基于粒子群优化的融合特征选择钻速预测模型研究

胥知画[1]，姜　杰[1*]，周长春[2]，李　谦[3]，任　军[1]
（1.成都理工大学机电工程学院，四川 成都 610059；2.成都环境工程建设有限公司，四川 成都 610000；3.成都理工大学环境与土木工程学院，四川 成都 610059）

摘要：传统的钻速预测模型经常受到数据维度过高和特征相关性等问题的制约，导致钻速预测效率和精度受限。为了解决这些问题，本文提出了一种基于粒子群优化（PSO）的融合特征选择钻速预测算法模型。在数据预处理的基础上，首先以3个关键参数threshold_1、threshold_2和threshold_3为优化目标，通过结合历史数据和粒子群优化算法构建适应度函数，从而建立钻速预测模型。接着，使用实际钻井数据对所提出的钻速预测方法进行验证，并与传统的机器学习算法模型进行对比实验。实验结果表明，所提出的粒子群优化融合特征选择算法在特征选择方面具有更高的效率和准确性，使用优化后的融合特征优选结果所训练的4个机器学习钻速预测模型精度相较于优化前分别提升了59%、1%、3%和1%，相较于使用全部特征所训练的模型分别提升了24%、2%、4%和3%。本文为钻井工程中提取到的特征参数过多时提供了一种有效的特征选择方法，对特征选择算法在工程领域的实际应用具有一定的指导意义。

关键词：钻速预测模型；特征选择；粒子群优化；机器学习
中图分类号：P634　　**文献标识码**：A　　**文章编号**：2096-9686(2025)02-0134-10

Research on a rate of penetration (ROP) prediction model based on feature selection integrated with particle swarm optimization (PSO)

XU Zhihua[1], JIANG Jie[1*], ZHOU Changchun[2], LI Qian[3], REN Jun[1]
(1.*School of Mechanical and Electrical Engineering, Chengdu University of Technology, Chengdu Sichuan* 610059, *China*; 2.*Chengdu Environmental Engineering Construction Co., Ltd, Chengdu Sichuan* 610000, *China*; 3.*College of Environment and Civil Engineering, Chengdu University of Technology, Chengdu Sichuan* 610059, *China*)

Abstract: Traditional rate of penetration (ROP) prediction models have often been constrained by issues such as high data dimensionality and feature correlation, resulting in limited efficiency and accuracy of ROP prediction. To address these issues, a ROP prediction algorithm model based on particle swarm optimization (PSO) with integrated feature selection has been proposed in this paper. Based on data preprocessing, 3 key parameters, threshold_1, threshold_2, and threshold_3, have been chosen as optimization targets, and a fitness function has been constructed by combining historical data and the PSO algorithm, thereby establishing the ROP prediction model. Subsequently, the proposed ROP prediction method has been validated using actual drilling data and compared with traditional machine learning algorithm models. Experimental results show that the proposed PSO-based integrated feature selection algorithm achieves higher efficiency and accuracy in feature selection. Compared to before optimization, the accuracy of the 4 machine learning ROP prediction models trained using the optimized integrated feature selection results is improved by 59%, 1%, 3%, and 1%, respectively. Compared to models trained using all features, the accuracy has been improved by 24%, 2%, 4%, and 3%, respectively. This paper provides

收稿日期：2024-06-27；**修回日期**：2024-07-30　　**DOI**：10.12143/j.ztgc.2025.02.018
基金项目：国家自然科学基金项目“量化月壤扰动特征的模块化月球钻进力学模型研究”（编号：42072344）；四川省自然科学基金青年基金项目“基于数字孪生的动态时变钻进工况自适应迁移模型研究”（编号：2024NSFSC0817）
第一作者：胥知画，女，汉族，2001年生，硕士研究生，机器人工程专业，研究方向为人工智能在钻井中的应用，四川省成都市成华区二仙桥东三路1号，1596720585@qq.com。
通信作者：姜杰，女，汉族，1985年生，讲师，博士，长期从事数据挖掘与人工智能在钻井中的应用研究工作，四川省成都市成华区二仙桥东三路1号，jiangjie13@cdut.edu.cn。
引用格式：胥知画，姜杰，周长春，等.基于粒子群优化的融合特征选择钻速预测模型研究[J].钻探工程，2025，52(2)：134-143.
XU Zhihua, JIANG Jie, ZHOU Changchun, et al. Research on a rate of penetration (ROP) prediction model based on feature selection integrated with particle swarm optimization (PSO)[J]. Drilling Engineering, 2025, 52(2): 134-143.

an effective feature selection method for cases where too many feature parameters have been extracted in drilling engineering. It offers significant guidance for the practical application of feature selection algorithms in the engineering field.

Key words: ROP prediction model; feature selection; PSO; machine learning

0　引言

随着能源需求的不断增长和石油资源的日益枯竭,钻井工程环境变得更加复杂。常规钻井方法在效率、成本和安全性方面的不足,制约了行业的进一步发展。因此,如何提高钻井效率、降低成本、减少事故风险成为了油气行业的关键挑战[1-2]。钻速预测作为钻井过程中的关键环节之一,其准确性和可靠性直接影响着钻井的效率和安全。因此,如何建立高效、精准的钻速预测模型成为了当前地质勘探与油气开发领域的研究热点[3-4]。特征选择是构建有效钻速预测模型的关键步骤之一[5-7]。通过对钻井过程中的各种影响因素进行分析和筛选,选取最具代表性和区分度的特征,可以大大提高钻速预测模型的准确性和稳定性[8]。然而,由于钻井过程中存在众多的影响因素和数据特征,传统的特征选择方法往往难以充分挖掘数据的潜在信息,导致构建的预测模型效果不尽人意,需要根据钻井实际需求构建一种面向钻速预测的特征选择方法。

近年来,数据挖掘和机器学习等技术在持续进步[9],钻井工程中的钻进参数特征选择算法引起了较为广泛的关注[10]。Moraveji等[11]根据实际钻井数据,对钻速与井深、钻压、转速、钻头射流冲击力、屈服点、塑性黏度比以及10 min~10 s凝胶强度比之间的数学关系进行了研究,并建立了相应的预测模型。Sarah等[12]对数据挖掘方法以及几种机器学习算法进行了比较,评估这些方法在预测ROP方面的准确性和有效性。康文豪等[13]为了能够使预测模型具有较好的拟合效果,提出了一种双层特征选择方法选择特征。Alsabaa等[14]基于机器学习技术开发了一套全自动化系统,用于预测合成油基钻井液(平面流变型)的流变特性以及监测钻井液的流变特性。甘超等[15]引入增量学习等技术来实现钻速预测模型的动态更新,提出了一种模型预训练的钻速动态预测方法。

综上,传统模型的重点在于对钻井的整个过程进行建模分析,力求厘清不同参数如何影响钻速。此类模型的主要缺点在于影响钻速的因素繁多,难以准确地建立物理模型来描述。为了解决传统特征选择方法的不足[16-17],本文提出了一种基于粒子群优化的融合特征选择算法。通过将粒子群优化算法应用于特征选择,可以高效地搜索到最佳特征子集,从而提升预测模型的准确性和稳定性。最后,通过比较BP神经网络(Back Propagation, BP)、决策树回归(Decision Tree, DT)、梯度提升决策树(Gradient Boosting Decision Tree, GBDT)以及随机森林(Random Forest, RF)4种模型的预测效果,证明了所建立的粒子群优化算法对融合特征选择算法在提高预测模型准确性方面的有效性。

1　数据预处理

1.1　数据来源

本文数据采集自中国南海某区块的8口井钻井数据,该区块地层岩性复杂[18],导致施工作业环境复杂多变,因此由传感器所采集的原始数据较为“混乱”[19]。原始钻井参数主要有两种获取方式:一种为静态方式,此部分参数一般人为不可控制,通常由现场工程师填写。另一种为动态方式,是在钻井过程中安装传感器采集与控制的参数,如钻压、流量、钻速等[20]。本数据采集过程中按照每1 m钻深采集一组数据,为实时传输数据,采集后存储到数据库中。将采集到的数据整合为一个D/den行×n列数据矩阵[21],其中n为钻井参数数量,D为井深,den为不同的钻井参数采集时最大密度。南海某片区数据经整合后共有44种不同类型的参数数据21912条。表1所示为参数缩写信息。

1.2　数据处理

传感器采集数据会随着钻井施工进行而呈现不同特点。现有的钻井数据是采集的原始钻井数据,包含钻井过程中的各个阶段的各类数据[22]。对参数进行观测分析,可发现如下特点:(1)数据基本稳定;(2)存在数字为0的情况;(3)存在突变的情况;(4)不同参数之间数量级差异巨大。针对以上特点,本文采用了缺失值处理、标准化处理等方法对原始数据进行处理[23]。

1.2.1　缺失值处理

采用插值填充法进行缺失值处理。插值填充

表1 参数信息

Table1 Parameter information

参数类型	参数名称/缩写
井眼位置	井眼位置/NO*;深度/D
施工工艺	井径/dia*;钻速/ROP;钻压/WOB;转速/RPM;扭矩/T;泵压/SPP;泵量/Q;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;漏斗黏度/MV;塑性黏度/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;pH值/pH*;泥饼厚度/MT*;含砂量/SA*;氯离子含量/CLC;钙离子含量/CAC*;固相含量/SO;膨润土含量/SOC;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。

法是指采用算法计算进行插值填充,多用于当某种参数的采集密度与其他参数的采集密度不一致时的情况。其操作步骤为:以密度最大的参数数据作为基准,进行缺失值的填补。常见的缺失值插值填充算法有拉格朗日插值填充、k近邻(KNN)插值填充、随机森林填充等算法[24]。

本文选择插值填充法中的k近邻插值填充对井眼中的垂深缺失数据进行填补,其步骤为:(1)找出存在缺失值参数的样本;(2)选择存在缺失参数值样本行的其余数据,同时选择近邻数k值;(3)采用欧式距离计算其余样本与缺失值样本之间的距离;(4)选出k个距离最近的样本点的缺失参数取值对缺失样本进行补全,本文中缺失值为连续参数,因此采用计算平均值的方法进行补全。欧氏空间中每个样本点与被填补点的距离计算见式(1)[25]:

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

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。

以钻井液出口温度为例,其缺失值及补全效果如图1、图2所示。

1.2.2 标准化数据处理

选出k个距离最近的样本点的缺失参数取值对缺失样本进行补全,若缺失参数为离散值则采用投票法决定填补值,若缺失参数为连续值则计算其平均值进行补全。本文中缺失值为连续参数,因此采用计算平均值的方法进行补全。由于各钻井参数的取值量纲并不相同,甚至部分钻井参数取值之间存在巨大的数量差异,直接使用存在量纲差异的数据建立机器学习机械钻速预测模型会造成误差加

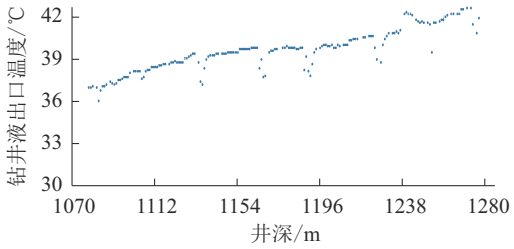

图1 出口钻井液温度缺失情况

Fig.1 Condition of missing drilling fluid outlet temperature value

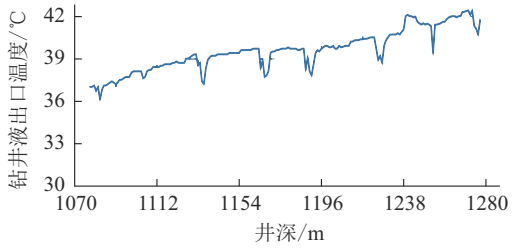

图2 出口钻井液温度缺失值处理效果

Fig.2 Treatment effect of missing drilling fluid outlet temperature value

大而精度降低。因此,本文进行数据标准化处理采用的Z-score标准化方法,其计算公式见式(2)[26]:

$$x_{new}=(x_{old}-\mu)\sigma_{xlist} \tag{2}$$

式中:x_{new}——标准化后的数据;x_{old}——标准化前的原始数据;μ——平均值;σ_{xlist}——该变量所有原始数据的标准差。

采用Z-score标准化方法进行处理之后的结果如图3所示,可以看到两个参数的取值范围都到了0附近,量纲一致。

1.3 数据分组

本文将整合并经过预处理之后的钻井数据集按

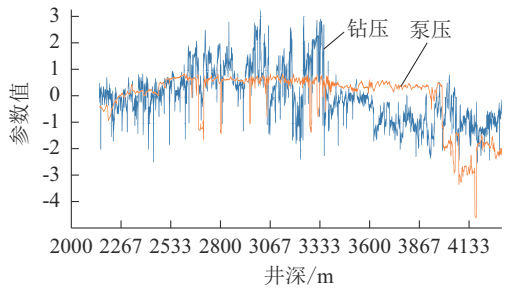

图 3　钻压泵压 Z-score 标准化处理结果

Fig.3　Z-score standardized treatment results of weight on bit pump pressure

照研究思路分成 his_data（历史钻井数据）和 neighbour_data（邻井数据）2 大类型分别进行研究。his_data（历史钻井数据）由在 1、2、3、4、5 和 6 号井采集数据组成。此部分数据主要用于特征优选算法设计，并在此数据集中建立不同机器学习钻速预测模型进行钻速预测。neighbour_data（邻井数据）由在 7 和 8 号井所采集的数据组成，此部分数据集也可认为是历史钻井数据集。一方面用于对基于 his_data 数据所设计的特征优选算法进行优化，同时利用此数据集对优化后的特征优选算法进行验证；另一方面在邻井数据集上对所设计的特征优选算法研究也可以验证特征优选算法在邻井数据上的泛化能力。

2　基于粒子群优化的钻井参数特征优选模型

2.1　融合特征选择算法概述

融合特征选择算法是以融合学习思想为基础，再综合皮尔逊相关性分析法、方差过滤法和互信息法，去除掉低相关性参数和中相关性参数，最终只选择与钻速（ROP）具有高相关性的参数组与方差过滤和互信息法选出的参数的交集参数作为模型的输入特征参数组[27]。操作步骤可分为 5 步（图 4）。

（1）对数据进行皮尔逊相关性计算，然后根据

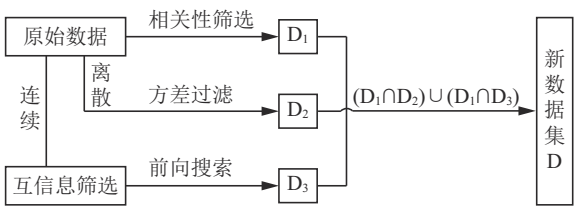

图 4　特征选择过程示意

Fig.4　Schematic diagram of the feature selection process

相关性将所有特征参数分为高、中、低 3 个相关性组别。设定高相关性阈值 threshold_1，以确定出与钻速具有高相关性的高相关性参数组。

（2）对所有特征参数中的离散型参数进行方差过滤，根据方差值设置方差过滤阈值 threshold_2，选择出方差较高的特征参数。

（3）对所有特征参数中的连续型参数进行互信息估计量计算，并按互信息值大小排序。设定互信息阈值 threshold_3，对特征进行进一步选择。

（4）采用前向搜索策略来优化互信息筛选结果，从原始特征集中依次选出能最大化提升模型性能的特征参数，直到选完所有能提升模型精度的参数，将剩余无助于提升精度的参数筛除。

（5）将相关性过滤结果分别与方差过滤以及互信息过滤的结果首先进行交集操作，然后合并两个交集后的参数组，这样就得出了特征选择融合算法的最终结果。

2.2　粒子群算法概述及适应度函数建立

粒子群优化算法（PSO）是通过模拟鸟群觅食的搜索过程[28-29]，不断调整粒子的位置和速度，以找到最优解。具体来讲，粒子群优化算法由一群初始化的粒子组成，每个粒子代表一个候选解，它们的位置和速度分别表示各个粒子解的参数和搜索方向。每个粒子在搜索过程中根据自身历史最优位置和整个群体的历史最优位置来更新其位置和速度，不断搜索最优解[30]。各粒子通过当前位置和速度计算下一步的位置和速度，并更新自身及群体的历史最优位置。位置和速度的更新公式如式（3）所示：

$$\begin{cases} v_i(t+1)=wv_i(t)+c_1r_1(pbest_i-x_i(t))+ \\ \qquad\qquad c_2r_2(gbest-x_i(t)) \\ x_i(t+1)=x_i(t)+v_i(t+1) \end{cases} \tag{3}$$

式中：$v_i(t)$——粒子 i 在 t 时刻的速度；w——惯性权重；c_1——个体自身的学习因子；r_1、r_2——均为[0,1]之间的随机数；$pbest_i$——粒子 i 的历史最优位置；$x_i(t)$——粒子 i 在 t 时刻的位置；c_2——整体的学习因子；$gbest$——整个群体的历史最优位置。

粒子群优化算法的流程如图 5 所示，图中实线箭头为逻辑流，虚线箭头为数据流。其中粒子的优劣由适应度函数 $f(x)$ 决定。

粒子群优化算法主要涉及以下参数：（1）粒子种群数 m：在粒子群优化算法中，粒子表示待优化问

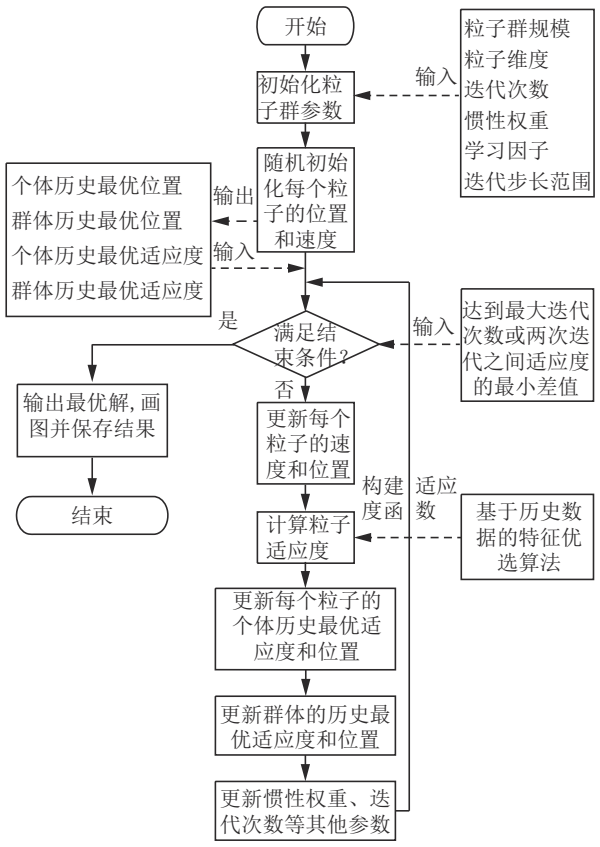

图5 粒子群优化算法流程

Fig.5 Flow chart of PSO algorithm

题的一个解,粒子数量影响算法的搜索范围和速度。通常情况下取50～1000,粒子数量越多,搜索空间越广,同时也会增加计算量和收敛时间。(2)维数:维数表示待优化问题的自变量数量,即解空间的维度。维数的大小显著影响算法的性能和搜索空间的范围。(3)位置:表示粒子在空间中的位置,通常用一个向量来表示。(4)速度:表示粒子在解空间中的运动速度,同样用向量表示。(5)适应度函数:用于评价每个粒子的解的优劣,即衡量解的质量。适应度函数是粒子群优化算法的核心部分,决定了每个粒子的移动方向和速度。(6)个体最佳位置:表示每个粒子的历史最佳位置,即该粒子在搜索过程中找到的最好解,个体最佳位置是用来帮助粒子探索更好的解。(7)全局最佳位置:表示所有粒子历史最佳位置中最好的位置,即整个种群中找到最好的解,全局最佳位置是用来指引整个种群朝向更优的解。(8)惯性权重:是一个控制粒子运动速度和粒子探索能力的参数,惯性权重通常随着迭代次数的逐渐增加而减小。(9)学习因子:包括加速度系数、个体加速度因子和全局加速度因子,它们是控制粒子运动速度和运动方向的参数,其值需要根据问题的特点进行调整。将训练好的基于历史数据的特征优选算法模型作为粒子群优化算法的适应度函数,以基于历史数据的特征优选算法模型中的3个关键参数threshold_1、threshold_2和threshold_3为优化目标,即需要寻找使得基于历史数据特征优选结果进行机器学习建模,使得建模误差最小的3个关键参数最优取值。由此建立适应度函数,当其值越低时,适应度越高,越接近优化目标。

2.3 参数设置及粒子群优化

为探究能使得特征优选结果建模误差最小的最优参数,粒子群寻优的目标是使得特征优选结果的机器学习钻速预测模型建模均方根误差*RMSE*最小,需要优化的3个参数分别是高相关性分组阈值threshold_1、方差过滤阈值threshold_2和互信息筛选阈值threshold_3。设置粒子群优化算法的自变量空间维度为3,位置限制即3个自变量的取值范围与适应度函数中3个关键参数的取值范围保持一致。表2展示的是3个参数在基于历史数据的特征优选算法中的取值以及在粒子群优化过程中的寻优空间,同时,每个粒子的速度限制V_m设置为每个自变量变化范围的20%。对于粒子群优化算法的参数设置:粒子群大小m为20,惯性权重为0.75,个体学习因子c_1和群体学习因子c_2均为2,最大迭代次数设定为100次。在基于历史数据的特征优选算法的基础上,通过jupyter notebook编写程序实现了粒子群优化算法。

表2 自变量取值范围及速度限制

Table 2 Value range and speed limit of argument

参 数	threshold_1	threshold_2	threshold_3
原算法中取值	0.6	中位数	中位数
取值范围	0.4～1	0～100	0～1
速度限制	0.12	20	0.2

粒子群优化算法经过100次迭代优化,迭代过程如图6所示。观察到在大约第6次迭代时,适应度函数的均方根误差(*RMSE*)达到了最低点。此时*RMSE*为3.43,最小适应度取值对应的3个关键参数取值见表3。将优化后的3个关键参数作为设计的基于历史数据的特征优选算法中,以neighbour_

data 数据集作为测试进行特征优选，如表 4 所示，特征优选结果显示共选出 15 个参数。其中井眼位置参数和地质情况参数各 1 个，各占所选参数的 6.67%；施工工艺参数 5 个，占所选参数的 33.33%；钻井液性质参数 8 个，占所选参数的 53.33%。

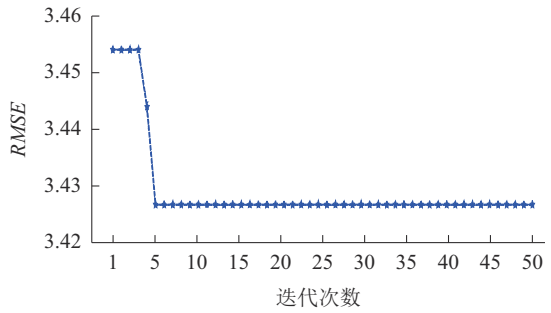

图 6　粒子群优化算法迭代过程

Fig.6　Iterative process of PSO algorithm

表 3　粒子群优化后 3 个关键参数取值

Table 3　Value of the three key parameters after PSO

参数	threshold_1	threshold_2	threshold_3
取值	0.46	0.5	0.52

表 4　优化后的特征优选算法在测试数据上的特征优选结果

Table 4　The feature optimization result of the optimized feature optimization algorithm on the test data

特征优选结果	D6;Q;MW;PV;YP;SM;FL;WOH;MO;PP;D3;SO;MI;D;dia
R^2	0.92
RMSE	3.25

3　基于粒子群优化的钻速预测模型验证

3.1　10 折交叉验证

首先，将原始数据集随机分成 10 等份，以确保每份数据的分布特性相似。然后，每次选择其中 1 份作为验证集，剩下的 9 份作为训练集进行模型训练。通过循环这一过程，可以获得 10 个独立的模型，从而评估模型在新数据上的表现。这个过程重复 10 次，每次使用不同的数据分组作为验证集。每轮验证的性能通过准确率、精确度、召回率和 F1 分数等指标进行评估。最终，通过取每轮验证的性能指标的平均值，得到模型的总体性能评估结果。这一方法的原理如图 7 所示。

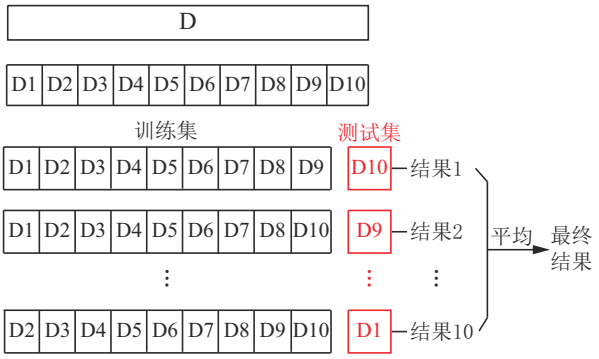

图 7　10 折交叉验证的计算原理

Fig.7　10 fold cross validation calculation schematics

10 折交叉验证的优点在于充分利用了数据集，确保每个样本都恰好出现在验证集中 1 次。这有助于减小因数据划分不同而引入的偶然性，提高模型性能评估的稳定性[31]。

3.2　模型评估

3.2.1　决定系数（R^2）

以决定系数（R^2）作为模型的评估指标。R^2 是衡量回归模型对观测数据拟合程度的统计指标，当 R^2 接近 1 时，表示拟合效果较好[32]。具体计算公式如式（4）所示：

$$R^2=1-\left[\sum_{i=1}^{n}(y_i-\hat{y}_i)^2\right]/\left[\sum_{i=1}^{n}(y_i-\bar{y})^2\right] \quad (4)$$

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

3.2.2　均方根误差（*RMSE*）

均方根误差（*RMSE*）是用来衡量预测值与真实值之间的偏差的一种统计量，*RMSE* 是预测值和真实值之间偏差平方和的均值的算术平方根，能够反映出预测误差。*RMSE* 值越接近 0，表明模型性能越好，预测精度越高[32]。计算公式如式（5）所示：

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \quad (5)$$

3.3　粒子群优化结果钻速预测验证

为了验证经过粒子群优化后的特征优选算法的有效性，本文设计对比试验对经过粒子群优化之后的特征优选结果进行分析。首先，利用 neighbour_data 数据集上的所有钻井参数、所设计的融合特征的优选结果以及利用粒子群优化算法对其优化之后的特征优选结果分别训练 4 个不同的机器学习钻速预测模型，并对其进行评估，同时利用

这些模型分别在邻井数据集上进行钻速预测。最后，对利用全部特征所训练的钻速预测模型、利用融合特征优选结果所训练的钻速预测模型和利用基于粒子群优化之后的特征优选结果所训练的钻速预测模型性能以及预测误差进行比较和分析。分别使用邻井数据的全部特征参数、融合特征算法的特征优选结果和使用粒子群算法优化之后的融合特征优选结果作为模型的不同输入，分别建立了BP、DT、GBDT以及RF 4种模型来进行钻速预测。这4种模型在测试集上的精度(R^2)和均方根误差($RMSE$)对比结果见图8～11，平均精度(R^2)和均方根误差($RMSE$)的汇总列于表5。

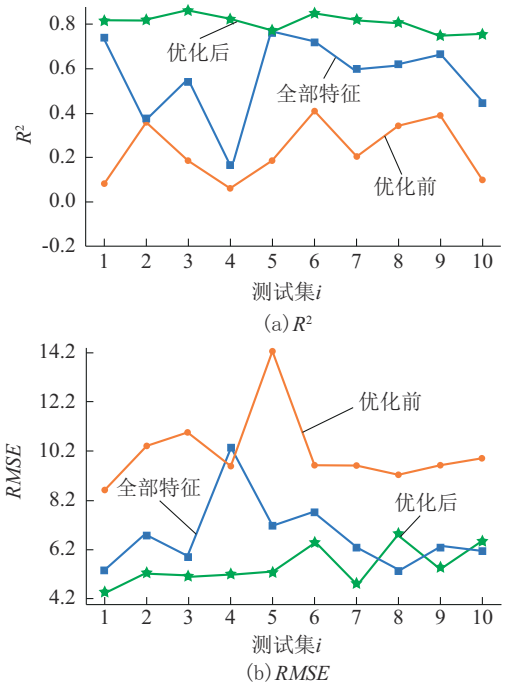

图8 BP神经网络钻速预测模型10折交叉测试集评估

Fig.8 Evaluation of 10 fold cross test set of BP ROP prediction model

从精度和均方根误差两个角度分析可知，BP神经网络算法在此数据集上利用全特征、优化特征优选结果前与优化后的特征优选结果所训练的钻速预测模型性能均较差，但是使用粒子群方法对融合特征优选方法仍然有一定的优化效果。由特征优选结果作为输入建立的DT、GBDT以及RF模型性能优于使用全部特征所建立的模型，而经过粒子群优化之后的特征优选结果模型性能又优于特征优选结果模型。3个不同输入的钻速预测模型的性能

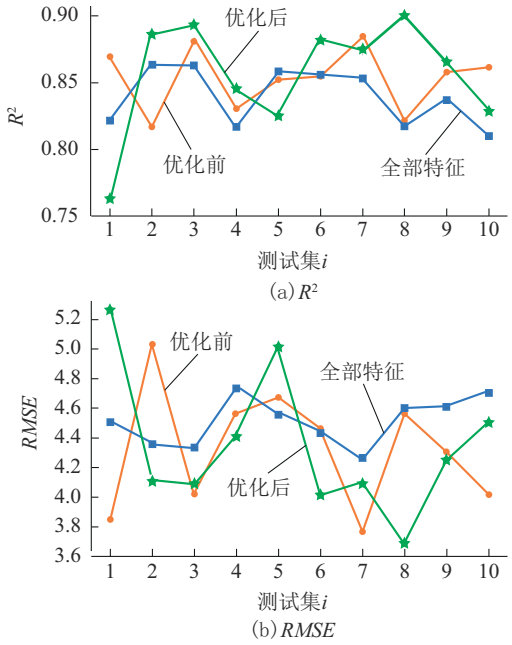

图9 决策树钻速预测模型10折交叉测试集评估

Fig.9 Evaluation of 10 fold cross test set of DT ROP prediction model

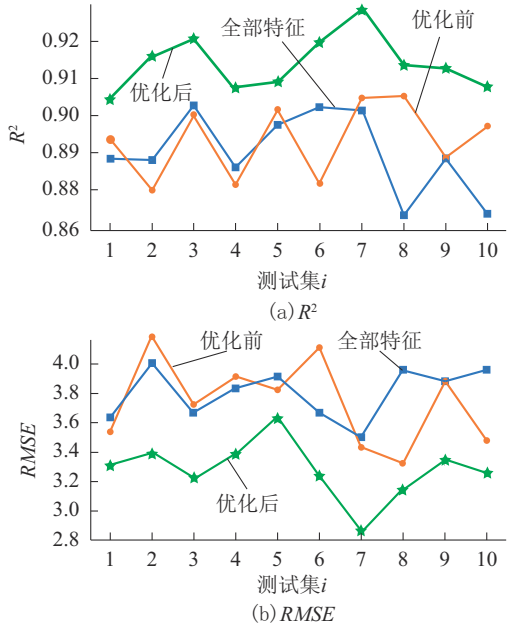

图10 GBDT钻速预测模型10折交叉测试集评估

Fig.10 Evaluation of 10 fold cross test set of GBDT ROP prediction model

表现验证了本文所使用的粒子群优化算法对所设计的融合特征优选方法优化效果。

3.4 实验结果分析

图12、图13所示为使用邻井数据集上的全部特

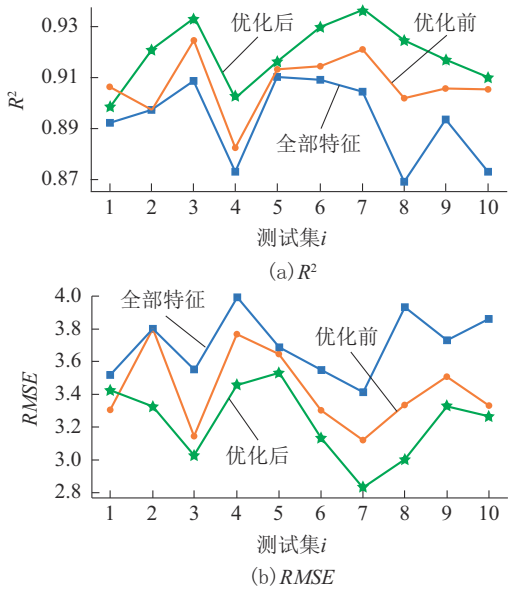

图 11　随机森林钻速预测模型 10 折交叉测试集评估

Fig.11　Evaluation of 10 fold cross test set of RF ROP prediction model

表 5　不同钻速预测模型测试集精度与误差平均值

Table 5　The average of R^2 and *RMSE* of test set of different prediction models

模　型	对比类型	平均 R^2	平均 *RMSE*
BP 神经网络	全部特征	0.56	6.72
	优化前特征	0.21	10.18
	优化后特征	0.80	5.53
决策树	全部特征	0.84	4.51
	优化前特征	0.85	4.32
	优化后特征	0.86	4.35
GBDT	全部特征	0.88	3.80
	优化前特征	0.89	3.74
	优化后特征	0.92	3.28
随机森林	全部特征	0.89	3.70
	优化前特征	0.91	3.43
	优化后特征	0.92	3.23

征参数、使用所设计的融合特征优选结果和使用粒子群优化之后的特征优选结果所训练的不同机器学习钻速预测模型及建模精度和误差的对比。结合训练的不同模型在 10 折交叉验证过程中的 10 个测试集上评估得分的变化情况,验证了所设计的融合特征优选算法有更好的拟合效果。从图中的对比结果,可以看出经过粒子群优化之后的融合特征优选结果所训练的模型比优化前的融合特征优选结果所训练的模型有更好的预测性能。

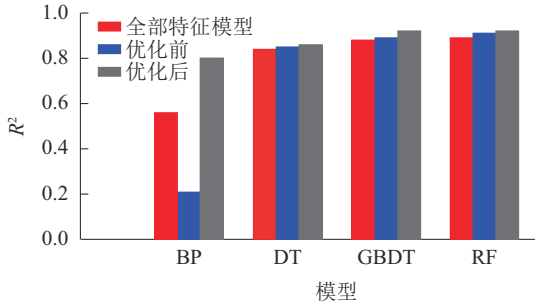

图 12　不同模型全特征、优化前和优化后的 R^2

Fig.12　The R^2 of different model full features, pre-and post-optimization

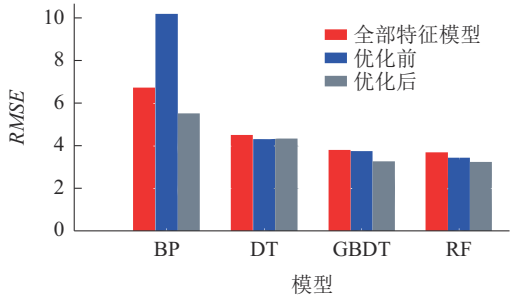

图 13　不同模型全特征、优化前和优化后的 *RMSE*

Fig.13　The *RMSE* of different model full features, pre-and post-optimization

图 12、图 13 的精度和误差柱状图统计也说明了 BP 神经网络算法不适合用于在此数据集上训练机器学习钻速预测模型,本文利用 BP 神经网络算法在全部特征、优化前特征和优化后特征训练的模型精度分别为 0.56、0.21 和 0.80,误差分别为 6.72、10.18 和 5.53,可见精度均较低,误差均偏大。

从图 12、图 13 分析还可得知,在剩下的 DT、GBDT 和 RF 3 个钻速预测模型中,由 3 个不同参数数据作为输入所训练的 DT 钻速预测模型的精度均低于由决策树集成的 RF 和 GBDT 钻速预测模型,同时误差也都更高。从图中还可以发现,在 3 个不同参数集作为输入所训练 RF 钻速预测模型的精度分别为 0.89、0.91 和 0.92,高于 3 个不同 GBDT 钻速预测模型的 0.88、0.89 和 0.92;同时 3 个不同 RF 钻速预测模型的均方根误差分别为 3.70、3.43 和 3.23,低于 GBDT 钻速预测模型的 3.80、3.74 和 3.28,证明了粒子群优化算法能够使得融合特征选择算法选出包含更多目标信息的钻井参数特征。

4 结论与建议

精确预测机械钻速有助于优化钻井流程、提高作业效率并降低施工成本。本文以南海某区块的8口井钻井数据为例,基于粒子群优化算法进行融合钻井特征参数优选,然后建立了4种机器学习钻速预测模型对机械钻速进行预测,主要结论如下:

(1)针对钻速预测的机器学习建模,本文介绍的融合特征选择算法能从众多特征参数中准确挑选出对模型贡献最大的参数,具有高效筛选能力。通过深入数据分析,确定对模型性能有显著影响的关键特征,从而提升模型的预测精度和泛化能力。

(2)本文建立的预测模型结果显示:BP神经网络算法不适合在该数据集上训练ROP预测模型,即使进行了粒子群优化,其准确性仍然较差,误差也比较显著。相比之下,使用DT、GBDT和RF算法训练的ROP预测模型精度相较于优化前和使用全部特征所训练的模型都有所提升,证明了所建立的粒子群优化算法对融合特征优选算法有较好的优化效果。

(3)粒子群优化算法对初始解的选择非常敏感,初始解的质量直接影响了算法的收敛性和最终结果。如果初始解选择不当,可能导致算法陷入局部最优解,影响特征选择的效果,在未来的研究中可以针对这一问题进行优化。

参考文献(References):

[1] 郭旭涛.智能钻井技术研究现状[J].现代工业经济和信息化,2022,12(3):150-151,154.
GUO Xutao. Research status of intelligent drilling technology [J]. Modern Industrial Economy and Informationization, 2022, 12(3):150-151,154.

[2] 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. DOI: 10.1155/2016/3575380.

[3] 张菲菲,崔亚辉,于琛,等.基于机器学习的钻井工况识别技术现状及发展[J].长江大学学报(自然科学版),2023,20(4):53-65,143.
ZHANG Feifei, CUI Yahui, YU Chen, et al. Recent developments and future trends of drilling status recognition technology based on machine learning [J]. Journal of Yangtze University (Natural Science Edition), 2023,20(4):53-65,143.

[4] 谭扬.机器学习算法在石油钻井领域的应用优化研究[D].北京:北京邮电大学,2019.
TAN Yang. Research on application and optimization of machine learning algorithm in oil drilling field[D]. Beijing: Beijing University of Posts and Telecommunications, 2019.

[5] 蒲先渤,李泽群,尹飞,等.基于PCA-LM-BP神经网络的岩石可钻性预测研究[J].钻探工程,2023,50(6):63-68.
PU Xianbo, LI Zequn, YIN Fei, et al. Research on rock drill ability prediction based on PCA-LM-BP neural network [J]. Drilling Engineering, 2023,50(6):63-68.

[6] Deng Y, Chen M, Jin Y, et al. Theoretical and experimental study on the penetration rate for roller cone bits based on the rock dynamic strength and drilling parameters[J]. Journal of Natural Gas Science and Engineering, 2016,36, Part A:117-123.

[7] 王亚飞,张占荣,刘华吉,等.基于模型融合的钻进参数识别岩石类型研究[J].钻探工程,2023,50(2):17-25.
WANG Yafei, ZHANG Zhanrong, LIU Huaji, et al. Data?driven model for the identification of the rock type by drilling data [J]. Drilling Engineering, 2023,50(2):17-25.

[8] 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 drilling rate prediction method based on Machine Learning and Multi-source data preprocessing [J]. China Petroleum and Chemical Standard and Quality, 2021,41(20):133-136.

[9] 王胜,赖昆,张拯,等.基于随钻振动信号与深度学习的岩性智能预测方法[J].煤田地质与勘探,2023,51(9):51-63.
WANG Sheng, LAI Kun, ZHANG Zheng, et al. Intelligent lithology prediction method based on vibration signal while drilling and deep learning[J]. Coal Geology & Exploration, 2023,51(9):51-63.

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

[11] Moraveji M K, Naderi M. Drilling rate of penetration prediction and optimization using response surface methodology and bat algorithm[J]. Journal of Natural Gas Science and Engineering, 2016,31:829-841.

[12] Sabah M, Talebkeikhah M, Wood D A, et al. A machine learning approach to predict drilling rate using petrophysical and mud logging data[J]. Earth Science Informatics, 2019,12(3):319-339.

[13] 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].重庆理工大学学报(自然科学),2022,36(7):303-309.
KANG Wenhao, XU Tianqi, WANG Yangguang, et al. Short-term wind power prediction based on double-layer feature selection and CatBoost-Bagging integration [J]. Journal of Chongqing University of Technology:Natural Science, 2022,36(7):303-309.

[14] Alsabaa A, Gamal H, Elkatatny S, et al. Machine learning model for monitoring rheological properties of synthetic Oil-Based mud[J]. ACS Omega, 2022,7(18):15603-15614.

[15] 甘超,汪祥,王鲁朝,等.基于区域多井数据优选与模型预训练的深部地质钻探过程钻速动态预测方法[J].钻探工程,2023,50(4):1-8.
GAN Chao, WANG Xiang, WANG Luchao, et al. Dynamic prediction method of rate of penetration (ROP) in deep geological drilling process based on regional multi-well data optimization and model pre-training[J]. Drilling Engineering, 2023,50(4):1-8.

[16] 邓少贵,张凤姣,陈前,等.基于混合机器学习算法的页岩薄互

层识别方法[J].石油学报，2023，44(7)：1097-1104.
DENG Shaogui, ZHANG Fengjiao, CHEN Qian, et al. Identification of shale thin interbeds based on hybrid machine learning algorithm[J]. Acta Petrolei Sinica, 2023,44(7):1097-1104.

[17] 曾凡辉，胡大淦，张宇，等.数据驱动的页岩油水平井压裂施工参数智能优化研究[J].石油钻探技术，2023，51(5)：78-87.
ZENG Fanhui, HU Dagan, ZHANG Yu, et al. Research on Data-Driven intelligent optimization of fracturing treatment parameters for shale oil horizontal wells[J]. Petroleum Drilling Techniques, 2023,51(5):78-87.

[18] 曾小龙，李谦，魏宏超，等.基于南海巨厚塑性泥岩地层特征的钻速预测模型[J].煤田地质与勘探，2023，51(11)：159-168.
ZENG Xiaolong, LI Qian, WEI Hongchao, et al. Rate-of-penetration(ROP)prediction model based on formation characteristics of extremely thick plastic mudstone in South China Sea[J]. Coal Geology & Exploration, 2023,51(11):159-168.

[19] 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 sampled data cleaning technology based on regression model[J]. Electronics Optics & Control, 2022,29(4):117-120.

[20] Wang X, Gan C, Cao W H. A novel drilling rate of penetration (ROP) prediction method using data pre-processing techniques and T-S fuzzy inference[C]//2021 40th Chinese Control Conference (CCC). Shanghai China, 2021:1261-1266..

[21] 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)：85-95.
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):85-95.

[22] 匡俊寨，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.

[23] 姜杰，霍宇翔，张颢曦，等.基于数字孪生的智能钻探服务平台架构[J].煤田地质与勘探，2023，51(9)：129－137.
JIANG Jie, HUO Yuxiang, ZHANG Haoxi, et al. Architecture of intelligent service platform for drilling based on digital twin[J]. Coal Geology & Exploration, 2023,51(9):129-137.

[24] 乔永坚，刘晓琳，白亮.面向高维特征缺失数据的K最近邻插补子空间聚类算法[J].计算机应用，2022，42(11)：3322-3329.
QIAO Yongjian, LIU Xiaolin, BAI Liang. K-nearest neighbor imputation subspace clustering algorithm for high-dimensional data with feature missing[J]. Journal of Computer Applications, 2022,42(11):3322-3329.

[25] 曹凯鑫，汤猛猛，葛建鸿，等.大气污染物PM2.5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)：229-305.
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.

[26] 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].现代隧道技术，2022，59(2)：38-44，52.
WANG Shuangjing, WANG Yujie, LI Xu, et al. Study of standardized pre-processing method of TBM tunnelling data[J]. Modern Tunnelling Technology, 2022,59(2):38-44,52.

[27] 周长春，姜杰，李谦，等.基于融合特征选择方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.

[28] 张涛，李艳萍，刘晓宇，等.基于自适应粒子群优化最小二乘支持向量机的深层变质岩测井岩性识别[J].地球物理学进展，2023，38(1)：382-392.
ZHANG Tao, LI Yanping, LIU Xiaoyu, et al. Lithology interpretation of deep metamorphic rocks with well logging based on APSO-LSSVM algorithm[J]. Progress in Geophysics, 2023,38(1):382-392.

[29] 张宁，刘祺，邵俊杰，等.基于模糊粒子群算法的钻杆运移装置控制系统研究[J].煤田地质与勘探，2023，51(3)：177-185.
ZHANG Ning, LIU Qi, SHAO Junjie, et al. Research on control system of drill pipe conveying device based on fuzzy particle swarm optimization[J]. Coal Geology & Exploration, 2023,51(3):177-185.

[30] 黄宇，顾智勇，张中印，等.基于差分量子粒子群优化算法的作业车间调度[J].科学技术与工程，2022，22(29)：12848-12854.
HUANG Yu, GU Zhiyong, ZHANG Zhongyin, et al. Job-shop scheduling problem based on differential quantum particle swarm optimization algorithm[J]. Science Technology and Engineering, 2022,22(29):12848-12854.

[31] Ru Z L, Zhao H B, Zhu C X. Probabilistic evaluation of drilling rate index based on a least square support vector machine and Monte Carlo simulation[J]. Bulletin of Engineering Geology and the Environment, 2019,78(5):3111-3118.

[32] 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.
LÜ Xiaoling, SONG Jie. Big data mining and statistical machine learning[M]. Beijing: China Renmin University Press, 2016.

（编辑　王跃伟）