文章编号: 1671-251X(2024)12-0136-09 DOI: 10.13272/j.issn.1671-251x.2024070022

基于特征选择与 BO-GBDT 的工作面瓦斯涌出量预测方法

马文伟[1,2]

(1. 中煤科工集团沈阳研究院有限公司, 辽宁 抚顺 113122;
2. 煤矿安全技术国家重点实验室, 辽宁 抚顺 113122)

摘要: 影响工作面瓦斯涌出量的特征众多, 利用主成分分析等方法对原始数据降维, 可节省计算资源, 但会改变数据集的原始特征结构, 损失部分原始数据特征的细节信息。针对该问题, 建立梯度提升决策树(GBDT)瓦斯涌出量预测模型, 利用 5 种特征选择算法对数据集进行特征过滤, 分析每种特征组合在 GBDT 模型中的拟合度、计算时间及预测结果, 优选出包装法为最佳的特征选择算法; 结合现场实际, 优选出 8 种特征进行瓦斯涌出量预测, 结果表明, 特征数量的多少与预测结果的准确性和泛化性并不呈正比关系, 冗余特征或无关特征的存在反而会降低模型的预测准确性。为进一步提高模型精度, 通过 5 种超参数寻优算法对 GBDT 模型进行超参数寻优, 对比分析每一种超参数组合下 GBDT 模型的预测性能, 结果表明: 寻优算法本身对 GBDT 模型的准确性和泛化性影响较小, 但基于树结构 Parzen 估计器 (TPE) 的贝叶斯优化(BO)算法所得出的最优超参数组合在 GBDT 模型中具有最高的准确率和相对较少的优化时间, 其优化性能最佳, 以此建立 BO-GBDT 模型。将特征选择后的数据集划分出训练集及测试集, 利用 BO-GBDT 模型进行工作面瓦斯涌出量预测, 并与随机森林、支持向量机、神经网络模型进行对比, 结果表明: BO-GBDT 模型具有更高的准确性和泛化性, 其平均相对误差为 2.61%, 相比随机森林、支持向量机、神经网络模型分别降低了 35.56%, 37.41%, 32.03%, 能够满足现场工程应用需求, 为矿井安全生产提供理论指导。

关键词: 瓦斯涌出量预测; 特征选择; 梯度提升决策树; 贝叶斯优化; 超参数优化; 机器学习

中图分类号: TD712.5 文献标志码: A

Prediction method of gas emission in working face based on feature selection and BO-GBDT

MA Wenwei[1,2]

(1. CCTEG Shenyang Research Institute, Fushun 113122, China;
2. State Key Laboratory of Coal Mine Safety Technology, Fushun 113122, China)

Abstract: Gas emission in the working face is influenced by a variety of factors. Dimensionality reduction methods, such as Principal Component Analysis, can reduce computational resources but may alter the original feature structure, leading to a loss of some detailed information in the dataset. To address this issue, a gradient boosting decision tree (GBDT) model for gas emission prediction was developed. Five feature selection algorithms were applied to filter the dataset, and the model fit, computational time, and prediction accuracy of each feature combination in the GBDT model were analyzed. The wrapping method was identified as the most effective feature selection algorithm. Based on field conditions, 8 optimal features were selected for prediction. The results indicated that the number of features did not necessarily correlate with the prediction's accuracy or

收稿日期: 2024-07-07; **修回日期**: 2024-12-22; **责任编辑**: 胡娴。

基金项目: 国家科技重大专项资助项目(2016ZX05045-004-001)。

作者简介: 马文伟(1985—), 男, 山西大同人, 副研究员, 硕士, 主要从事矿井瓦斯灾害防治及煤矿智能化方面的研究工作, E-mail: 120598723@qq.com。

引用格式: 马文伟. 基于特征选择与 BO-GBDT 的工作面瓦斯涌出量预测方法[J]. 工矿自动化, 2024, 50(12): 136-144.

MA Wenwei. Prediction method of gas emission in working face based on feature selection and BO-GBDT[J]. Journal of Mine Automation, 2024, 50(12): 136-144.

扫码移动阅读

generalization capability. In fact, redundant or irrelevant features reduced the model's prediction accuracy. To further improve performance, five hyperparameter optimization algorithms were applied to the GBDT model. A comparative analysis of prediction performance for each hyperparameter combination was conducted. The results showed that the optimization algorithm itself had minimal impact on the accuracy and generalization of the GBDT model. However, the optimal hyperparameter combination, obtained through the tree-structured Parzen estimator (TPE) based Bayesian optimization (BO) algorithm, provided the highest accuracy and relatively short optimization time, yielding the best optimization performance. Thus, the BO-GBDT model was established. After feature selection, the dataset was divided into training and testing sets, and the BO-GBDT model was used to predict gas emission in the working face. Comparison with random forest, support vector machine, and neural network models showed that the BO-GBDT model achieved the highest accuracy and generalization, with an average relative error of 2.61%. This was 35.56%, 37.41%, and 32.03% lower than the random forest, support vector machine, and neural network models, respectively. The BO-GBDT model meets the field engineering application requirements and provides theoretical guidance for ensuring safe mining production.

Key words: gas emission prediction; feature selection; gradient boosting decision tree; Bayesian optimization; hyperparameter optimization; machine learning

0　引言

瓦斯涌出量是矿井进行瓦斯灾害防治与管理、矿井通风系统优化设计的重要基础数据。AQ 1018—2006《矿井瓦斯涌出量预测方法》[1]明确指出，新建矿井或生产矿井新水平，都必须进行瓦斯涌出量预测，以确定新矿井、新水平、新采区投产后瓦斯涌出量大小，预测结果作为矿井和采区通风设计、瓦斯抽放及瓦斯管理的依据。在煤矿开采过程中，矿井瓦斯涌出的主要来源为回采工作面，因此，快速、精准预测回采工作面的瓦斯涌出情况，并据此采取相应措施降低工作面瓦斯涌出量，能够有效降低事故发生概率，减少人员伤亡和财产损失。

《矿井瓦斯涌出量预测方法》中指出，回采工作面的瓦斯涌出主要来源于开采层与邻近层瓦斯涌出，分别给出了瓦斯涌出量计算经验公式。依据经验公式计算瓦斯涌出量的方法较简单，操作方便，故现场应用广泛。但该方法依赖于煤层原始瓦斯含量、残存瓦斯含量的精确测定，以及采出率、瓦斯涌出影响系数、瓦斯排放率等参数的确定，人为影响因素较多，导致预测结果与现场实际情况不符的状况时有发生。

随着计算机技术的快速发展，国内外学者提出了基于各种算法的工作面瓦斯涌出量预测方法[2-5]。文献[6]建立了基于增强分类与回归树（Classification and Regression Tree, CART）算法的采煤工作面瓦斯涌出量的量化模型，有效提高采煤工作面瓦斯涌出量预测精度。文献[7-9]构建了基于随机森林算法的回采工作面瓦斯涌出量预测模型，一定程度上提高了预测精度和效率。文献[10]通过最小角回归（Least Angle Regression, LARS）算法实现降维，采用最小绝对值压缩和选择算子（Least Absolute Shrinkage and Selection Operator, LASSO）回归模型进行回采工作面瓦斯涌出量仿真预测，在预测精度及泛化能力上有所提高。文献[11]提出了基于因子分析法与 BP 神经网络的工作面瓦斯涌出量预测模型，为复杂影响因素影响下的工作面瓦斯涌出量预测提供了新思路。文献[12]采用主成分分析获得影响回采工作面瓦斯涌出量的主成分，并采用主成分分量进行多步线性回归，预测工作面瓦斯涌出量，具有较高的精度。

影响工作面瓦斯涌出量的特征众多，利用主成分分析等方法对原始数据降维，可节省计算资源，提高计算效率[13-15]，但会改变数据集的原始特征结构，损失部分原始数据特征的细节信息，使得在数据采集过程中难以把握数据的完整性和准确性，易造成数据过度采集，导致采集成本大幅增加且数据质量参差不齐。针对该问题，本文采用多种特征选择算法，在原始特征空间的基础上选择最优特征子集，不改变数据集的特征意义，同时使其具有更好的数据可收集性和可读性；建立梯度提升决策树（Gradient Boosting Decision Tree, GBDT）模型，并通过贝叶斯优化（Bayesian Optimization, BO）算法对模型超参数进行优化，构建基于特征选择与 BO−GBDT 的工作面瓦斯涌出量预测模型。

1　工作面瓦斯涌出量样本采集与分析

工作面瓦斯涌出的影响因素主要包括煤体瓦斯

赋存因素、煤层赋存条件因素、回采工艺及管理因素等。煤体瓦斯赋存因素主要包括煤层原始瓦斯含量、邻近层原始瓦斯含量等；煤层赋存条件因素主要包括煤层埋深、煤层厚度、煤层倾角、邻近层煤厚、层间岩性等；回采工艺及管理因素主要包括工作面长度、回采高度、采出率、日产量等。

收集中煤新集刘庄矿业有限公司、开滦（集团）有限责任公司钱家营煤矿、陕西黄陵二号煤矿有限公司、山西马堡煤业有限公司的 73 组回采工作面瓦斯涌出量相关特征数据进行实验[8-9,14,16-17]，部分数据见表 1。其中 X_1 为瓦斯含量，X_2 为煤层埋深，X_3 为开采层厚度，X_4 为煤层倾角，X_5 为回采高度，X_6 为日进尺，X_7 为工作面长度，X_8 为采出率，X_9 为日产量，X_{10} 为邻近层瓦斯含量，X_{11} 为邻近层煤层厚度，X_{12} 为邻近层与本煤层间距，X_{13} 为邻近层与本煤层层间岩性，Y 为标签，指回采工作面绝对瓦斯涌出量。刘庄矿业有限公司的数据来源于该矿 13−1 号煤层、11−2 号煤层、8 号煤层，钱家营煤矿的数据来源于该矿 7 号煤层、9 号煤层、12−1 号煤层，黄陵二号煤矿有限公司的数据来源于该矿 2 号煤层，马堡煤业有限公司的数据来源于该矿 8 号煤层、15 号煤层。

表 1　回采工作面瓦斯涌出量样本数据

Table 1　Gas emission sample data of mining working face

序号	X_1/(m³·t⁻¹)	X_2/m	X_3/m	X_4/(°)	X_5/m	X_6/m	X_7/m	X_8/%	X_9/t	X_{10}/(m³·t⁻¹)	X_{11}/m	X_{12}/m	X_{13}	Y/(m³·min⁻¹)
1	3.90	499	4.3	15	4.3	10	280	0.93	17 217	3.10	2.80	52	5.89	2.71
2	3.16	502	2.7	8	2.7	10	290	0.93	11 197	2.80	1.79	48	4.90	2.84
3	3.40	522	3.4	12	3.4	8	280	0.95	10 891	2.15	1.72	14	4.71	3.20
4	2.96	540	2.8	10	2.8	8	290	0.95	9 289	2.44	2.20	20	4.24	3.60
5	3.68	513	3.5	12	3.5	9	285	0.94	12 838	3.28	1.80	19	4.54	3.10
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
71	2.46	448	2.3	11	2.3	4.33	159	0.95	1 998	2.01	1.69	17	4.65	4.07
72	3.12	541	2.6	13	2.6	3.82	166	0.94	2 207	2.3	1.81	14	4.72	4.92
73	4.65	630	6.3	12	6.3	2.81	170	0.93	3 457	3.34	1.62	19	4.65	8.05

分析表 1 可知，X_1—X_{13} 特征之间本身存在重复、冗余等问题。例如，除放顶煤等特殊开采工艺外，开采层厚度 X_3 与回采高度 X_5 一般均一致，本数据集中这 2 项特征数值基本一致，存在冗余。回采高度 X_5、日进尺 X_6、工作面长度 X_7 的乘积与日产量 X_9 相等，存在冗余。为进一步分析数据集中各特征之间及特征与标签之间的相关性，绘制相关性热图，如图 1 所示。

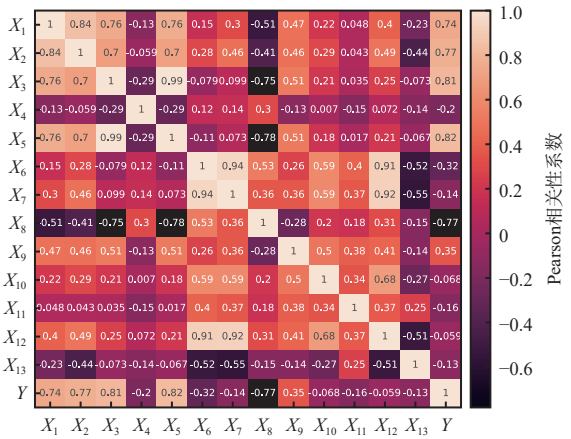

图 1　数据集特征与标签相关性热图

Fig. 1　Heatmap of correlation between features and labels in dataset

Pearson 相关性系数越大，代表对应的 2 个参数相关性越强。分析图 1 可知，2 个特征因素之间、单个特征与标签之间的相关性各不相同，总体相关性较低。统计发现，2 个特征之间或者单个特征与标签的相关性超过 0.80 的为 7 组，占总量的 7.7%；相关性超过 0.70 的为 11 组，占总量的 12.09%。

若直接采用所有特征进行瓦斯涌出量预测，可能会降低模型精度，增加模型复杂度，影响计算效率。因此，有必要通过特征选择对原始数据集进行处理，减少不相关和冗余特征的影响。同时，降低特征数量也会极大降低后续数据集的获取难度，降低工人及技术人员采集数据的劳动强度。

2　特征选择

特征越多并不直接意味着会使模型性能变好，反之会使模型更复杂，训练时间更长，带来“维度灾难”[18]。在表 1 所示的数据集中，影响工作面瓦斯涌出量的因素较多，且在实际数据收集工作中难度较大，因此，在不影响模型预测准确性的前提下，以较少特征进行瓦斯涌出量预测，对降低现场工作量具有重要意义。特征选择也称特征子集选择，

是指从 M 个特征中选择 N 个更易于理解的特征，挖掘底层数据中隐藏的有用信息，以提升模型预测性能，同时降低特征维度，大幅度提升模型的计算效率。

特征选择算法主要包括方差过滤法、F 检验法、互信息法、嵌入法、包装法。方差过滤法基于特征的方差与特征区分度的相关性进行特征选择。F 检验法通过捕捉每个特征与标签之间的线性关系，计算 2 组或多组数据之间的方差比值，检验数据之间是否存在显著差异，进而进行特征选择。互信息法通过估计每个特征与标签之间任意关系的互信息量，以判断特征的取舍。嵌入法通过机器学习算法和模型进行训练，得到每个特征的权值系数，并依据系数大小进行特征选择。包装法利用目标函数作为黑盒，找到最佳特征子集，并通过反复创建模型保留最佳特征或剔除最差特征，直到资源耗尽，完成特征选择。

为找出最优特征选择算法，对比分析每种算法计算结果在 GBDT 模型中的表现，结果见表 2，R^2 为决定系数。

表 2　不同特征选择算法的特征选择结果

Table 2　Feature selection results of different feature selection algorithms

特征选择算法	特征													R^2	计算时间/s
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}		
方差过滤法	√	√	√	√	√	√	√	×	×	×	√	√	×	0.846 0	0.13
F检验法	√	√	√	×	√	√	×	√	√	×	×	×	×	0.849 1	0.13
互信息法	√	√	√	√	√	√	√	√	√	√	√	√	√	0.848 3	0.16
嵌入法	√	√	√	×	×	√	×	√	×	×	√	×	×	0.851 5	0.43
包装法	√	√	×	×	√	√	×	√	×	√	×	×	×	0.851 5	0.14
未进行特征选择	√	√	√	√	√	√	√	√	√	√	√	√	√	0.848 3	0.16

由表 2 可知，采用嵌入法和包装法进行特征选择后，特征由原来的 13 个降为 6 个，特征数量降低 53.85%，通过 GBDT 算法默认参数所建立的模型评估指标 R^2 由原来的 0.848 3 提高为 0.851 5，包装法在提高模型拟合度的同时，计算时间也由 0.16 s 缩短为 0.14 s。因此本文采用包装法所得特征选择结果进行下一步分析。

进一步分析表 2 可知，特征中瓦斯含量 X_1、煤层埋深 X_2、开采层厚度 X_3、回采高度 X_5、日进尺 X_6、采出率 X_8 均被 4 种以上的特征选择算法选择，其中只有开采层厚度未被包装法选择，由于本数据集中开采层厚度与回采高度基本一致，故按照包装法所选，只选择回采高度特征。

与本煤层瓦斯涌出量相关的其他特征中，煤层倾角、工作面长度与工作面瓦斯涌出量相关性相对较弱，日产量与回采高度、日进尺、采出率等特征冗余，结合包装法选择结果，剔除煤层倾角、工作面长度、日产量这 3 个特征。

与邻近层瓦斯涌出相关的特征包括邻近层瓦斯含量、邻近层煤层厚度、邻近层间距、层间岩性等级，由表 2 可知，层间岩性等级与工作面瓦斯涌出量相关性最弱。考虑到现场实际中，邻近层工作面瓦斯涌入会对本煤层工作面瓦斯涌出量造成较大影响，故保留邻近层瓦斯含量、邻近层煤层厚度、邻近层间距这 3 个与邻近层瓦斯涌出相关性较强的特征。

经过特征选择后，共保留 8 个特征，包括瓦斯含量 X_1、煤层埋深 X_2、回采高度 X_5、日进尺 X_6、采出率 X_8、邻近层瓦斯含量 X_{10}、邻近层煤层厚度 X_{11}、邻近层与本煤层间距 X_{12}。相比原数据集，特征数量减少 38.46%，通过 GBDT 算法默认参数所建立的模型评估指标 R^2 由原来的 0.848 3 降低为 0.840 0，仅降低 0.98%，计算时间由 0.16 s 缩短为 0.14 s。

3　模型构建及超参数优化

3.1　GBDT 算法原理

GBDT 是一种基于 Boosting 算法的集成决策模型[19-20]。核心思想是通过多个决策树的迭代训练，逐步优化预测结果，提升模型准确性，集成模型的最终输出结果受所有决策树的影响。

算法实现过程：

1）给定训练数据集 T，$T=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_n,y_n)\}$，其中 $\boldsymbol{x}_i(i=1,2,\cdots,n)$为特征向量，$y_i$为目标值，$n$ 为训练样本数量。

2）通过计算所有目标的平均值得到初始预测值，作为后续迭代的基础。初始化常数模型$F_0(x)$：

$$F_0(x)=\frac{1}{n}\sum_{i=1}^{n}y_i \tag{1}$$

式中x为$\boldsymbol{x}_1$—$\boldsymbol{x}_n$组成的 n 维特征空间。

3）计算第 $m-1$ 次迭代模型 $F_{m-1}(\boldsymbol{x}_i)$ 与 y_i 的残差 $\gamma_{i,m}$，$m=1, 2, \cdots, M$，M 为迭代总次数。残差反映当前模型尚未拟合的部分，后续模型将针对这些残差构建决策树进行拟合。

$$\gamma_{i,m} = y_i - F_{m-1}(\boldsymbol{x}_i) \tag{2}$$

4）以 $(\boldsymbol{x}_i, \gamma_{im})$ 为新的训练数据，拟合回归树 $h_m(x)$。回归树的构建过程是通过划分特征空间，使得每个叶节点内的数据具有相似目标值。在回归树构建过程中，不断寻找最优划分特征和划分点，以最小化叶节点内的设定度量。

5）更新第 m 次迭代模型 $F_m(x)$：

$$F_m(x) = F_{m-1}(x) + \alpha_m h_m(x) \tag{3}$$

式中 α_m 为学习率，表示每棵新生成的决策树对最终预测值的影响程度。

6）经过 M 次迭代后，得到最终 GBDT 模型 $F_M(x)$：

$$F_M(x) = F_0(x) + \sum_{m=1}^{M} \alpha_m h_m(x) \tag{4}$$

3.2　超参数优化

BO 是一种基于贝叶斯推断的全局优化算法，核心思想是在参数空间中建立目标函数的后验模型，并使用该模型来指导下一步的参数选择，从而在尽可能少的迭代次数内找到函数的全局最优解或局部最优解[21-23]。

超参数优化的目的是提高 GBDT 模型的预测精度。基于 BO 算法的超参数优化流程如图 2 所示。

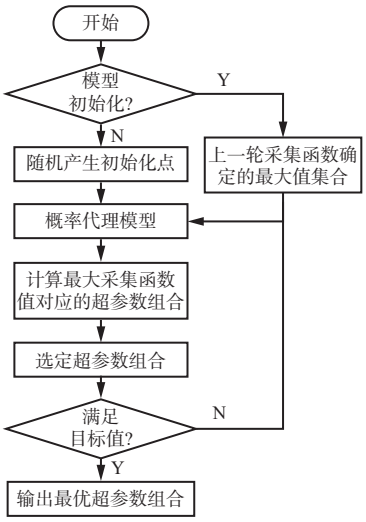

图 2　基于 BO 算法的超参数优化流程

Fig. 2　Hyperparameter optimization process based on Bayesian optimization（BO）algorithm

1）判断模型是否初始化，若未进行初始化，即为优化流程的初始阶段，通过随机生成初始化样本点构建模型；若已完成初始化，即之前已有迭代，则利用上一轮采集函数确定的最大值集合继续优化。

2）利用步骤 1）中的初始化点或上一轮得到的最大值集合及其对应的目标函数值构建概率代理模型，估计已有数据点目标函数在整个超参数空间中的分布。

3）根据概率代理模型，选择采集函数，通过计算最大采集函数值所对应的超参数组合，确定下一个评估点。

4）使用选定的超参数组合训练模型，并在验证集中评估模型，得到对应的目标函数值。

5）如果该目标函数值满足终止条件，即停止迭代，输出最优超参数组合；如果不满足终止条件，则循环执行步骤 2）—步骤 5），直至目标值达到设定标准或设置的计算资源耗尽。

4　实验分析

4.1　GBDT 超参数优化

以特征选择结果作为新的数据样本集，将数据集中的 60 个数据作为训练样本集，剩余 13 个数据作为测试验证样本集。借助 Python 语言，结合 Sklearn 机器学习库中的 GradientBoostingRegressor 构建 GBDT 回归预测模型。GBDT 模型的超参数数量较多，结合本文数据集的特点，剔除对模型影响相对较小的参数，选择 8 个超参数作为优化目标，见表 3。

表 3　GBDT 模型超参数

Table 3　Gradient boosting decision tree（GBDT）model hyperparameters

序号	超参数名称	含义
1	n_estimators	弱学习器最大个数
2	learning_rate	学习率
3	max_features	划分时考虑的特征数量
4	Subsample	子采样比例
5	loss	损失函数选择
6	criterion	衡量每个决策树节点分裂质量的评价指标
7	max_depth	每棵子树的深度
8	min_impurity_split	最小基尼不纯度

常用的超参数优化算法包括网格搜索算法、随机搜索算法、BO 算法、遗传算法、粒子群算法、梯度下降类优化算法等。网格搜索算法通过穷举所有超参数组合进行寻优，较为直观但计算量较大。随机搜索算法通过在超参数空间随机抽取组合进行评估，计算量较小。BO 算法通过构建不同目标函数（高斯函数等）调整搜索策略，智能寻求最优解。遗

传算法通过模拟生物进化过程进行搜索，计算量较大，适用于离散优化。粒子群算法通过模拟鸟群觅食行为进行搜索，对参数设置较为敏感，易陷入局部最优解。梯度下降算法需要计算目标函数关于超参数的梯度信息，不适用于高维空间。

本文数据集较小，但需优化的超参数空间较大，要构建能够有效利用有限数据且在较大参数空间内高效搜索最优解的算法。相对来讲，网络搜索、随机搜索及 BO 算法能够在有效的计算资源条件下，通过选择合理的参数调整范围和步长，有效探索参数空间。因此，选择这 3 种算法进行对比分析，为更好地对比 BO 算法的性能，将其分为基于高斯过程的 BO 算法、基于树结构 Parzen 估计器 (Tree-structured Parzen Estimator, TPE) 的 BO 算法、基于 Optuna 的 BO 算法 3 种，并以 R^2 和寻优时间作为评价指标。

由于表 4 中的超参数较多，所形成的超参数寻优空间较大，所以利用网格搜索在整个超参数空间寻优时，需耗费较多时间和资源。为尽可能降低资源消耗，同时考虑到基于高斯过程的 BO 算法不支持字符串寻优，在寻优过程中，首先通过试算确定超参数中 max_features, loss, criterion 的最佳选择。为了进行算法的统一评估，其余超参数寻优时，根据对参数区间及步长的控制，保持每次参数范围调整时总的参数空间规模相同。5 种寻优算法的寻优结果见表 4。可看出，基于 TPE 的 BO 算法得出的最优超参数准确率最高，其 R^2 最大，为 0.927 2，寻优时间仅为 2.36 min。

表 4　超参数寻优算法性能对比

Table 4　Performance comparison of hyperparameter optimization algorithms

超参数	网格搜索	随机搜索	基于高斯过程的BO算法	基于TPE的BO算法	基于Optuna的BO算法
max_features	log2	sqrt	log2	sqrt	sqrt
loss	absolute_error	absolute_error	absolute_error	quantile	absolute_error
criterion	friedman_mse	squared_error	squared_error	friedman_mse	squared_error
n_estimators	790	783	208	374	421
learning_rate	0.01	0.01	0.106 0	0.296 1	0.214 8
subsample	0.6	0.8	0.562 7	0.311 2	0.473 9
max_depth	6	5	33	2	42
min_impurity_split	0	0.888 9	0.059 4	2.525 0	3.350 1
R^2	0.903 4	0.901 6	0.926 6	0.927 2	0.926 6
寻优时间/min	108.9	4.42	8.63	2.36	4.58

4.2　模型预测结果分析

4.2.1　超参数寻优结果对比分析

为分析不同超参数组合对预测结果的影响，利用原始数据集中 60 组训练集，分别对基于不同超参数组合的 GBDT 模型进行训练，并分别对 13 组测试集进行预测，将预测值与真实值进行对比，对比曲线如图 3 所示。可看出基于不同超参数组合的 GBDT 模型所计算的预测值与真实值变化趋势相同，预测值之间数据差距较小，说明计算所得的超参数可使 GBDT 模型具有较好的准确性和泛化性。

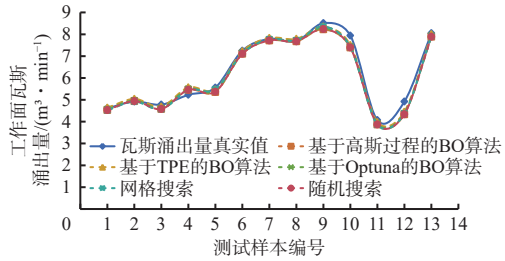

图 3　基于不同超参数组合的 GBDT 模型预测值与真实值对比

Fig. 3　Comparison of predicted and actual values in GBDT models under different hyperparameters combinations

为进一步对比分析预测值与真实值之间的差异，计算真实值与预测值之间的相对误差，如图 4 所示。对图 4 进行统计分析，得出最大相对误差为 11.79%，最小相对误差为 0，平均相对误差最大值为 3.53%，最小值为 2.70%，见表 5。

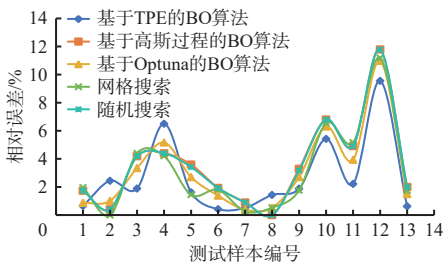

图 4　基于不同超参数组合的 GBDT 模型相对误差对比

Fig. 4　Comparison of relative errors in GBDT models under different hyperparameters combinations

表 5　基于不同超参数组合的 GBDT 模型相对误差统计

Table 5　Statistical relative errors in GBDT models under different hyperparameter combinations

超参数优化算法	最大相对误差/%	平均相对误差/%
网格搜索算法	11.17	3.15
随机搜索算法	11.78	3.51
基于高斯过程的BO算法	11.79	3.53
基于TPE的BO算法	9.55	2.70
基于Optuna的BO算法	10.97	3.13

由表 5 可看出，采用基于 TPE 的 BO 算法进行超参数优化后，GBDT 模型最大相对误差为 9.55%，平均相对误差为 2.70%，均小于其余 4 种算法。同时，基于 TPE 的 BO 算法寻优时间较短，优化性能较好。

4.2.2　特征选择对比分析

为分析不同特征选择对模型预测精度的影响，使用方差过滤法、F 检验法、互信息法、嵌入法、包装法选出的特征，以及基于包装法并考虑现场实际所选择的 8 个特征，通过基于 TPE 的 BO 算法进行超参数优化，再通过 GBDT 模型进行学习和预测。预测数据与真实数据对比曲线如图 5 所示。

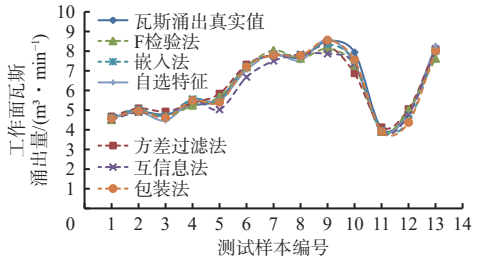

图 5　不同特征组合下 GBDT 模型预测值与真实值对比

Fig. 5　Comparison of predicted and actual values in GBDT models under different feature combinations

分析图 5 可知，针对同一数据集，不同特征组合下 GBDT 模型预测值与真实值之间变化趋势相同，可见特征数量与预测结果的准确性和泛化性之间并不呈正比关系，冗余特征或无关特征的存在反而会降低模型的预测准确性。

不同特征组合下 GBDT 模型相对误差曲线如图 6 所示，相对误差统计见表 6。分析图 6 和表 6 可知，采用包装法并自选特征时，GBDT 模型的平均相对误差和最大相对误差均最小，分别为 2.61%，7.18%。互信息法特征选择算法所得到的特征数量最多，为 13 项，但平均相对误差反而最大，为 4.30%，最大相对误差为 9.53%。这进一步说明特征数量与预测结果的准确性和泛化性并不呈正比关系。采用包装法并自选特征时，能够保证在较少特征数量的前提下取得较好的模型预测准确性。

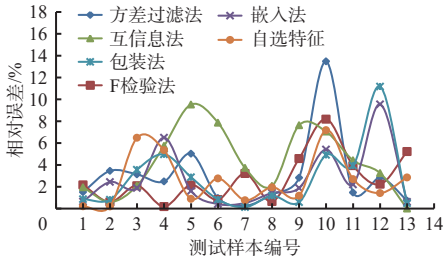

图 6　不同特征组合下 GBDT 模型的相对误差对比

Fig. 6　Comparison of relative errors in GBDT models under different feature combinations

表 6　不同特征组合下 GBDT 模型的相对误差统计

Table 6　Statistical relative errors in GBDT models under different feature combinations

特征选择方法	最大相对误差/%	平均相对误差/%
方差过滤法	13.48	3.04
F检验法	8.19	2.77
互信息法	9.53	4.30
嵌入法	9.55	2.70
包装法	11.18	2.79
包装法+自选	7.18	2.61

4.2.3　模型预测结果对比分析

为验证通过特征选择和超参数优化后 GBDT 模型的准确性和泛化性，利用 python 中的机器学习库分别建立随机森林、支持向量机及神经网络 3 种预测模型，进行对比分析，结果如图 7 所示。可看出 4 种模型的预测数据与真实数据均保持大致相同的趋势，而 GBDT 模型预测结果与真实值的曲线拟合度相对较高。

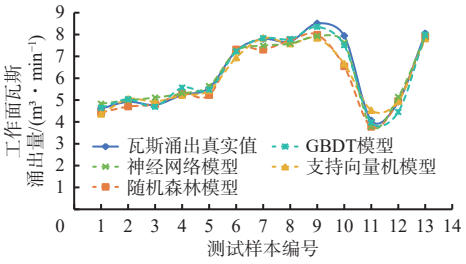

图 7　4 种模型的预测数据与真实数据对比

Fig. 7　Comparison of predicted data and actual values of four models

4 种预测算法的相对误差如图 8 所示，可看出 GBDT 模型和神经网络模型的相对误差变化趋势较为平缓，且均处于较低水平，而随机森林模型和支持向量机模型的相对误差变化幅度较大。

4 种模型的相对误差统计见表 7，可见 GBDT 模型的平均相对误差较随机森林模型、支持向量机模型、神经网络模型分别降低了 35.56%，37.41%，32.03%，具有较高的精确性。

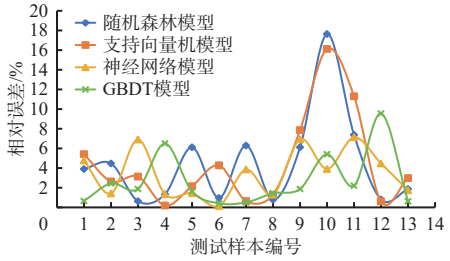

图 8　4 种模型的相对误差对比

Fig. 8　Comparison of relative errors of four models

表 7　4 种模型的相对误差统计

Table 7　Statistical relative error in four models

预测模型	最大相对误差%	平均相对误差%
随机森林模型	12.59	4.05
支持向量机模型	12.72	4.17
神经网络模型	8.11	3.84
GBDT模型	7.18	2.61

5　结论

1）对比分析了 5 种不同特征选择算法的性能，优选出包装法为最佳特征选择算法，同时结合现场实际，在包装法选择特征的基础上增加邻近层瓦斯涌出相关特征，优选出 8 个特征作为新的特征组合，有效减少了数据集中特征的重复、冗余，也可降低现场数据收集工作量。

2）对比 5 种超参数寻优算法对 GBDT 模型进行超参数寻优，结果表明，基于 TPE 的 BO 算法具有最高的准确率和相对较少的计算时间，其优化性能最佳。

3）不同超参数寻优算法所得到的超参数组合在 GBDT 模型的预测结果与真实值具有相同的变化趋势，预测值之间数据差距较小，表明寻优算法对 GBDT 模型的准确性和泛化性影响较小。

4）对比 5 种特征选择算法及自选特征的特征组合在 GBDT 模型下的预测结果，结果表明特征数量的多少与预测结果的准确性和泛化性并不呈正比关系，冗余特征或无关特征的存在反而会降低模型预测的准确性。

5）将随机森林、支持向量机和神经网络模型与 GBDT 模型进行对比，结果表明 GBDT 模型在工作面瓦斯涌出量预测中具有更高的准确性和泛化性，平均相对误差为 2.61%，相比随机森林、支持向量机、神经网络模型分别降低了 35.56%，37.41%，32.03%，能够很好地满足现场工程应用需求。

参考文献（References）：

[1] AQ 1018—2006 矿井瓦斯涌出量预测方法[S].
AQ 1018—2006 The predicted method of mine gas emission rate[S].

[2] 王磊，刘雨，刘志中，等. 基于 IABC−LSSVM 的瓦斯涌出量预测模型研究[J]. 传感器与微系统，2022，41(2)：34-38.
WANG Lei，LIU Yu，LIU Zhizhong，et al. Research on prediction model for gas emission based on IABC-LSSVM[J]. Transducer and Microsystem Technologies，2022，41(2)：34-38.

[3] 张玉财，王毅，郭凯岩. 基于 WOA−LSTM 的工作面瓦斯涌出量预测研究[J]. 矿业安全与环保，2023，50(5)：50-55.
ZHANG Yucai，WANG Yi，GUO Kaiyan. Research on prediction of gas emission in working face based on WOA-LSTM[J]. Mining Safety & Environmental Protection，2023，50(5)：50-55.

[4] 荣统瑞，侯恩科，夏冰冰. 基于二次分解和 BO−BiLSTM 组合模型的采煤工作面瓦斯涌出量预测方法研究[J]. 煤矿安全，2024，55(5)：83-92.
RONG Tongrui，HOU Enke，XIA Bingbing. Research on prediction method of coal mining face gas outflow based on quadratic decomposition and BO-BiLSTM combination model[J]. Safety in Coal Mines，2024，55(5)：83-92.

[5] 徐耀松，白济宁，王雨虹，等. 基于 CEEMDAN−DA−GRU 的瓦斯涌出量预测模型[J]. 传感技术学报，2023，36(3)：441-448.
XU Yaosong，BAI Jining，WANG Yuhong，et al. Prediction model of gas emission based on CEEMDAN-DA-GRU[J]. Chinese Journal of Sensors and Actuators，2023，36(3)：441-448.

[6] 刘鹏，魏卉子，景江波，等. 基于增强 CART 回归算法的煤矿瓦斯涌出量预测技术[J]. 煤炭科学技术，2019，47(11)：116-122.
LIU Peng，WEI Huizi，JING Jiangbo，et al. Predicting technology of gas emission quantity in coal mine based on enhanced CART regression algorithm[J]. Coal Science and Technology，2019，47(11)：116-122.

[7] 汪明，王建军. 基于随机森林的回采工作面瓦斯涌出量预测模型[J]. 煤矿安全，2012，43(8)：182-185.
WANG Ming，WANG Jianjun. Gas emission prediction model of stope based on random forests[J]. Safety in Coal Mines，2012，43(8)：182-185.

[8] 张增辉，马文伟. 基于随机森林回归算法的回采工作面瓦斯涌出量预测[J]. 工矿自动化，2023，49(12)：33-39.
ZHANG Zenghui，MA Wenwei. Prediction of gas emission in mining face based on random forest regression algorithm[J]. Journal of Mine Automation，2023，49(12)：33-39.

[9] 成小雨，周爱桃，郭焱振，等. 基于随机森林与支持向量机的回采工作面瓦斯涌出量预测方法[J]. 煤矿安全，2022，53(10)：205-211.
CHENG Xiaoyu，ZHOU Aitao，GUO Yanzhen，et al.

Prediction method of gas emission based on random forest and support vector machine[J]. Safety in Coal Mines, 2022, 53(10): 205-211.

[10] 陈茜, 黄连兵. 基于 LASSO−LARS 的回采工作面瓦斯涌出量预测研究[J]. 煤炭科学技术, 2022, 50(7): 171-176.
CHEN Qian, HUANG Lianbing. Gas emission prediction from coalface based on least absolute shrinkage and selection operator and least angle regression[J]. Coal Science and Technology, 2022, 50(7): 171-176.

[11] 徐刚, 王磊, 金洪伟, 等. 因子分析法与 BP 神经网络耦合模型对回采工作面瓦斯涌出量预测[J]. 西安科技大学学报, 2019, 39(6): 965-971.
XU Gang, WANG Lei, JIN Hongwei, et al. Gas emission prediction in mining face by factor analysis and BP neural network coupling model[J]. Journal of Xi'an University of Science and Technology, 2019, 39(6): 965-971.

[12] 吕伏, 梁冰, 孙维吉, 等. 基于主成分回归分析法的回采工作面瓦斯涌出量预测[J]. 煤炭学报, 2012, 37(1): 113-116.
LYU Fu, LIANG Bing, SUN Weiji, et al. Gas emission quantity prediction of working face based on principal component regression analysis method[J]. Journal of China Coal Society, 2012, 37(1): 113-116.

[13] 肖鹏, 谢行俊, 双海清, 等. 基于 KPCA−CMGANN 算法的瓦斯涌出量预测研究[J]. 中国安全科学学报, 2020, 30(5): 39-47.
XIAO Peng, XIE Xingjun, SHUANG Haiqing, et al. Prediction of gas emission quantity based on KPCA-CMGANN algorithm[J]. China Safety Science Journal, 2020, 30(5): 39-47.

[14] 王媛彬, 李媛媛, 韩骞, 等. 基于 PCA−BO−XGBoost 的矿井回采工作面瓦斯涌出量预测[J]. 西安科技大学学报, 2022, 42(2): 371-379.
WANG Yuanbin, LI Yuanyuan, HAN Qian, et al. Gas emission prediction of the stope in coal mine based on PCA-BO-XGBoost[J]. Journal of Xi'an University of Science and Technology, 2022, 42(2): 371-379.

[15] 陈巧军, 余浩, 李艳昌, 等. 基于 KPCA−LSSVM 的回采工作面瓦斯涌出量的预测[J]. 中国安全生产科学技术, 2024, 20(4): 78-84.
CHEN Qiaojun, YU Hao, LI Yanchang, et al. Prediction of gas emission quantity in mining face based on KPCA-LSSVM[J]. Journal of Safety Science and Technology, 2024, 20(4): 78-84.

[16] 胡坤, 王素珍, 韩盛, 等. 基于 TLBO−LOIRE 的回采工作面瓦斯涌出量预测[J]. 应用基础与工程科学学报, 2017, 25(5): 1048-1056.
HU Kun, WANG Suzhen, HAN Sheng, et al. Gas emission quantity prediction of working face based on TLBO-LOIRE method[J]. Journal of Basic Science and Engineering, 2017, 25(5): 1048-1056.

[17] 洪林, 赫祥林, 董晓雷, 等. PCA−GA−ELM 煤矿瓦斯涌出量预测[J]. 辽宁工程技术大学学报(自然科学版), 2015, 34(7): 779-784.
HONG Lin, HE Xianglin, DONG Xiaolei, et al. Prediction of mine gas emission based on PCA-GA-ELM[J]. Journal of Liaoning Technical University (Natural Science), 2015, 34(7): 779-784.

[18] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.
ZHOU Zhihua. Machine learning[M]. Beijing: Tsinghua University Press, 2016.

[19] 祝元丽, 冯向阳, 闫庆武, 等. 基于 GBDT 的望奎县农田土壤有机碳主控因子研究[J]. 中国环境科学, 2024, 44(3): 1407-1417.
ZHU Yuanli, FENG Xiangyang, YAN Qingwu, et al. Spatial distribution and main controlling factors of soil organic carbon under cultivated land based on GBDT model in black soil region of Northeast China[J]. China Environmental Science, 2024, 44(3): 1407-1417.

[20] 黄桂灶, 马同鑫, 杨泽锋, 等. 基于 GBDT 算法的弓网动态匹配特性预测模型[J]. 振动与冲击, 2024, 43(16): 26-32, 50.
HUANG Guizao, MA Tongxin, YANG Zefeng, et al. A study on prediction model of dynamic matching characteristics of pantograph-catenary system based on the GBDT algorithm[J]. Journal of Vibration and Shock, 2024, 43(16): 26-32, 50.

[21] SNOEK J, LAROCHELLE H, ADAMS R P. Practical Bayesian optimization of machine learning algorithms[C]. Annual Conference on Neural Information Processing Systems, Lake Tahoe, 2012: 2951-2959.

[22] 李海霞, 宋丹蕾, 孔佳宁, 等. 传统机器学习模型的超参数优化技术评估[J]. 计算机科学, 2024, 51(8): 242-255.
LI Haixia, SONG Danlei, KONG Jianing, et al. Evaluation of hyperparameter optimization techniques for traditional machine learning models[J]. Computer Science, 2024, 51(8): 242-255.

[23] 崔榕峰, 马海, 郭承鹏, 等. 基于贝叶斯超参数优化的 Gradient Boosting 方法的导弹气动特性预测[J]. 航空科学技术, 2023, 34(7): 22-28.
CUI Rongfeng, MA Hai, GUO Chengpeng, et al. Prediction of missile aerodynamic data based on Gradient Boosting under Bayesian hyperparametric optimization[J]. Aeronautical Science & Technology, 2023, 34(7): 22-28.