：对于任意两个特征f_i、f_j，如果$\text{MIC}(f_i, f_j) > 0.8$，且特征$f_i$对目标变量相关性大于$f_j$，则视$f_j$为冗余特征并予以剔除。

2.2 嵌入式 LightGBM 特征重要性评估

针对复杂多变的 KR 脱硫工艺监测数据，单一度量难以表征特征性能。为弥补此不足，采用 LightGBM 算法作为评估基模型，首先基于预筛选特征计算其信息增益得分G，量化特征对目标变量的预测贡献度。

其次，提出基于数据扰动的特征敏感性评估机制。考虑到适度的噪声可增强模型对数据的鲁棒性并提升特征判别能力[13]，对特征值注入高斯噪声。核心原理在于关键特征受扰动时预测误差显著增大，而冗余或无关特征扰动时误差变化可忽略。处理流程如下：

1）数据标准化：使用 Min-Max 方法将特征数值转换至 [0, 1] 区间，消除量纲差异对特征选择的干扰；

2）高斯噪声添加：向特征集D依次添加噪声，生成d个扰动特征子集$\widetilde{D}$，如式 (4) ~ (6) 所示。

$$\widetilde{D} = \{f_1', f_2, \ldots, f_d\} \tag{4}$$

$$f_i' = x_1', x_2', \ldots, x_i', \ldots x_n' \tag{5}$$

$$x_i' = x + \epsilon_i, \epsilon_i \sim N\left(0, \sigma^2\right) \tag{6}$$

式中：d为当前特征维度；n为样本数目；ϵ_i表示高斯噪声服从均值为零、标准差为σ的正态分布。

3)预测误差求解: 利用 k-fold 交叉验证分别验证原特征集D和扰动特征集$\widetilde{D}$的预测误差, 记作E_{D}和E_{f_i};

4)敏感性评估: 通过衡量相对误差变化程度, 计算各特征的敏感性S_i, 如式 (7) 所示。

$$S_i = S(f_i) = \frac{1}{k}\sum_{j=1}^{k}\left|E_{f_i j} - E_{Dj}\right| \tag{7}$$

式中: E_{Dj}表示在第j折中使用原特征集D进行验证的误差; $E_{f_i j}$表示使用噪声特征集$\widetilde{D}$在第j折进行验证的误差。

最后, 基于信息贡献度G与数据敏感度S的熵值分析确定特征权重, 计算如式 (8) ~ (10) 所示。

$$H = -\sum_{i}^{d} p(i)\log_2(p(i)) \tag{8}$$

式中: $0 < p(i) < 1$, $p(i)$为第i个特征的概率。

$$w_G = \frac{1 - H(G)}{2 - H(G) - H(S)}, w_S = 1 - w_G \tag{9}$$

加权求和得到特征重要性系数I_i:

$$I_i = w_G \cdot G_i + w_S \cdot S_i \tag{10}$$

2.3　基于 GBT 序列向前的特征选择策略

常见特征选择策略包括遍历搜索、随机搜索和启发式搜索等, 为能够以较低的运算成本获得较高的性能评价指标, 设计 GBT 序列向前搜索策略, 如表 1 所示。该策略根据所获得特征重要性系数I_i进行降序排列, 并以系数的均值T作为阈值筛选初始特征子集, 然后遍历未选入特征与初始特征子集组合, 凭借 GBT 回归模型进行性能指标增量计算。

表 1　序列向前搜索过程

Table 1　Sequence forward search process

Iteration times	Current subset	Evaluation number ($a_0 > a_1 > a_2 > a_3$)	Optimal subset
1	Satisfy characteristic f_i	a_3	f_1f_2
	$f_1f_2f_3$	a_2	
2	$f_1f_2f_4$	a_0	$f_1f_2f_4$
	$f_1f_2f_5$	a_1	
3	$f_1f_2f_3f_4$	a_2	$Stop(f_1f_2f_4)$
	$f_1f_2f_4f_5$	a_1	

以决定系数(R^2)为例, 若$R^2(D_s + f_1) \geqslant R^2(D_s + f_2) \cdots \geqslant R^2(D_s + f_{d-s})$, 则$f_1$特征入选。迭代过程持续至性能无提升时终止, 通过定向特征扩展机制, 降低搜索复杂度。

3　试验及分析讨论

3.1　数据获取与预处理

数据来源于国内某钢厂铁水 KR 脱硫生产现场, 时间跨度为 6 个月, 总计 12 670 条记录, 并涵盖 127 个工艺参数, 部分数据如表 2 所示。原始数据存在较多缺失值和重复值, 考虑到应用插补法填补缺失值可能影响数据的真实性, 选择直接删除缺失值或重复值。同时, 根据钢厂脱硫工艺标准, 剔除脱硫后硫含量大于目标硫含量的记录, 并删除关键工艺参数为零的异常样本。经数据初步清洗, 剩余 5 054 条工艺数据。

表 2　某钢厂脱硫工艺数据实例

Table 2　An example of desulfurization process data in a steel plant

Number	Steel grade	Class	Group	Mixing speed/(r·min^{-1})	…	Process duration/min
1	M2A4-2	2	1	68	…	34
2	SPHC(MD)	2	4	75	…	33
3	B(H)	1	4	70	…	25
…	…	…	…	…	…	…
6 650	H2B1-1 B	1	1	65	…	29
…	…	…	…	…	…	…
12 670	B(H1)	2	4	74	…	31

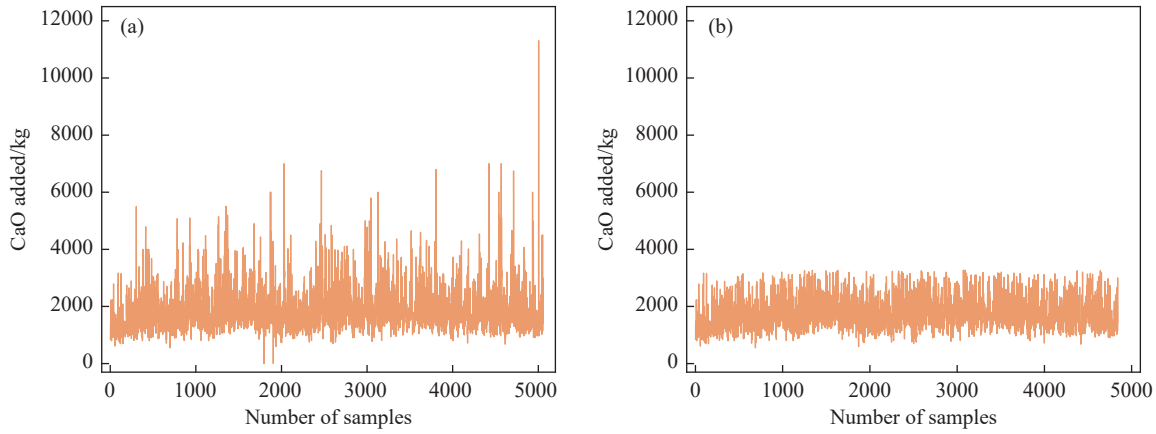
在铁水 KR 脱硫工艺中, 脱硫剂加入量与脱硫效率呈正相关。然而, 原始数据仅提供了脱硫前的初始硫含量和脱硫后的目标硫含量, 未直接提供硫含量的变化量。为此, 通过转换增添一项新特征列, 用来表示脱硫前后的硫含量差异, 记作 S_difference。此外, 采用编码方式将数据中的钢种类别转换为数字形式, 以便后续建模分析。进一步提升数据质量, 将采用 Pauta 准则对异常值进行识别与处理。如图 2 所示, 与未剔除异常的数据样本相比, 剔除边界后的数据范围更为合理, 并保留关键样本。最终, 得到 4 000 条数据用于特征性能评估。

为评价数据的准确性, 对清洗前后工艺参数进行统计分析, 部分离散型数据如表 3 所示。结果表明, 经过异常值剔除、缺失值删除和重复样本清理, 数据质量显著提升, 各参数的均值与中位数差异减小, 标准差整体下降, 分布更为集中, 为后续特征选择与建模提供可靠的数据基础。

3.2　特征变量选择

结合专家经验, 建立准确的脱硫剂加入量预测模型, 输入特征应严格限定为加剂前的工艺参数, 并将目标硫含量纳入特征集, 通过 MIC 方法量化输入特征间的信息冗余程度。经识别与过滤, 确定 11 个具有较高预测价值的关键特征。其中包括脱硫前温度 B_T、处理前硫含量 B_S、处理前硅含量 B_Si、搅拌桨使用次数 P_U、搅拌时间 S_T、搅拌桨插入深度 I_D、搅拌桨速度 P_S、液面高度 L_H、处理前

铁水质量 B_W、目标硫含量 T_S 以及脱硫前后的硫含量差 S_D。如图 3 所示，通过对预筛选特征进行信息增益量化分析可知，特征 S_D 的信息增益最大，表明该工艺参数对目标变量具有较强解释性。



(a) 未剔除异常值; (b) 剔除异常值

图 2　某特征异常值处理前后对比

Fig. 2　Schematic diagram of a characteristic abnormal value before and after processing

表 3　部分工艺参数统计分析

Table 3　Analysis of statistical indexes of some process parameters

Statistics		Temperature/℃	S content/%	Si content/%	Mixing speed/(r·min^{-1})	Depth/mm	CaO added/kg	Metal quality/kg
Before	Range	1 211 ~ 1 471	0.005 ~ 0.540	0.030 ~ 1.600	5 ~ 110	4 200 ~ 5 040	5 ~ 11 305	135 700 ~ 401 800
	Median	1 362	0.036	0.386	74	4 590	1 690.5	254 600
	Average	1 360	0.041	0.403	71.8	4 587	1 858.7	253 758
	Standard	32.3	0.022	0.161	9.1	214	724.8	11 840.6
After	Range	1 281 ~ 1 443	0.005 ~ 0.070	0.057 ~ 0.752	49 ~ 95	4 200 ~ 5 040	550 ~ 3 272	227 300 ~ 278 200
	Median	1 363	0.036	0.387	74	4 591	1 661	254 800
	Average	1 362	0.037	0.395	71.9	4 588	1 751.6	254 068
	Standard	29.8	0.012	0.126	8.9	214	493.8	9 810.7

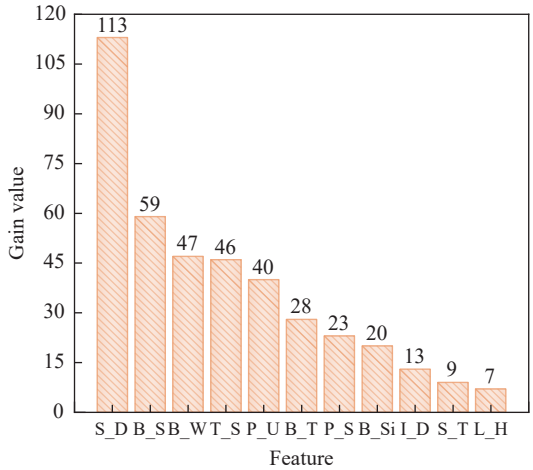


图 3　特征与目标变量的增益性度量

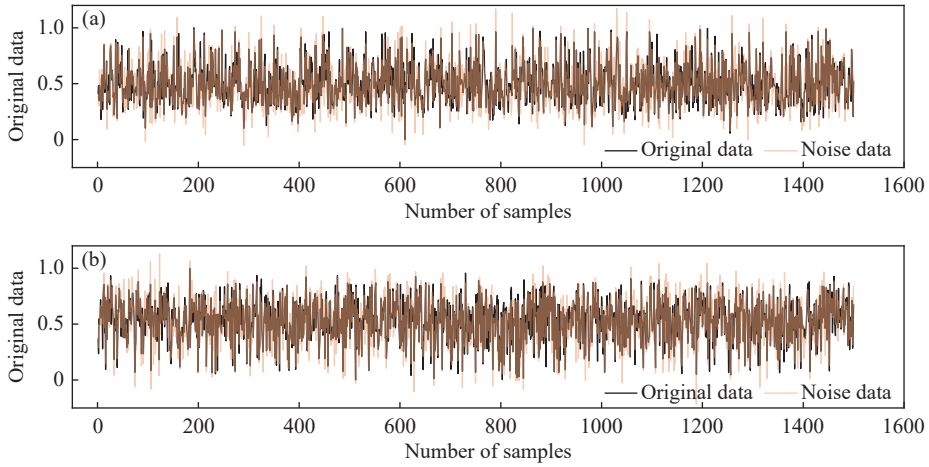
Fig. 3　Gain measure of feature and target variable

关于特征的敏感性，在标准化后的初选特征上添加均值为 0、标准差为 0.1 的高斯噪声。其标准差是通过反复试验所确定，保证在特征选择过程中不同因素所受到的干扰和不确定性程度相似，避免模型对特定特征的偏向。

以处理前硫含量与处理前铁水质量两个关键特征为例。如图 4 所示，可看出添加噪声后的特征仍保持原始数据的整体趋势，表明可控的随机波动为输入特征注入一定敏感性，有助于特征性能评估。

其次通过交叉验证进行 11 组对照数据的预测误差计算，利用公式 (7) 完成预筛选特征的敏感性评估，结果见表 4 所示。关于双重性能评估结果，运用基于信息熵理论的融合策略，合理分配权重（即 $w_G = 0.48$、$w_S = 0.52$）并加权，获取特征重要性系数 I_i，并设系数均值 $\theta = 0.0912$ 作为阈值，如图 5 所示。

由图 5 可知，满足阈值的特征共有 4 个，占总特征的 62.8%，优先作为目标预测的关键敏感特征。根据特征选择策略，采用均方根误差（RMSE）、平均绝对误差（MAE）以及决定系数（R^2）作为模型预测效果评价指标。如表 5 所示，随着特征个数的依次增加，预测效果呈现山峰型趋势。当特征子集由 S_D、T_S、B_S、B_W、P_U、B_Si 和 B_T 这 7 个特征构成时，模型的预测效果最佳。



(a) 处理前硫含量; (b) 处理前铁水质量

图 4　加噪前后的数据分布趋势

Fig. 4　Data distribution trend before and after adding noise

表 4　特征敏感性评估表

Table 4　Feature sensitivity assessment

Feature	S_i	Feature	S_i	Feature	S_i
B_T	5.07	S_T	1.84	B_W	7.66
B_S	6.75	P_S	3.41	T_S	15.45
B_Si	6.07	I_D	4.38	S_D	9.04
P_U	4.19	L_H	1.10		

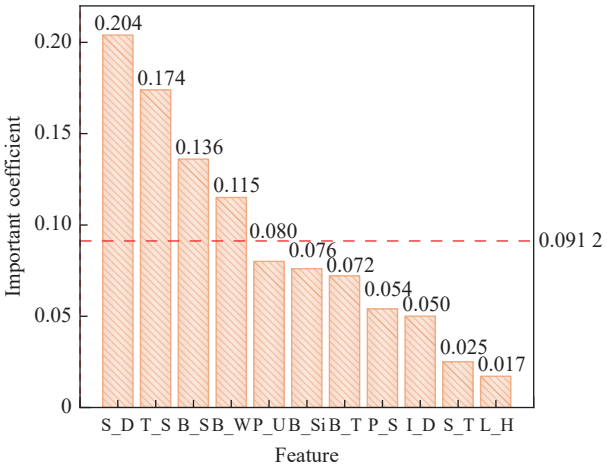


图 5　特征重要性系数 I_i

Fig. 5　Feature importance coefficient I_i

3.3　传统特征选择方法对比分析

为验证方法的可行性, 将采用预处理后的铁水 KR 脱硫数据, 分别与 Pearson 相关系数、Spearman 相关系数、MIC 最大互信息系数、LightGBM 嵌入式方法作对比, 如表 6 所示。

其次对不同特征选择方法以及全特征在回归模型下的预测效果进行系统性比较, 展示各方法的性能差异。具体模型包括 K 近邻(KNN)、深度神经网络(DNN)、随机森林(RF)、XGBoost 和支持向量回归(SVR), 模型的部分关键超参数设置如表 7 所示。与此同时, 将 RMSE、MAE、R^2 以及响应时间(t)作为评价指标。

表 5　特征子集评价表

Table 5　Feature subset evaluation

Optimal composition	R^2	RMSE	MAE
S_D、T_S、B_S、B_W	0.899 0	162.130 3	114.756 3
S_D、T_S、B_S、B_W、P_U	0.907 5	159.502 8	111.716 3
S_D、T_S、B_S、B_W、P_U、B_Si	0.907 5	155.106 9	109.309 5
S_D、T_S、B_S、B_W、P_U、B_Si、B_T	0.910 7	152.403 2	107.498 2
S_D、T_S、B_S、B_W、P_U、B_Si、B_T、I_D	0.907 3	155.335 2	109.323 7
S_D、T_S、B_S、B_W、P_U、B_Si、B_T、I_D、P_S	0.897 3	163.447 3	114.996 5
S_D、T_S、B_S、B_W、P_U、B_Si、B_T、I_D、P_S、L_H	0.902 2	155.118 2	108.806 7
S_D、T_S、B_S、B_W、P_U、B_Si、B_T、I_D、P_S、L_H、S_T	0.897 8	163.045 0	115.729 1

表 6　不同特征选择方法所选特征

Table 6　Features selected by different feature selection methods

Method	Feature subset
Pearson	S_D、B_S、T_S、L_H
Spearman	B_T、B_S、B_Si、T_S、B_W、L_H
MIC	B_S、B_W、B_Si、B_T、T_S、L_H、P_U
LightGBM	S_D、B_S、B_W、T_S、P_U、B_T、P_S、B_Si、I_D
IMFS	S_D、T_S、B_S、B_W、P_U、B_Si、B_T

从表 8 可知, 提出的特征选择方法在多个预测模型中, 均表现出较优的预测效果。在决定系数 R^2 指标方面, DNN、RF 和 XGBoost 模型分别达到 0.908、0.912 和 0.913, 均优于其他特征选择方法, 表明该方法更利于模型拟合。在误差分析方面, IMFS 方法也展现出优越的性能。以 KNN 模型为例, 其 RMSE 为 167.97, 较 MIC(203.51)和 Light-

GBM(182.01)特征选择方法分别降低了 17.5% 和 7.7%。此外,方法在确保较高的预测精度的同时,提高了模型的运行速度。以 SVR 模型为例,相较于全特征方法(3.895 s),节省了 27.0% 的计算时间。尽管与 Pearson 和 Spearman 特征选择方法相比,所选特征数量有所增加,但其模型运行时间相近。

表 7 模型超参数设置

Table 7 Super parameter setting of the model

KNN	DNN			RF		XGBoost		SVR		
K	Batch_size	Lr	Epochs	N_estimators	Max_depth	N_estimators	Lr	Kernel	Degree	Coef0
5	32	0.01	320	100	10	100	0.1	poly	3	1.0

表 8 传统特征选择方法对脱硫剂加入量预测的评价结果

Table 8 Evaluation results of the traditional feature selection method for the prediction of desulfurizer addition

Method	Model	R^2	RMSE	MAE	t/s
Pearson	KNN	0.869	187.48	156.66	0.085
	DNN	0.879	185.39	150.68	0.086
	RF	0.893	163.96	133.29	1.378
	XGBoost	0.884	175.86	148.28	0.133
	SVR	0.881	177.17	149.36	2.128
Spearman	KNN	0.865	191.24	160.53	0.012
	DNN	0.881	183.4	147.92	0.097
	RF	0.891	167.37	133.52	1.554
	XGBoost	0.889	170.77	139.25	0.152
	SVR	0.883	174.76	146.34	2.374
MIC	KNN	0.856	203.51	174.35	0.013
	DNN	0.894	167.73	134.07	0.103
	RF	0.893	164.48	132.12	1.815
	XGBoost	0.883	184.24	148.37	0.143
	SVR	0.889	169.77	138.53	2.43
LightGBM	KNN	0.873	182.01	153.25	0.018
	DNN	0.897	163.55	131.82	0.113
	RF	0.896	161.26	130.64	2.092
	XGBoost	0.894	162.73	130.88	0.15
	SVR	0.899	156.17	127.35	2.384
IMFS	KNN	0.884	167.97	146.44	0.013
	DNN	0.908	152.58	118.41	0.103
	RF	0.912	144.65	113.86	1.754
	XGBoost	0.913	142.79	112.34	0.145
	SVR	0.897	158.96	126.48	2.843
All features	KNN	0.729	330.98	298.47	0.216
	DNN	0.773	287.17	254.38	0.137
	RF	0.770	290.71	257.17	2.482
	XGBoost	0.765	297.22	263.29	0.178
	SVR	0.787	271.56	239.36	3.895

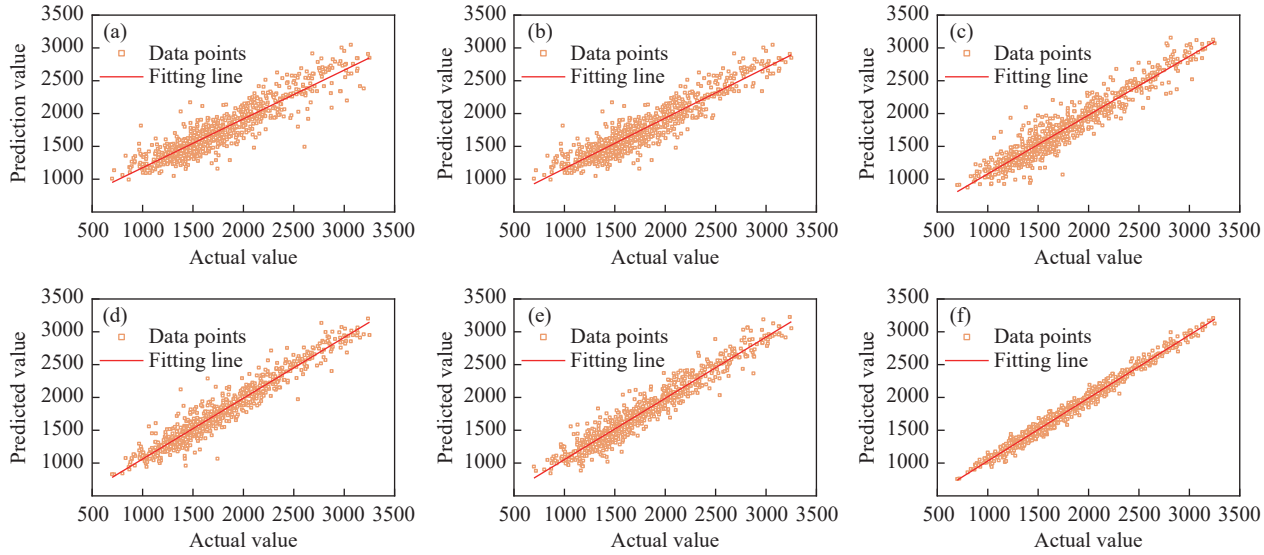
如图 6 所示,基于随机森林构建的预测模型中,不同特征选择方法对目标变量预测效果差异明显,凸显了特征选择在提升预测性能方面的重要性。试验结果表明,当使用全特征时(图 6(a)),预测值呈明显离散分布且偏离拟合线,说明原始特征集存在冗余信息。相比 Pearson 与 Spearman 相关系数方法(图 6(b)~(c)),预测分布有所改善,但部分区域仍存在偏差。相比之下,IMFS 方法(图 6(f))展现出更优异的性能,其预测值沿着拟合线分布,证实该方法能有效消除冗余特征并保留关键工艺信息。

3.4 深度学习特征提取方法对比分析

为进一步验证所提 IMFS 方法在高维、非线性工业数据场景下的有效性,引入三类典型的深度学习特征提取模型作为对比基准。此类方法摒弃显式的特征选择机制,依托神经网络固有的表示学习能力,自动地从原始数据中发掘深层特征模式与复杂非线性关系。对比模型包括人工神经网络(ANN)、广义回归神经网络(GRNN)以及一维卷积神经网络(1D-CNN)。所有模型均基于第 3.1 节预处理后的数据集训练,采用 5 折交叉验证,以均方误差(MSE)作为损失函数,并通过网格搜索优化关键超参数,以确保模型达到相对最优性能。

具体结构设置如下:ANN 采用四层全连接结构(64-32-16-1 神经元),每层后接批归一化、ReLU 激活函数及 Dropout(0.3)层;1D-CNN 由 1 个输入层、3 个卷积层和 2 个全连接层构成,卷积层后均跟随 ReLU 激活函数与最大池化;GRNN 基于高斯径向基核函数构建非线性加权回归结构,以前向传播实现快速预测。

结合表 8、9 的结果可知,尽管深度学习模型具备自动特征学习的潜力,但在样本规模有限且特征维度较高的工业数据场景中,若缺乏有效的特征筛选机制,其优势难以充分发挥。相比之下,IMFS 方法通过双重度量与搜索策略有效剔除冗余信息并突出保留关键工艺特征,从而显著提升预测精度和稳定性。



(a) 全特征; (b) Pearson 相关系数; (c) Spearman 相关系数; (d) MIC 最大互信息系数; (e) LightGBM 嵌入式; (f) IMFS

图 6　不同特征选择方法下的预测效果

Fig. 6　The prediction effect under different feature selection methods

表 9　深度学习特征提取方法的预测性能

Table 9　Prediction performance of deep learning feature extraction methods

Algorithm	Optimizer	Batch size	Lr	Epochs	R^2	RMSE	MAE
ANN	Adam	64	0.01	500	0.84	203.64	129.60
GRNN			1.0		0.82	215.77	156.01
1D-CNN	Adam	32	0.005	800	0.86	182.80	133.32

4　结论

1)以国内某钢铁厂的实际生产数据作为研究对象, 提出一种基于重要性度量的集成式特征选择方法 IMFS。该方法充分利用最大互信息系数, 合理地揭示了特征与因变量之间的相关性, 以达到过滤无关特征、去除冗余特征的目的。

2)为实现目标预测精度与特征数量的协同优化, 构建了嵌入式关键参数识别环节。通过采用 LightGBM 算法对候选特征进行双重评估, 引入熵权法对参数的增益性和敏感性进行加权融合, 并结合 GBT 序列向前搜索策略, 最终获取最优特征子集。

3)试验结果表明, IMFS 方法能够选出数量适中且具有高预测准确性的特征子集, 将复杂特征数由 127 个降为 7 个, 有效地解决了高维度、特征冗余以及目标弱相关的问题, 为脱硫剂加入量特征优选提供可靠的依据。

参考文献

[1] GAO J, CUI L, WANG W, *et al*. Prediction of sulfur content during steel refining process based on machine learning methods[J]. Steel Research International, 2024, 96(3): 2400662-2400662.

[2] GONG H J, LIANG X T, ZHOU Z C, *et al*. Application of rotary injection desulfurization technology in hot metal pretreatment[J]. Iron Steel Vanadium Titanium, 2020, 41(1): 173-178.
(龚洪君, 梁新腾, 周遵传, 等. 旋转喷吹脱硫技术在铁水预处理上的应用研究[J]. 钢铁钒钛, 2020, 41(1): 173-178.)

[3] ADHIWIGUNA IBGS, KARAGÜLMEZ G, KESKIN O, *et al*. Investigation on applicability of lime as desulfurization agent for molten cast iron[J]. Steel Research International, 2025, 96(1): 2400416.

[4] ZHENG Y, ZUO K L. Prediction model of desulfurizer consumption based on BP neural network and regression[J]. Iron Steel Vanadium Titanium, 2017, 38(4): 130-134.
(郑毅, 左康林. 基于 BP 神经网络和回归的脱硫粉剂预报模型[J]. 钢铁钒钛, 2017, 38(4): 130-134.)

[5] LIU Z X, DU J Q, LUO J G, *et al*. Review on stability feature selection[J]. Computer Engineering and Applications, 2025, 61(7): 81-95.
(刘梓萱, 杜建强, 罗计根, 等. 稳定性特征选择研究综述[J]. 计算机工程与应用, 2025, 61(7): 81-95.)

(下转第 64 页)

[13] HE J G, DENG A Y, XU X J, *et al*. Effect of electromagnetic stirring position on liquid steel flow and liquid level fluctuation in continuous casting mold for wide thick slab[J]. Continuous Casting, 2022, 4: 50-58.
（何建国, 邓安元, 许秀杰, 等. 电磁搅拌宽厚板结晶器内钢液流动和液面波动[J]. 连铸, 2022, 4: 50-58.）
[14] XIE X X, LUO S, CHEN Y, *et al*. Effect of electromagnetic stirring on flow, solidification and liquid level fluctuations in slab mold[J]. Continuous Casting, 2025, 44(2): 7-14.
（解晓晓, 罗森, 陈耀, 等. 电磁搅拌对板坯结晶器内钢液流动、凝固和液面波动的影响[J]. 连铸, 2025, 44(2): 7-14.）
[15] SUN X H, LI B, LU H B, *et al*. Steel slag interface behavior under multifunction electromagnetic driving in a continuous casting slab mold[J]. Metals, 2019, 9(9): 983-999.
[16] LI B, LU H B, ZHONG Y B, *et al*. Numerical simulation for the influence of EMS position on fluid flow and inclusion removal in a slab continuous casting mold[J]. ISIJ Int., 2020, 60(6): 1204-1212.
[17] LIU G L, LU H B, LI B, *et al*. Influence of M-EMS on fluid flow and initial solidification in slab continuous casting[J]. Materials, 2021, 14(13): 3681-3697.
[18] XU L, PEI Q W, LI N, *et al*. Study on the effect of multi area controllable electromagnetic braking on behavior of non-uniform molten steel flow and steel-slag interface in the mold[J]. Iron Steel Vanadium Titanium, 2025, 46(1): 112-123.
（许琳, 裴群武, 李楠, 等. 多区域独立可控电磁制动对结晶器内钢液非均匀流动与渣金界面行为影响的研究[J]. 钢铁钒钛, 2025, 46(1): 112-123.）
[19] XU L, WANG E G, KARCHER C, *et al*. Numerical simulation of the effects of horizontal and vertical EMBr on jet flow and mold level fluctuation in continuous casting[J]. Metall. Mater. Trans. B, 2018, 49(5): 2779-2793.
[20] XU L, KARCHER C, WANG E G. Numerical simulation of melt flow, heat transfer and solidification in CSP continuous casting mold with vertical-combined electromagnetic braking[J]. Metall. Mater. Trans. B, 2023, 54(4): 1646-1664.
[21] XU L, HAN Z F, KARCHER C, *et al*. Melt flow, heat transfer and solidification in a flexible thin slab continuous casting mold with vertical-combined electromagnetic braking[J]. J. Iron Steel Res. Int., 2024, 31(2): 401-415.
[22] ZHANG A H, MA D Z, JIAN W W, *et al*. Numerical simulation of argon blowing behavior inside an independent adjustable combination electromagnetic brake mold[J]. Continuous Casting, 2024, 4: 38-46.
（张安昊, 马丹竹, 建伟伟, 等. 独立可调式组合电磁制动结晶器内吹氩行为数值模拟[J]. 连铸, 2024, 4: 38-46.）

编辑 邓淑惠

（上接第 53 页）

[6] WANG N, LI X F, NIE L D, *et al*. High-precision vehicle energy consumption prediction using mutual information feature selection[J]. Journal of Tongji University (Natural Science), 2024, 52(S1): 39-45.
（王宁, 李秀峰, 聂辽栋, 等. 基于 MI 特征选择的车辆能耗高精度预测方法[J]. 同济大学学报 (自然科学版), 2024, 52(S1): 39-45.）
[7] YAN X M, CHEN C, WANG N, *et al*. Prediction of desulfurization rate during LF refining process based on random search and AdaBoost model[J]. Journal of Materials and Metallurgy, 2023, 22(5): 430-436, 443.
（严旭梅, 陈超, 王楠, 等. 基于随机搜索算法和 AdaBoost 模型预测 LF 精炼过程脱硫率[J]. 材料与冶金学报, 2023, 22(5): 430-436, 443.）
[8] FANG Y F, DAN B B, WU J W, *et al*. Method for predicting desulfurizer dosage based on ensemble learning[J]. Journal of Wuhan University of Science and Technology, 2024, 47(5): 361-367.
（方一飞, 但斌斌, 吴经纬, 等. 基于集成学习的脱硫剂加入量预测方法[J]. 武汉科技大学学报, 2024, 47(5): 361-367.）
[9] XU M, LEI H, HE J Y, *et al*. Predicting the endpoint steel temperature of RH refining using improved XGBoost[J]. Journal of Materials and Metallurgy, 2023, 22(5): 437-443.
（徐猛, 雷洪, 何江一, 等. 利用改进 XGBoost 预测 RH 精炼终点钢水温度[J]. 材料与冶金学报, 2023, 22(5): 437-443.）
[10] GU T Y, GUO J S, LI Z X, *et al*. Detecting associations based on the multi-variable maximum information coefficient[J]. IEEE Access, 2021, 9: 54912-54922.
[11] JU Y, SUN G Y, CHEN Q H, *et al*. A model combining convolutional neural network and LightGBM algorithm for ultra-short-term wind power forecasting[J]. IEEE Access, 2019: 28309-28318.
[12] LI Y Z, DAI W, ZHANG W F. Bearing fault feature selection method based on weighted multidimensional feature fusion[J]. IEEE Access, 2020, 8: 19008-19025.
[13] ZHANG S G, ZHOU T, SUN L, *et al*. *v*-Support vector regression model based on Gauss-Laplace mixture noise characteristic for wind speed prediction[J]. Entropy, 2019, 21(11): 1056.

编辑 邓淑惠