基于融合特征选择算法的钻速预测模型研究

周长春[1]，姜 杰[2]，李 谦[1]，朱海燕[3]，李之军[1]，鲁柳利[4]
（1.成都理工大学环境与土木工程学院，四川 成都 610059；2.成都理工大学机电工程学院，四川 成都 610059；
3.成都理工大学能源学院，四川 成都 610059；4.成都工业学院大数据与人工智能学院，四川 成都 611730）

摘要：钻速预测是钻井优化的重要组成部分，机器学习算法是当前实现准确钻速预测的重要手段，准确的特征选择是保证机器学习精度的关键途径。基于南海某井眼的实际钻井数据，本文采用一种融合特征选择法从钻井特征参数中选出井径、钻井液出口温度、钻井液入口密度、钻井液出口密度、K值、塑性粘度、滤失量、上覆压力、孔隙压力、和喷嘴等效直径共10种参数。将优选出的参数作为模型输入，引入集成的梯度提升树（Gradient Boosting Decision Tree，GBDT）算法建立机械钻速预测模型。将建立的模型与常规机器学习算法模型进行对比试验。试验结果显示，所提出的融合特征选择算法模型精度较全特征模型高2%，较常用机器学习模型平均高14.5%，该研究为钻井参数的准确、快速寻优提供了有效解决方案，对提高钻进速率具有一定的指导意义和实际应用价值。

关键词：钻速预测；机器学习；融合特征选择；梯度提升树算法（GBDT）
中图分类号：P634 **文献标识码**：A **文章编号**：2096-9686(2022)04-0031-10

Research on drilling rate prediction model based on fusion feature selection algorithm

ZHOU Changchun[1], JIANG Jie[2], LI Qian[1], ZHU Haiyan[3], LI Zhijun[1], LU Liuli[4]
(1.*College of Environment Civil Engineering, Chengdu University of Technology, Chengdu Sichuan* 610059, *China*;
2.*School of Mechanical and Electrical Engineering, Chengdu University of Technology, Chengdu Sichuan* 610059, *China*;
3.*College of Energy, Chengdu University of Technology, Chengdu Sichuan* 610059, *China*;
4.*School of Big Data and Artificial Intelligence, Chengdu Technological University, Chendu Sichuan* 611730, *China*)

Abstract: ROP prediction is an important part of drilling optimization, machine learning algorithms are currently an important means to achieve accurate ROP prediction, and correct feature selection is the key way to ensure machine learning accuracy. Based on the actual drilling data of a well in the South China Sea, this research uses a fusion feature selection method to select 10 drilling characteristic parameters, including well diameter, outlet temperature, inlet density, outlet density, K value, plastic viscosity, filtration loss, overburden pressure, pore pressure, and nozzle equivalent diameter. The optimized parameters are taken as model inputs, and the integrated Gradient Boosting Decision Tree (GBDT) algorithm is introduced to establish a ROP prediction model. The established model is compared with the conventional machine learning algorithm model, and the test results show that the accuracy of the proposed fusion feature selection algorithm model is 2% higher than that of the full feature model, and the average accuracy is 14.5% higher than that of the commonly used machine learning model. The research provides an effective solution for the accurate and rapid optimization of drilling parameters, and have guiding significance and practical

收稿日期：2022-04-25；**修回日期**：2022-06-17 **DOI**：10.12143/j.ztgc.2022.04.005
基金项目：中海石油（中国）有限公司项目“南海西部油田上产2000万方钻完井关键技术研究”子课题“乐东10区超高温高压井综合提速技术研究”（编号：CNOOC－KJ135ZDXM38ZJ05ZJ）；四川省科技支撑计划应用基础研究项目“四川深层页岩气产能大数据挖掘和智能评估方法研究”（编号：2021YJ0360）
第一作者：周长春，男，汉族，1995年生，硕士研究生在读，岩土工程专业，从事人工智能在钻探施工中应用的研究工作，四川省成都市成华区二仙桥东三路1号，zcc@stu.cdut.edu.cn.。
引用格式：周长春，姜杰，李谦，等.基于融合特征选择算法的钻速预测模型研究[J].钻探工程，2022，49(4)：31-40.
ZHOU Changchun, JIANG Jie, LI Qian, et al. Research on drilling rate prediction model based on fusion feature selection algorithm [J]. Drilling Engineering, 2022, 49(4): 31-40.

application value for improving the drilling rate.

Key words: ROP prediction; machine learning; fusion feature selection; Gradient Boosting Decision Tree(GBDT)

0 引言

我国能源生产重点方向正在向超深层发展,随着钻井的深度增加,钻头进入更加复杂的地层,会使施工难度加大、钻井速度减慢、成本升高。在国内外的研究中,机械钻速一直是作为钻井作业整体水平的直观反映,准确预测机械钻速可以有效计算钻井成本和钻井时间,从而优化钻井参数、合理安排钻机工作人员,并为钻井设计人员提供依据[1]。

传统的钻速预测研究中,一些研究人员考虑岩性、竖井直径和转速等作为主要因素,通过对多元化回归的分析,获得钻速方程[2]。还有一些研究人员制作模拟和动态模型,通过试验模拟钻探时的冲击强度来调整及预测钻速[1]。随着大数据及计算机技术的发展及其被应用到油气行业,采用机器学习技术对机械钻速进行预测已成为智能钻井行业研究的有效方法和重要手段[3]。如Amer等[4]将钻压、转速、排量、扭矩、泵量、泥浆密度和立管压力作为输入参数输入到基于人工神经网络的钻速预测模型。赵颖等[5]以南海YL8-3-1井为例,使用井眼深度、钻压、大钩位置、扭矩、出入口钻井液密度和温度等基于极限学习机建立了海上钻井机械钻速预测模型。对于特征选择方法的研究方面:李莉等[6]在特征选择阶段采用核主成分分析剔除源项目中的冗余数据的方法进行建模,结果表明所选择特征会使得建模精度有一定的提高。周翔等[7]提出了大数据环境下的投票特征选择算法可以有效解决特征选择问题。康文豪等[8]提出了一种双层特征选择法进行特征选择,其结果是所选特征使得预测模型有较好的拟合效果。此外,针对机械钻速预测研究,Dupriest等[9]强调了特征选择在建模过程中的重要性。Shi等[10]通过对钻头钻进机制进行研究确定了包括表面测量、钻头特性、水力学变量和地层特性等10个参数作为人工神经网络模型输入进行了研究。

综上,很多研究通过优化智能算法来提升模型精度,亦有很多研究者对大数据中特征选择方法进行了研究,然而专门针对机械钻速预测来完成特征选择部分的智能方法研究却相对较少。在进行钻速预测研究时,海量的钻井参数会耗费大量的计算资源和时间,且不易得到理想的模型精度,故亟需针对机械钻速特征选择进行专门研究。因此,本文提出一种融合特征选择法进行参数优选,再选用梯度提升树(Gradient Boosting Decision Tree,GBDT)算法进行钻速预测,并针对参数优选结果与预测精度设计对比试验进行验证。

1 基于融合特征选择钻速预测模型总体架构设计

本文先对采集到的数据进行整合预处理,然后基于设计的融合特征选择算法进行特征优选,最后针对特征优选结果建立GBDT钻速预测模型并设计对比试验进行验证,如图1所示。

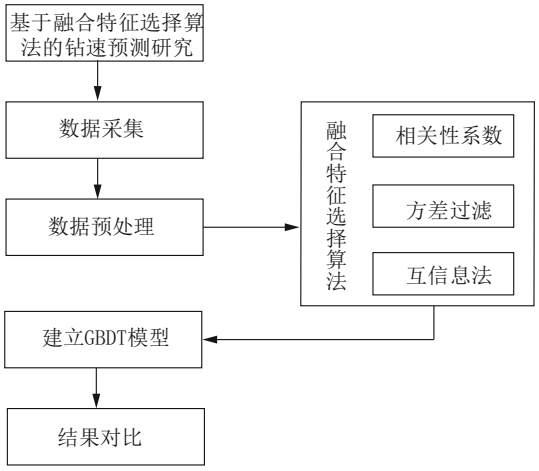

图1 融合特征选择算法钻速预测模型研究

Fig.1 Research on ROP prediction model with fusion feature selection algorithm

2 数据预处理

2.1 数据采集

令钻井参数数量为n,井深为D,不同的钻井参数采集时最大密度为d,则整合后的数据矩阵为一个D/d行×n列矩阵[11]。在本文所使用的南海某井眼钻井数据共5大类43种不同的参数共3967条,表1所示为参数缩写信息和参数分类信息。

2.2 数据清洗

数据清洗就是指利用数据分析将采集到的“脏数据”转化为符合要求的数据[12-13]。对于钻井“脏数据”的清洗过程包括异常值的检测、删除以及缺失数

表1　参数信息

Table 1　Parameter information

参数类型	参数名称/缩写
井眼参数	深度/D;井径/d
施工工艺	钻速/ROP;钻压/WOB;转速/RPM;泵量/Q;扭矩/T;泵压/SPP;钻时/BT;泵时/PT;大钩载荷/WOH;钻井液入口温度/TI;钻井液出口温度/TO;钻井液出口密度/MO;钻井液入口密度/MI
钻井液性质	屈服值/YP;密度/MW;漏斗粘度/MV;塑性粘度/PV;3转读数/D3*;6转读数/D6*;10 s静切力/SS;10 min静切力/SM;滤失量/FL;泥饼厚度/MT*;氯离子含量/CLC;钙离子含量/CAC*;含砂量/SA*;膨润土含量/SOC;固相含量/SO;pH值/pH*;流型指数/N;稠度系数/K
地质情况	地震速度/EV;孔隙压力/PP;破裂压力/FP;上覆压力/OP;岩性/TYP*
钻头参数	钻头内排磨损分级/WI*;钻头外排磨损分级/WO*;喷嘴等效直径/NS*;提速钻具使用/ST*;喷嘴数量/NN*

注:带*参数为离散型参数(本文将参数取值为点集的参数定义为离散型参数),其余为连续型参数

据的插值补全。观察采集到的3697条原始数据,发现前面的967条数据中有大量参数未采集到,因此判定为无效数据,采用删除策略后剩余3000条数据。由于所采集数据缺失部分为离散值,因此采用k近邻填补法(KNN),即计算欧几里得空间中每个样本点与被填补点的距离,选出k个距离最近的样本点的类别,采用投票法决定填补值,距离计算采用欧式距离,计算式如式(1)所示[14]。

$$d(x,y)=\sqrt{\sum_{i=1}^{N}(x_{i1}-x_{i2})^2} \tag{1}$$

式中:d——欧式距离;N——N维空间;x_{i1}——第1个点的第i维坐标;x_{i2}——第2个点的i维坐标。

2.3　数据标准化处理

补齐数据之后,由于参数数据间较大的量纲差距会给后续的机器学习建模的模型性能造成隐患,因此需要对数据做标准化处理来缩小量纲差距,其计算式如式(2)所示[15]。

$$x_{\text{new}}=\frac{x_{\text{old}}-\mu}{\sigma_{x\text{list}}} \tag{2}$$

式中:x_{new}——完成标准化的数据;x_{old}——标准化前的原始数据;μ——平均值;$\sigma_{x\text{list}}$——原始数据同一变量所有数据标准差。

以钻压和钻井液出口温度为例,标准化处理之后效果展示如图2所示。

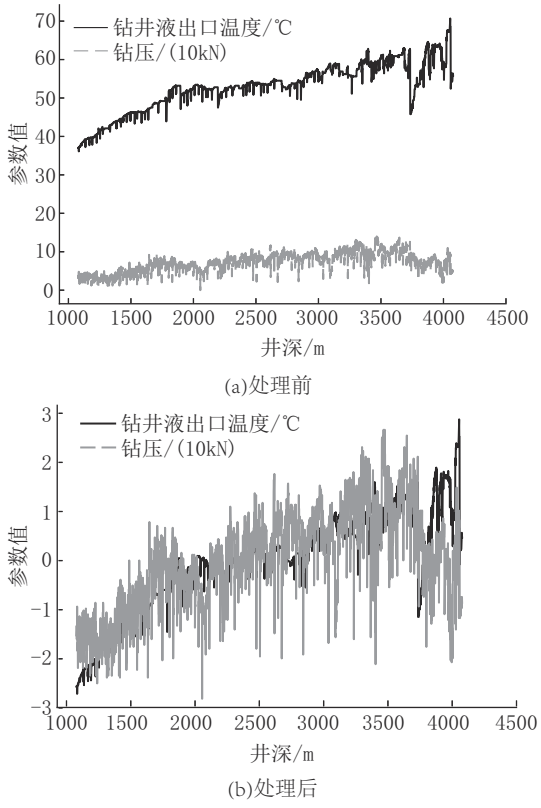

图2　标准化处理前后对比

Fig.2　Comparison before and after standardization

3　融合特征选择算法设计

3.1　相关性分析

相关性分析的主要目的在于判定输入与输出变量之间的相关性以指导建模时下一步该采取何种操作,本文采用皮尔逊相关系数计算方法对所选变量进行相关性分析,筛选出高相关性参数组作为特征选择工作的第一步,计算方法如式(3)所示[16]。

$$\rho_{ab}=\frac{\operatorname{cov}(a,b)}{\sigma_a\sigma_b}=\frac{\sum_{i=1}^{n}(a_i-\bar{a})(b_i-\bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i-\bar{a})^2}\sqrt{\sum_{i=1}^{n}(b_i-\bar{b})^2}} \tag{3}$$

式中：ρ_{ab}——a、b变量之间的相关性；cov(a,b)——变量a、b的协方差矩阵；σ_a、σ_b——变量a,b各自的标准差；a_i、b_i——变量a、b数据集中第i个变量值；$\bar{a}$、$\bar{b}$——变量a、b平均值；n——变量a、b的数据集大小。

ρ_{ab}的取值在区间[－1,1]上，取值为正时，表示两个参数之间呈现正的相关性，反之则表示两个参数呈负相关性，ρ_{ab}的绝对值越靠近1，说明a、b之间的相关性越高，越靠近0，则说明两个变量之间的相关性越低，计算表1中钻速ROP参数与除钻速之外的所有其他参数之间的相关性，计算结果如图3、图4所示。

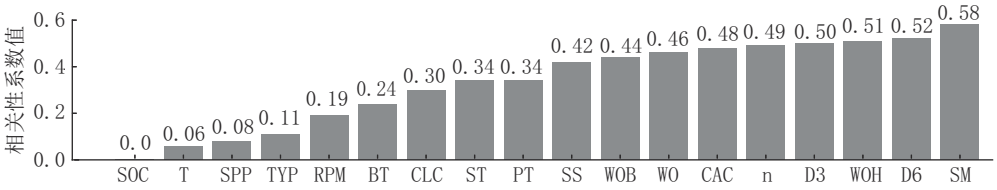

图3 低、中相关性参数组

Fig.3 Low and medium correlation parameter groups

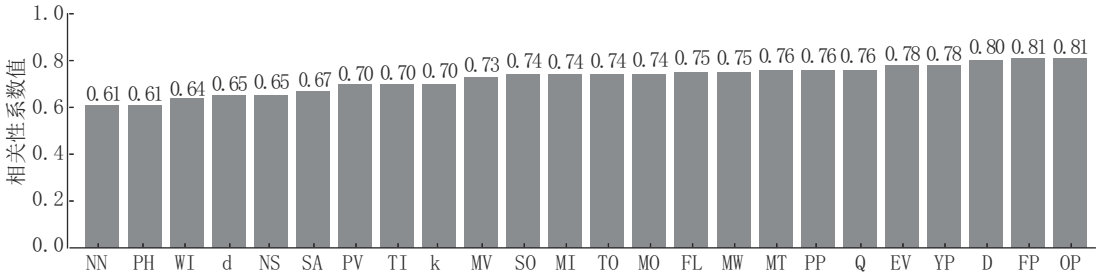

图4 高相关性参数组

Fig.4 High correlation parameter group

对计算结果进行统计，可按照皮尔逊相关性系数将除钻速之外的其他参数与钻速的相关性分为高相关性、中相关性和低相关性3类[16]。

(1)高相关性参数：总共有24种，占所有参数的55.81%，该类参数与钻速的相关性系数计算结果的绝对值均位于[0.6,0.81]区间内。

(2)中相关性参数：总共有15种，占所有参数的34.88%，该类参数与钻速的相关性系数计算结果的绝对值均位于[0.1,0.6]区间内。

(3)低相关性参数：总共有3种，占所有参数的9.31%，该类参数与钻速的相关性系数计算结果的绝对值均小于0.1。

从相关性系数计算结果可以看到传统经验中如岩性等参数的相关性系数取值较低，这是因为皮尔逊相关性分析对线性相关的参数更为敏感，更容易选出线性关系更明显的特征，因此传统钻速研究中非线性相关的参数相关性系数值会相对较低。

3.2 方差过滤

在机器学习建模过程中，引入的参数相关性越高，建立高精度机器学习预测模型所需要的参数数量越少[17]。因此，使用方差过滤法选择少量的包含更多信息量的参数，以提升模型的效率和精度。其原理是对于离散型特征，对方差进行计算，然后按计算结果保留贡献较大的特征。其操作步骤是先对离散型特征参数进行方差计算，观察计算结果发现，特征方差以岩性(TYP)为界呈明显的两级分布，因此以TYP方差2.6157为阈值，选择方差大于和等于阈值的特征，方差计算结果如表2所示。

3.3 互信息法

离散型特征选择结束之后，用互信息法从30个连续型参数中选出特征量相对较少且互信息估量较高的参数组，互信息定义如式(4)所示，其估计量取值区间位于[0,1]，其值越大，表明变量与标签之间的相关性越大[18]。

表2　离散型参数方差

Table 2　Discrete parameter variance

参数	方差	参数	方差
SA	0.0018	PH	0.4942
MT	0.0082	TYP*	2.6157
WO	0.1016	NS*	8.1302
ST	0.2255	D3*	9.1330
NN	0.2483	D6*	11.9211
WI	0.4627	CAC*	173.0132

注:带*号特征为方差过滤法选择结果,其余为被过滤参数

$$I(X;Y)=\iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\mathrm{d}x\mathrm{d}y \quad (4)$$

式中:$p(x,y)$——X与Y的联合概率分布;$p(x)$、$p(y)$——边缘概率分布。

操作步骤是先对30个连续型特征进行离散化处理,然后计算出每一个参数的互信息估计量并排序,计算结果如表3所示,最后利用前向搜索策略结合模型后验法,即依次向模型输入特征,每输入一个特征对模型进行一次评价,当模型性能提升时则选择当前特征,当模型性能下降则过滤掉特征。前向搜索过程如图5所示,图中折线上三角点对应参数为互信息法结合前向搜索策略选择特征参数,其余点对应参数为被过滤参数。

表3　互信息量估计量

Table 3　Mutual information estimator

参数	互信息量	参数	互信息量	参数	互信息量
WOB*	0.2540	SPP	0.5371	CLC*	0.6918
T	0.2924	SM	0.5544	SO	0.6966
BT	0.3107	K*	0.5585	MI*	0.7096
PT	0.3311	EV	0.5868	MO	0.7146
SOC	0.3552	RPM*	0.5934	MW*	0.7220
d*	0.4053	MV	0.6185	FP	0.7269
n*	0.4471	YP	0.6198	Q	0.7357
SS	0.4643	PV*	0.6252	OP*	0.7382
TI	0.4881	WHO	0.6269	D	0.7427
TO*	0.5157	FL*	0.6518	PP*	0.7563

注:带*特征参数为互信息法前向搜索策略特征选择结果,其余参数为被过滤参数

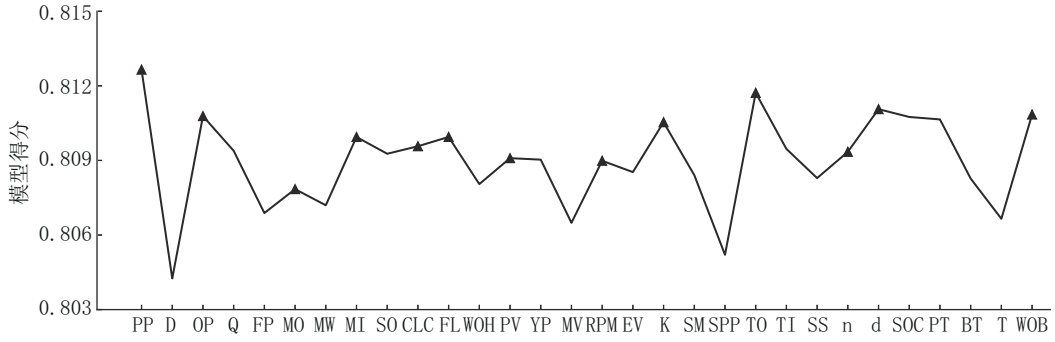

图5　基于前向搜索的互信息特征筛选

Fig.5　Mutual information feature screening based on forward search

3.4　融合特征选择算法步骤及评价

融合皮尔逊相关性分析法、方差过滤法和互信息法进行特征选择,其操作步骤如图6所示。

操作可分为4步:

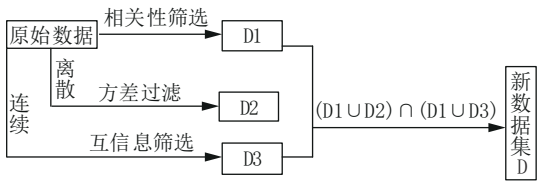

图6　特征选择过程示意

Fig.6　Schematic diagram of the feature selection process

(1)对经清洗之后的数据进行皮尔逊相关性计算,按照皮尔逊相关性原理将所有特征参数划分为高相关性参数组、中相关性参数组和低相关性参数组,然后选择与钻速具有高相关性的高相关性参数组作为特征选择的融合算法的第一步选择;

(2)将所有特征参数中的离散类型参数按照方差过滤法原理进行方差过滤,然后选择方差值高的特征参数作为特征选择的融合算法的第二步选择;

(3)将所有特征参数中连续类型参数按照互信息法计算原理进行互信息估计量计算并按互信息估量值的大小进行排序,然后使用前向搜索策略结合

模型验证来进一步进行特征筛选。

(4)将通过相关性过滤结果的参数组分别与方差过滤结果参数组和互信息过滤参数组结果分别取交集,最后将2个交集参数组取并集作为特征选择的融合算法的最终选择结果,它们与钻速的相关性系数、方差及互信息量如表4所示。

表4 融合特征选择算法特征选择结果

Table 4 Feature selection results with fusion feature selection algorithm

参数类型	参数	相关性系数	互信息量	方差
连续型参数	PP	0.76	0.7563	
	OP	0.81	0.7382	
	MW	0.75	0.7220	
	MI	0.74	0.7096	
	FL	0.75	0.6518	
	PV	0.70	0.6252	
	K	0.70	0.5585	
	TO	0.74	0.5157	
	d	0.65	0.4053	
离散型参数	NS	0.65		8.1302

在设计的融合特征选择算法中,利用皮尔逊相关性系数方法和方差过滤方法能够有效去除数据中的无关特征,使得模型的输入参数间会存在较大耦合。因此进行的第三步操作:将互信息法与前向搜索策略结合能够有效剔除部分相互耦合的特征。

4 基于融合特征选择结果的GBDT钻速预测模型

4.1 GBDT算法模型介绍

GBDT算法属于集成学习算法的一种,它融合了装袋法(Bagging)与提升法(Boosting)的思想,由Firedman在2001年提出,既可用来解决分类问题,也可用来解决回归问题[19]。GBDT算法由多个基学习器$f(x)$、残差构成的损失函数$L(x,y)$以及加法集成策略$H(x)$构成,其原理如图7所示,为方便展示,图中用虚线框表示多个基学习器及其预测结果。

GBDT算法的基学习器由决策树组成,单棵决策树的结构越复杂,GBDT算法的整体复杂度也会更高,使得计算缓慢且易过拟合。

$$f_0(x)=\arg\min_{\alpha}\sum_{i=1}^{m}L(y_i,\alpha) \tag{5}$$

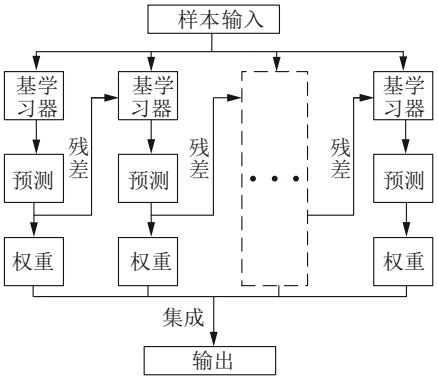

图7 GBDT算法原理示意

Fig.7 Schematic diagram of GBDT algorithm principle

式中:L——损失函数用来度量预测值与真实值之间的误差;m——样本个数;$\arg\min_{\alpha}$——损失函数取得最小值时,计算α取值;f_0——基学习器。

选择平方误差(squared_error)作为GBDT算法的损失函数,因为此函数一阶导数连续,易于被优化,是一个鲁棒的损失函数,式(6)为其计算表达式:

$$L[y_i,f(x_i)]=\frac{1}{2}\sum[y_i-f(x_i)]^2 \tag{6}$$

式中:$L[y_i,f(x_i)]$——损失函数;y_i、$f(x_i)$——分别为每个样本(x_i,y_i)的真实值和拟合值。

在此基础上,将损失值的负梯度作为残差估计值,利用梯度提升技术对残差进行拟合:

$$R_{ik}=-\frac{\partial L[y_i,f(x_i)]}{\partial f(x_i)}=y_i-f(x_i) \tag{7}$$

式中:R_{ik}——残差估计值;k——第k($k=1,2,\cdots\cdots K$)次迭代。

GBDT算法对基学习器进行集成时遵循的原则是依据上一个基学习器$f_{k-1}(x)$的结果,计算损失函数$L(y_i,f(x_i))$,并使用损失函数自适应的影响下一个基学习器$f_k(x)$的构建,集成模型的输出结果。其操作步骤是先确定每个叶节点区域对应损失函数最小化的最佳拟合值ε_{ik},然后更新学习器$f_k(x)$,最终构建GBDT模型如式(8)所示[19]。

$$\left.\begin{aligned}\varepsilon_{ik}&=\arg\min_{\alpha}\sum_{x_i\in C_{ik}}[y_i-f_{k-1}(x_i)-\varepsilon]^2\\ f_k(x)&=f_{k-1}(x)+\sum_{i}^{I}\varepsilon_{ik}\eta\\ H(x)&=f_0(x)+\sum_{k=1}^{K}\sum_{i=1}^{I}\varepsilon_{ik}\eta\end{aligned}\right\} \tag{8}$$

式中:η——学习率;C_{ik}($i=1,2,\cdots\cdots I$)——得到的

第k棵树的叶节点区域;ε_{ik}——每个叶子点区域确定使对应损失函数最小化的最佳拟合值;$H(x)$——GBDT模型最终拟合结果。

4.2　模型设计

导入经融合特征选择算法所确定的特征参数进行机器学习建模,采用10折交叉验证法降低模型过拟合风险,使用决定系数(R^2)、均方根误差($RMSE$)和相对误差($MAPE$)等指标对模型进行评估,部分数据展示如表5所示。

表5　模型输入部分数据

Table 5　Some model input data

PP	OP	MW	MI	FL	PV	k	TO	d	NS	ROP
0.999532	2.032095	1.06	1.05	7.8	9	0.1	37	17.5	45.25	46.03
0.999532	2.03227	1.06	1.05	7.8	9	0.1	37	17.5	45.25	49.53
0.999532	2.032445	1.06	1.05	7.8	9	0.1	37.1	17.5	45.25	50.3
0.999532	2.03262	1.06	1.05	7.8	9	0.1	37	17.5	45.25	43.73
0.999532	2.032795	1.06	1.05	7.8	9	0.1	37	17.5	45.25	32.6

4.2.1　10折交叉验证

将数据集等比例划分成10份,以其中的一份作为测试数据,其余9份作为训练数据,每次试验选取不同的测试集,剩下的作为训练集,重复进行10次试验,最后把10次测试集得分平均作为最终得分,其原理如图8所示[20]。

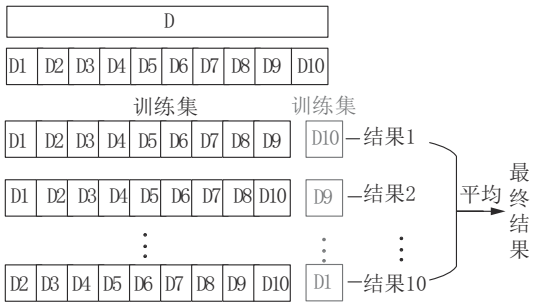

图8　10折交叉验证原理示意

Fig.8　Schematic diagram of the 10-fold cross-validation principle

4.2.2　模型评估

4.2.2.1　决定系数(R^2)

决定系数是指回归直线对观测值的拟合程度,R^2越接近1,表明拟合程度越好[20]。其计算式为:

$$R^2=1-\left[\sum_{i=1}^{n}(y_i-\hat{y}_i)^2/\sum_{i=1}^{n}(y_i-\bar{y})^2\right] \quad (9)$$

式中:y_i——真实值;$\bar{y}$——真实平均值;$\hat{y}_i$——预测值。

4.2.2.2　均方根误差($RMSE$)和相对误差($MAPE$)

均方根误差是预测值与真实值偏差的平方和的均值的平方根,其计算式如式(10)所示;相对误差是指误差与真实值的百分比,其计算式如式(11)所示,它能够表示预测值的可信程度[20]。二者均能表示预测值与真实值的偏离程度,其取值越接近于0,表示模型的性能越好,预测精度越高。

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \quad (10)$$

$$MAPE=(1/n)\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right|\times 100\% \quad (11)$$

10次试验的评分如表6所示,R^2最高能达到0.88的预测精度,平均达到0.85的精度。从误差的角度来看,平均均方根误差为4.57,平均相对误差为16%,表明模型预测精度较好,预测偏差较小,能够在一定程度上对机械钻速进行准确预测。

表6　GBDT模型下10折交叉验证试验R^2

Table 6　10-fold cross-validation test R^2 under GBDT model

次数	1	2	3	4	5	6	7	8	9	10	平均值
R^2	0.82	0.88	0.88	0.85	0.82	0.87	0.88	0.82	0.78	0.85	0.85
$RMSE$	4.75	4.13	3.69	4.78	5.22	4.29	3.95	4.82	5.74	4.29	4.57
$MAPE$/%	18	12	15	16	17	16	13	18	18	14	16

为了展示预测结果与真实值的拟合关系，提取出10次测试集的预测值绘制回归直线拟合关系图，如图9所示。此时R^2为0.85，*RMSE*和*MAPE*分别为4.57和16%，可以观察到所有的数据都分布在拟合线的周围，表明模型有不错的预测精度。

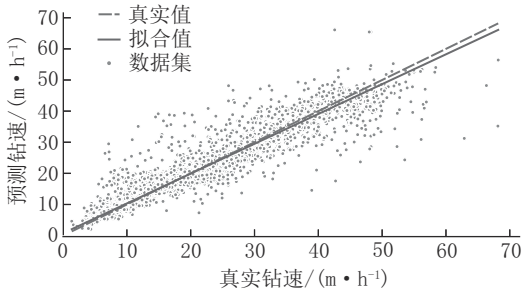

图9 GBDT预测真实值拟合关系

Fig.9 Fitting relationship between GBDT predictions and true values

取10折交叉验证时划分为10部分数据中的第1部分和第2部分测试集的预测值和真实值对比，绘制GBDT模型预测值和真实值的关系图(图10)，可以看到钻速预测值与真实值吻合，同样表明模型的拟合效果较好。

4.3 对比试验

为验证融合特征选择算法在预测性能上的优势以及GBDT模型相较于传统机器学习算法模型的优势，建立全特征GBDT模型，并与特征选择结果的常用机器学习算法模型进行对比试验。

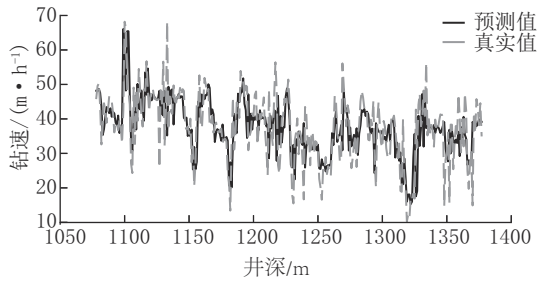

图10 钻速预测值与真实值对比

Fig.10 Comparison between the predicted ROP and the actual ROP

4.3.1 全特征模型

选择所有特征，使用10折交叉验证法，建立GBDT模型，通过比较模型在测试集上的各评估指标，发现使用全部特征作为模型输入时，模型在测试集上的泛化能力R^2得分为0.83，*RMSE*和*MAPE*得分分别为4.81和19%，融合特征选择结果建模与之相比，R^2提升了2%，而*RMSE*和*MAPE*分别降低了0.24和3%，如表7所示。图11为每个测试集的3个模型评估指标得分，可见经过特征选择得分均优于由全部特征所建立的模型，表明融合特征选择算法能为提高模型精度做出贡献。

表7 模型评估指标

Table 7 Model evaluation metrics

对比类型	R^2	*RMSE*	*MAPE*/%
特征选择	0.85	4.57	16
全部特征	0.83	4.81	19

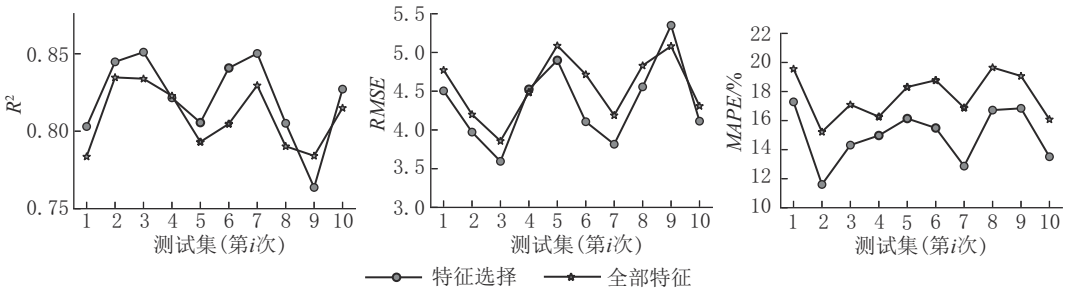

图11 全特征模型与特征选择模型测试集得分对比

Fig.11 Comparison of test set scores between the full feature model and the feature selection model

4.3.2 传统机器学习模型

选择适用于高维特征计算的支持向量回归、人工神经网络中具有代表性的BP神经网络回归、适用于处理线性关系的线性回归以及树模型的基础决策树回归算法结合10折交叉验证进行对比试验，各模型平均得分如表8所示，与GBDT模型相比，GBDT模型的R^2分别比支持向量回归、BP神经网络回归、线性回归和决策树回归高22%、18%、16%和

7%,*RMSE* 分别低了 2.44、2.01、1.92 和 0.85,*MAPE*分别低了17%、14%、13%和1%。

10个测试集各模型评估指标对比如图12所示。

表8　不同机器学习算法模型评估平均得分

Table 8　Average evaluation scores of different machine learning algorithm models

模　型	R^2	*RMSE*	*MAPE*/%
GBDT回归	0.85	4.57	16
支持向量回归	0.63	7.01	33
BP神经网络回归	0.67	6.58	30
线性回归	0.69	6.49	29
决策树回归	0.78	5.42	17

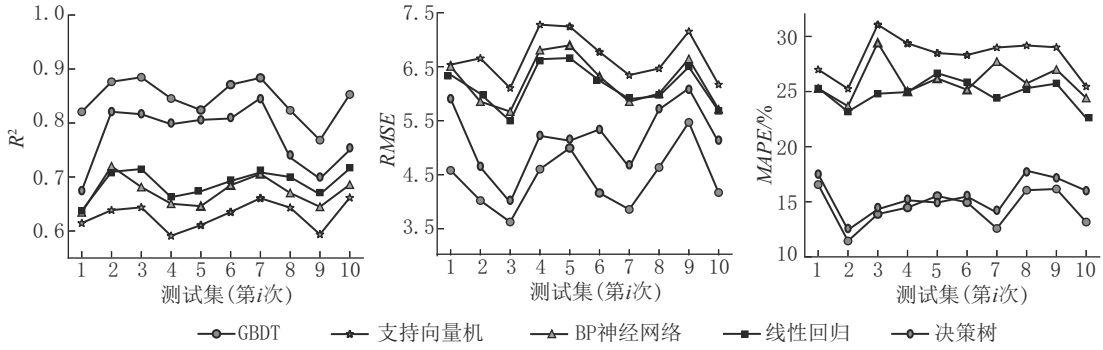

图12　GBDT模型与常见机器学习算法模型测试集对比

Fig.12　Comparison of the test sets between the GBDT model and the common machine learning algorithm model

(1)针对钻速预测机器学习建模之前特征的选择,本文提出的融合特征选择算法能够准确地从大量特征参数中选择出对模型贡献最大的参数,从而降低特征空间的维度,与使用全部特征所建立的模型相比,经过融合特征选择算法选择的特征参数所建立的模型的精度优于使用全部特征所建模型的精度,表明融合特征选择算法能够为机械钻速准确预测选择出合适的参数,且该算法能够为智能钻井机械钻速预测提供科学依据。

(2)本文所建立的梯度提升回归树模型在测试集上能够达到85%的精度,即表明模型有较好的泛化性能,能够较好地拟合机械钻速,与常用的机器学习算法相比,GBDT算法模型的决定系数R^2均高于常用算法模型,且均方根误差*RMSE*和相对误差*MAPE*均低于常用算法模型,表明GBDT模型预测性能比传统机器学习模型更具优势,也说明GBDT模型在未知数据上具有更好的泛化能力。

(3)本文所融合的多种特征选择方法能够有效剔除数据中的无关特征,但并不能解决参数间的耦合问题,因此本文在融合的方法中结合了前向搜索策略,能够在一定程度上减少参数间的耦合。不足之处在于该算法侧重于对具有物理意义的参数进行选择,因此并没有针对最终的特征选择结果进行特征信息研究,将来的研究中可对此进一步优化。

试验结果表明,与常用机器学习算法相比,GBDT算法模型的R^2均高于常用算法模型且*RMSE*和*MAPE*均低于常用算法模型,说明在此井眼中,GBDT模型对机械钻速的拟合效果更好,在测试集上具有更好的泛化性能。

5　结论

准确的机械钻速预测是提高钻进效率、降低钻井成本的重要手段。本文以南海某井眼钻井数据为例,融合相关性分析、方差过滤、互信息法并结合前向搜索策略进行特征选择,然后建立GBDT模型对机械钻速进行预测,主要结论如下:

参考文献(References):

[1]　于洋,黄凯,李卉.基于机器学习和多源数据预处理技术的机械钻速预测方法研究[J].中国石油和化工标准与质量,2021,41(20):133-136.
YU Yang, HUANG Kai, LI Hui. Research on prediction method of ROP based on machine learning and multi-source data preprocessing technology[J]. China Petroleum and Chemical Standard and Quality, 2021,41(20):133-136.

[2]　Barbosa L F F M, Nascimento A, Mathias M H, et al. Machine learning methods applied to drilling rate of penetration prediction and optimization—A review[J]. Journal of Petroleum Science and Engineering, 2019,183:106332.

[3] 张维罡.基于机器学习算法的石油钻速研究[J].化工管理,2021(20):89-90.
ZHANG Weigang. Research on ROP of petroleum based on machine learning algorithm[J]. Chemical Management, 2020(20):89-90.

[4] Amer M M, Dahab A S, El-Sayed A A H. An ROP predictive model in nile delta area using artificial neural networks[C]// SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition. OnePetro, 2017.

[5] 赵颖,孙挺,杨进,等.基于极限学习机的海上钻井机械钻速监测及实时优化[J].中国海上油气,2019,31(6):138-142.
ZHAO Ying, SUN Ting, YANG Jin, et al. Extreme learning machine-based offshore drilling ROP monitoring and real-time optimization[J]. China Offshore Oil and Gas, 2019, 31(6):138-142.

[6] 李莉,石可欣,任振康.基于特征选择和TrAdaBoost的跨项目缺陷预测方法[J].计算机应用,2022,42(5):1554-1562.
LI Li, SHI Kexin, REN Zhenkang. Cross-project defect prediction method based on feature selection and TrAdaboost[J]. Journal of Computer Applications, 2022,42(5):1554-1562.

[7] 周翔,翟俊海,黄雅婕,等.大数据环境下的投票特征选择算法[J/OL].小型微型计算机系统,2022:1-9.
ZHOU Xiang, ZHAI Junhai, HUANG Yajie, et al. Voting feature selection algorithm in big data environment[J/OL]. Journal of Chinese Computer Systems, 2022:1-9.

[8] 康文豪,徐天奇,王阳光,等.双层特征选择和CatBoost-Bagging集成的短期风电功率预测[J/OL].重庆理工大学学报(自然科学),2022:1-8.
KANG Wenhao, XU Tianqi, WANG Yangguang, et al. Short-term wind power prediction based on double-layer feature selection and catboost-bagging integration[J/OL]. Journal of Chongqing University of Technology(Natural Science), 2022:1-8.

[9] Dupriest F E, Koederitz W L. Maximizing drill rates with real-time surveillance of mechanical specific energy[C]//SPE/IADC Drilling Conference. OnePetro, 2005.

[10] Shi X, Liu G, Gong X, et al. An efficient approach for real-time prediction of rate of penetration in offshore drilling[J]. Mathematical Problems in Engineering, 2016:20-16.

[11] 李谦,周长春,朱海燕,等.生产数据的整合与初步分析在钻井中的应用实例[J].钻探工程,2021,48(S1):97-107.
LI Qian, ZHOU Changchun, ZHU Haiyan, et al. Application of integration and preliminary analysis of production data in drilling[J]. Drilling Engineering, 2021,48(S1):97-107.

[12] 李洪烈,夏栋,王倩.基于回归模型的采集数据清洗技术[J].电光与控制,2022,29(4):117-120.
LI Honglie, XIA Dong, WANG Qian. A sample data clean technology based on regression model[J]. Electronics Optics & Control, 2022,29(4):117-120.

[13] 匡俊搴,赵畅,杨柳,等.一种基于深度学习的异常数据清洗算法[J].电子与信息学报,2022,44(2):507-513.
KUANG Junqian, ZHAO Chang, YANG Liu, et al. An outlier cleaning algorithm based on deep learning[J]. Journal of Electronics & Information Technology, 2022,44(2):507-513.

[14] 曹凯鑫,汤猛猛,葛建鸿,等.大气污染物PM_(2.5)缺失数据插值方法的比较研究:基于北京市数据[J].环境与职业医学,2020,37(4):299-305.
CAO Kaixin, TANG Mengmeng, GE Jianhong, et al. Comparison of methods to interpolate missing PM2.5 values: Based on air surveillance data of Beijing[J]. Journal of Environmental and Occupational Medicine, 2020,37(4):229-305.

[15] 王双敬,王玉杰,李旭,等.TBM掘进数据标准化预处理方法研究[J/OL].现代隧道技术,2022:1-8.
WANG Shuangjing, WANG Yujie, LI Xu, et al. Research on standardized preprocessing method of TBM tunneling data[J/OL]. Modern Tunnelling Technology, 2022:1-8.

[16] 屈峰涛.基于大数据和智能算法的钻井参数优选模型与应用研究[D].西安:西安石油大学,2021.
QU Fengtao. Research on establishment and application of drilling parameter optimization model based on big data and intelligent algorithms[D]. Xi'an: Xi'an Shiyou University, 2021.

[17] 李谦,曹彦伟,朱海燕.基于人工智能的钻速预测模型数据有效性下限分析[J].钻探工程,2021,48(3):21-30.
LI Qian, CAO Yanwei, ZHU Haiyan. Discussion on the lower limit of data validity for ROP prediction based on artificial intelligence[J]. Drilling Engineering, 2021,48(3):21-30.

[18] 殷豪,翟广松,王鹏,等.基于互信息特征选取-变分模态分解和长短时记忆网络的日前耦合市场电价预测[J/OL].电网技术,2022:1-9.
YIN Hao, ZHAI Guangsong, WANG Peng, et al. Electricity price forecast of day-ahead coupled market based on mutual information feature selection and variational mode decomposition and LSTM[J/OL]. Power System Technology, 2022:1-9.

[19] 陈陆,吴桦.基于GBDT的船舶油耗预测模型设计[J].电子设计工程,2022,30(2):91-95.
CHEN Lu, WU Ye. Prediction model of ship fuel consumption based on GBDT[J]. Electronic Design Engineering, 2022,30(2):91-95.

[20] 吕晓玲,宋捷.大数据挖掘与统计机器学习[M].北京:中国人民大学出版社,2016.
SONG Xiaoling, SONG Jie. Big Data Mining and Statistical Machine Learning[M]. Beijing: China Renmin University Press, 2016.

(编辑 李艺)