

<https://doi.org/10.3799/dqkx.2026.218>



基于最优相关性的钻速建模算法

魏思维¹, 李 谦^{1*}, 姜 杰², 何俊杰¹, 李同意¹

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

摘 要: 针对复杂地层钻井过程中钻速(ROP)预测精度低、参数间存在非单调、非线性强耦合的难题——传统 Pearson/Spearman 等相关系数主要衡量线性或单调关系, 难以有效刻画此类复杂依赖, 易导致关键特征被低估。旨在开发一种动态最优相关性建模算法, 以提升钻速预测的准确性和工程适用性。基于 10 口油气井的 21 912 组工程数据, 综合采用中位数填充、IQR 离群值处理和 Savitzky-Golay 滤波进行数据预处理; 通过对比 Pearson、Spearman、Kendall 和 Chatterjee 四种相关性理论, 识别影响钻速的关键参数(如钻井液密度、固相含量和井深); 引入滑动窗口技术(窗口大小 400、步长 100)实现局部动态建模, 并以随机森林模型为核心进行回归预测。实验表明, 以 Chatterjee 算法选取排名前 5 的特征时表现最优, 结合滑动窗口与随机森林的架构在第 10 口测试井中窗口测试集平均 R^2 达 0.985, 较全局静态建模($R^2=0.96$)显著提升。得到的最优相关性钻速建模算法能自适应地层突变, 有效捕捉参数间的局部关联, 为钻速实时优化提供高精度解决方案, 为钻井工程从经验试错向数据驱动决策的转变提供有效技术支撑。

关键词: 预测钻速; 相关性; 机器学习; 动态建模; 原始数据处理。

中图分类号: TD41

文章编号: 1000-2383(2026)08-3102-16

收稿日期: 2026-05-13

Correlation-Optimized Modeling Algorithm for Rate of Penetration

Wei Siwei¹, Li Qian^{1*}, Jiang Jie², He Junjie¹, Li Tongyi¹

1. College of Environment and Civil Engineering, Chengdu University of Technology, Chengdu 610059, China

2. School of Mechanical and Electrical Engineering, Chengdu University of Technology, Chengdu 610059, China

Abstract: Aiming at the problems of low prediction accuracy of rate of penetration (ROP) during drilling in complex formations and strong non-monotonic, nonlinear coupling among parameters: traditional correlation coefficients such as Pearson and Spearman mainly measure linear or monotonic relationships, which cannot effectively characterize such complex dependencies and tend to result in underestimation of key features, this paper aims to develop a dynamic optimal correlation modeling algorithm. This approach is designed to enhance both the accuracy and engineering applicability of ROP prediction. Using 21,912 data sets from ten oil and gas wells, we performed data preprocessing including median imputation for missing values, IQR-based outlier removal, and Savitzky-Golay smoothing. By comparing four correlation theories: Pearson, Spearman, Kendall, and Chatterjee: key parameters affecting ROP (such as mud density, solid content, and well depth) were identified. A sliding window technique

基金项目: 四川省自然科学基金项目青年基金, 基于数字孪生的动态时变钻进工况自适应迁移模型研究(No. 2024NSFSC0817); 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”(No. CNOOC-KJ135ZDXM38ZJ05ZJ)。

作者简介: 魏思维(2003—), 男, 硕士, 从事数据分析与人工智能在钻探钻井中的应用研究, ORCID: 0009-0002-5180-7691, E-mail: 3210408869@qq.com

*** 通讯作者:** 李谦, ORCID: 0000-0003-1535-7734. E-mail: liqian2014@cdut.edu.cn

引用格式: 魏思维, 李谦, 姜杰, 何俊杰, 李同意, 2026. 基于最优相关性的钻速建模算法. 地球科学, 51(8): 3102-3117.

Citation: Wei Siwei, Li Qian, Jiang Jie, He Junjie, Li Tongyi, 2026. Correlation-Optimized Modeling Algorithm for Rate of Penetration. *Earth Science*, 51(8): 3102-3117.

(window size: 400, step size: 100) was introduced to achieve local dynamic modeling, with a Random Forest model serving as the core for regression prediction. Experimental results demonstrate that the Chatterjee algorithm performed best in screening the top 5 features. The architecture combining the sliding window technique with the Random Forest model achieved an average R^2 of 0.985 on the window test set of the 10th test well, representing a significant improvement over global static modeling ($R^2=0.96$). The resulting optimal correlation ROP modeling algorithm can adapt to abrupt formation changes and effectively capture local correlations between parameters. This provides a high-precision solution for real-time ROP optimization and it provides effective technical support for the transformation of drilling engineering from empirical trial-and-error to data-driven decision-making.

Key words: rate of penetration prediction; correlation; machine learning; dynamic modeling; basic data processing.

0 引言

针对钻速预测与优化这一难题,国内外学者将机器学习与传统理论相结合,取得了显著进展。Bourgoyne and Young(1974)最早建立了多元回归的经典框架,而 Hegde *et al.*(2017)的研究证明了数据驱动模型在准确性上优于物理模型。为进一步提升精度,甘超(2019)、Ehsan and Ebrahim(2022)及严顾鑫(2022)分别利用改进的蝙蝠算法和粒子群算法(PSO)优化传统模型或神经网络,显著降低了预测误差。在算法演进方面,张海军等(2022)与王若禹(2023)通过集成随机森林与遗传算法,有效解决了参数分类与优化问题;曾小龙等(2023)则进一步引入地层特征分类以提升模型适应性。随着深度学习的发展,侯霖莉等(2024)基于绿泥石矿物微量元素结合机器学习方法能够实现对矿床类型的判别,对于区域尺度找矿勘查快速评价具有重要的指示意义,朱景宝等(2025)提出了基于机器学习和迁移学习的现地地震动峰值预测方法,在一定程度上提高现地地震预警地震动峰值预测的可靠性,任辉(2023)、高家轩(2022)利用RNN和DNN实现了时序与高精度的钻速预测,焦通(2023)与李博志(2024)更是引入注意力机制改进CNN模型,在拟合优度上表现优异。此外,针对实时性需求,黄小龙等(2021)与曹祺(2023)构建了实时更新的优化算法,配合周长春等(2022)提出的特征选择策略,为应对复杂井下环境提供了智能化解决方案。

钻井工程数据具有高维度、强噪声、非平稳和工况多变等特点,这对其有效分析与处理提出了特殊挑战。近年来,针对钻井数据的清洗、降噪、缺失值补偿及特征提取等预处理方法已取得系列进展(如基于小波变换的去噪、主成分分析的降维等),为后续建模奠定了数据基础。在此背景下,数据科学中的相关性分析与维度约简技术进一步提供了关键的方法论支撑。殷辛瑜(2023)与宋金玉(2024)在生物信息及区间数

据领域的研究表明,优化相关性度量能有效减少冗余特征。针对复杂变量关系,张亚红(2017)与杨甲森(2019)利用最大信息熵与MIC技术提升了非线性关系的检测能力。特别值得注意的是,Chatterjee(2021)提出了一种新的相关系数 ϵ ,克服了传统Pearson系数在非单调或非线性关系中的局限性,为大规模数据分析提供了理论完备的工具。此类方法已在多领域得到验证,如庞力源(2023)在空间数据、彭绍炜(2023)在时间序列以及董丽洁(2024)在气象模式提取中的应用,均证明了先进统计学方法在解析复杂系统依赖关系方面的有效性。

综上所述,虽然当前的钻速预测模型已逐渐从单一回归向深度学习与集成算法演进,但在面对极端非线性与高维多源异构数据时,仍存在泛化能力不足与解释性欠缺的瓶颈。现有研究多侧重于算法模型的堆叠,而往往忽视了数据输入端特征相关性的深度挖掘。

因此,本文旨在结合最新的相关性分析理论与深度学习技术,构建地质—工程参数的跨尺度关联模型。特别地,针对钻井参数间普遍存在的非单调、非线性依赖关系(如钻压与钻速在不同岩性段呈现先增后减的复杂模式),传统Pearson系数无法检测非线性关联,Spearman与Kendall系数虽能捕捉单调关系但仍难以处理非单调或非函数型依赖。通过引入Chatterjee(2021)提出的新型相关系数 ϵ ,该系数能够探测任意形式的依赖关系(包括非单调、非对称关联),从而更全面地识别影响钻速的关键驱动因子。在此基础上,通过优化特征选择与算法架构,进一步提升钻速预测的精度与鲁棒性,为实现钻井作业的科学高效提供有力的理论与技术支撑。

1 分析数据综述

1.1 对原始数据集的量化描述

本文所采用的原始数据均来源于10口实际油

气井的工程记录,涵盖了钻井工程中涉及的40余项关键参数.根据参数特性,可将其划分为五大类别:操作参数、钻井液性能参数、地质条件参数、钻头参数以及提速钻具参数.所选取的10口井垂直深度范围分布在1 000 m至4 000余m之间.操作参数主要包括井眼直径、机械钻速、钻压、扭矩、大钩载荷、立管压力、泵排量、钻井液入口密度等共计14个指标.其中,钻井液入口密度是指钻井液在泵入钻柱入口处的实时密度,由在线密度计连续监测,反映动态循环条件下的密度值.钻井液性能参数则涉及钻井液密度、固相含量、滤失量、稠度系数(k)等18个特性指标.其中,钻井液密度是指钻井液在实验室条件下(常压、常温或特定温度压力)测得的静态密度,通常作为基准设计值.两者区分的原因在于:入口密度受循环压耗、温度、含气量等因素影响,会与静态设计密度产生偏差,同时纳入两个参数有助于模型捕捉动态循环效应对钻速的影响.地质条件参数包含了孔隙压力、破裂压力、上覆岩层压力及地层岩性等6类信息.钻头参数由喷嘴数量、等效直径、内排齿磨损程度和外排齿磨损程度共同描述.而提速工具参数则涵盖了液力提速马达、水力脉冲发生器以及钻头齿形结构等关键特征,在钻井作业过程中,针对每口井都按照每米间隔进行一次系统的数据采集与测试,10口井累计获取了所有参数的有效实测数据,最终构建的数据集总共包含21 912组有效数据记录.

1.2 基本的数据预处理

数据预处理的总体目的是将原始钻井记录转化为适用于机器学习建模的高质量数据集.与常规数据处理相比,钻井数据预处理具有以下特殊性:钻井数据为深度连续的时序信号,传感器采样频率不统一,易产生缺失和错位;井下工况复杂(如起下钻、划眼、循环等),不同工况下参数统计特性差异显著,需结合工程上下文处理;异常值往往源于真实物理事件(如振动、钻井液脉冲干扰),不能简单剔除,需结合IQR等稳健方法进行甄别.基于上述特点,本文设计了以下预处理流程:

(1)删除无效列:在数据预处理阶段,移除了“位置”“井号”以及“岩性”这3列特征.其中,“位置”和“井号”与目标变量钻速(ROP)之间缺乏直接的物理或工程关联,而“岩性”列所包含的信息已由数据集中已有的“岩性编号”充分表征.因此,删除这些列有助于精简特征空间,降低数据维度,从而提

升后续计算的效率,并减少不相关或冗余特征对模型训练的潜在干扰;

(2)进行初步的数据清洗:在进行数据预处理时,首先移除了所有元素均为空值(NAN)的行,同时删除了非空数据占比低于70%的列.全空行不包含任何有效信息,而缺失程度过高的列所承载的原始信息量有限,若保留可能会引入噪声.进行上述操作,能够有效过滤无效或信息不足的数据单元,从而提升数据集的整体质量与一致性,为后续建模或分析提供更可靠、更具代表性的数据基础;

(3)缺失值的填充:在处理数据中的缺失值时,采用了中位数填充法.选择中位数而非均值或众数的主要理由如下:钻井数据常含有离群值(如传感器尖峰干扰),均值对离群值敏感,而中位数具有稳健性,能更好地代表数据的典型水平;钻井参数多为连续变量,众数填充不适用;与线性插值相比,中位数填充不依赖时间顺序,避免因异常段插值传播误差.因此,中位数填充是在保证稳健性的前提下处理钻井数据缺失的合理选择.中位数是一种用于衡量数据集中趋势的统计量,其计算方式是将所有观测值按升序或降序排列后,取位于最中间的那个数.如果数据点个数为偶数,则中位数是中间两个数的平均值,数学定义为对于有序数据集 $x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \dots \leq x_{(n)}$, 计算如公式(1)所示:

$$\text{Median} = \begin{cases} x_{(\frac{n+1}{2})}, & \text{若 } n \text{ 为奇数} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}, & \text{若 } n \text{ 为偶数} \end{cases} \quad (1)$$

(4)离群值处理:离群值处理至关重要,其核心在于依据数据集的整体分布规律,甄别并修正那些显著偏离主流数据群的异常点,这些异常值通常源于多种工程实际情况:例如,传感器可能因钻头剧烈振动而发生瞬时过载,导致钻压读数异常偏高;也可能是起下钻等作业环节中,钻速实际已归零,但系统未能正确标注相应工况;此外,钻井液脉冲信号在传输过程中若受到干扰,亦可能引起扭矩等参数的突发性跳变,具体使用的方法是IQR(四分位距)方法,核心原理如下:设某参数序列为, $X = \{x_1, x_2, \dots, x_n\}$, 离群值处理流程的数学表达式为:

四分位数的计算如公式(2)所示:

$$\begin{cases} Q_1 = \text{quantile}(X, 0.25) \\ Q_3 = \text{quantile}(X, 0.75) \\ \text{IQR} = Q_3 - Q_1 \end{cases} \quad (2)$$

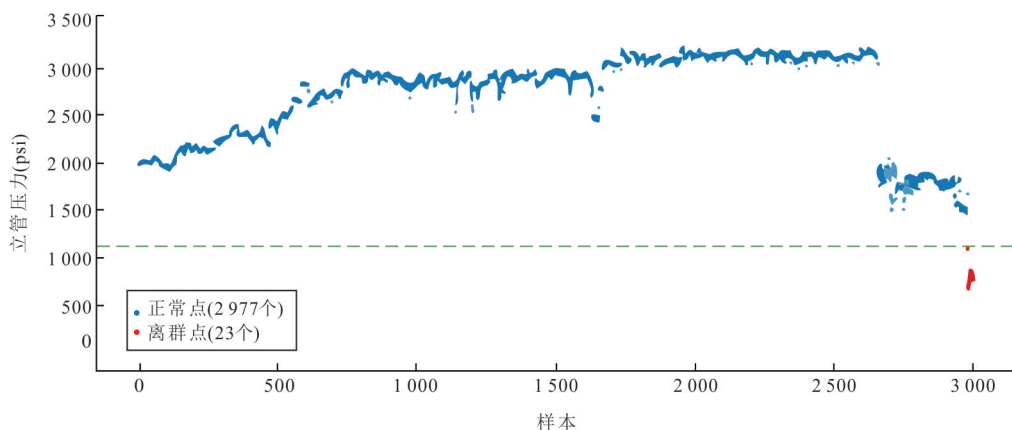


图1 离群值处理

Fig. 1 Outlier handling

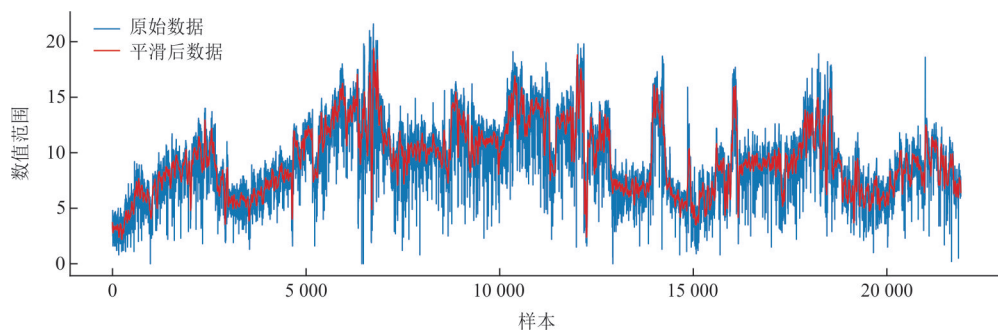


图2 数据平滑处理

Fig. 2 Data smoothing

边界值确定如公式(3)所示:

$$\begin{cases} \text{Lower Bound} = Q_1 - k \times IQR \\ \text{Upper Bound} = Q_3 + k \times IQR \end{cases} \quad (3)$$

其中:经验系数 $k=1.5$, Q_1 是将数据从小到大排序时,位于25%的数,即下四分位数; Q_3 是处于75%的数,即上四分位数.以下是一号井的立管压力的部分数据为例,(注:1 psi=0.006 894 76 MPa)其离群值处理具体情况如图1所示(图中红点为离群值):

(5)数据平滑处理:本文选用Savitzky-Golay滤波器对原始信号进行平滑去噪,该滤波器基于局部多项式最小二乘拟合原理,通过在滑动窗口内对数据进行多项式拟合,能在有效抑制高频随机噪声的同时,较好地保留信号本身的峰值、趋势等关键形态特征.具体情况如公式(4)所示:

$$y_i = \frac{1}{N} \sum_{k=-M}^M \alpha_k x_{i+k}, \quad (4)$$

其中: N 为窗口大小, α_k 为多项式系数.本文主要是通过滑动窗口内构建三次多项式(阶数过高会导致过拟合)用以逼近数据趋势,这里以钻压列为例,

数据平滑处理如图2所示:

(6)归一化处理:为消除钻井参数间量纲不统一对模型造成的偏差,本文采用Min-Max标准化方法对各特征列进行处理.该方法通过线性变换将原始数据缩放至 $[0, 1]$ 区间内,具体公式如式(5)所示:

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}. \quad (5)$$

在使用了Min-Max标准化之后,所有的特征会被压缩到相同的尺度空间,数据集将会平移到 $[0, 1]$ 的区间内,数据分布不变.

2 相关性理论与方法对比

2.1 4种基本的相关性理论

为全面评估钻井参数与钻速之间的关联类型,本文选取了4种具有代表性的相关性度量方法,其选取依据如下:Pearson相关系数作为经典的线性相关度量,是多数研究的基准;Spearman和Kendall相关系数均为非参数方法,能够检测单调(不一定线

性)关系,且对异常值不敏感,其中 Kendall 更适用于小样本和含重复值的数据;Chatterjee 相关系数是近年提出的新型度量,能够探测任意形式的依赖关系(包括非单调、非对称、非函数型),弥补了前三者的局限.通过这 4 种方法的对比,可以系统揭示参数与钻速之间从线性、单调到复杂非线性关联的全貌.

(1) Pearson 理论: Pearson 相关系数由卡尔·皮尔逊于 19 世纪末提出,用于衡量两个连续变量之间线性关系的强度.其核心思想是通过计算两变量的协方差与各自标准差乘积的比值,以标准化形式反映二者协同变化的程度,如公式(7)所示:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (6)$$

其中: x_i 和 y_i 表示第 i 个观测值, $\bar{x}$ 和 $\bar{y}$ 表示变量 x 和 y 的样本均值, n 代表样本量.该系数的取值范围在 -1 到 1 之间,当系数绝对值 $|r| \geq 0.7$ 时,通常认为变量间存在高度相关性,若 $0.3 \leq |r| < 0.7$,则为中度相关,而 $|r| < 0.3$,相关性较低.其计算假设数据服从双变量正态分布且具有同方差性,对异常值极为敏感,例如单个极端点可能使结果严重偏离真实相关性.

(2) Spearman 理论: Spearman 相关系数由查尔·斯皮尔曼于 20 世纪初提出,作为一种非参数统计方法,其核心在于通过分析变量秩次之间的关联,来评估两变量之间单调关系的强弱,无需数据满足特定的分布假设,该方法将原始数据转换为秩次,继而计算这些秩次之间的 Pearson 相关系数,其公式简化之后如式(8)所示:

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}, \quad (7)$$

其中: d_i 为变量对应数据的秩次差,秩次转化过程如下:

① 对每个变量的观测值分别独立排序(升序或降序);

② 相同值赋予平均秩次(如数值相同则取秩次均值);

③ 计算每对数据点的秩次差 $d_i = \text{rank}(x_i) - \text{rank}(y_i)$.

例如:若 $x = [10, 20, 20, 30]$,其秩次就是 $[1, 2.5, 2.5, 3]$ (因为第 2 个和第 3 个值相同)

该系数的取值范围在 -1 到 1 之间,当系数绝对值 $|\rho| \geq 0.7$ 时,通常认为变量间存在高度相关性,若

$0.4 \leq |\rho| < 0.7$,则为中度相关,而 $|\rho| < 0.4$ 时,则表明相关性较弱.

(3) Kendall 理论: Kendall 相关系数由莫里斯·肯德尔于 1938 年提出,其核心思想是通过比较变量间数据对的协同性与不协同性来衡量两个序列之间单调关系的强度.该系数在计算时,基于数据对的协同对数与不协同对数之差进行标准化处理.它具有对异常值不敏感、在小样本条件下仍能保持稳定等特点,因此特别适用于包含较多重复取值的顺序数据,计算如式(8)所示:

$$\tau = \frac{C - D}{\frac{1}{2}n(n-1)}, \quad (8)$$

其中: C 为协同对数(两变量排序一致), D 为不协同对数(排序相反),其中协同对指两变量排序一致的观测对,不协同对则相反,具体定义如下:

对任意两个数据点: (x_i, y_i) 和 (x_j, y_j) , 协同对 (Concordant): 若 $(x_j - x_i)(y_j - y_i) > 0$; 不协同对 (Discordant): 若 $(x_j - x_i)(y_j - y_i) < 0$; 结 (Tie): 若 $x_j = x_i$ 或 $y_j = y_i$.

该系数的取值范围在 -1 到 1 之间,当系数绝对值 $|\tau| \geq 0.5$ 时,通常认为变量间存在高度相关性,若 $0.2 \leq |\tau| < 0.5$,则为中度相关,而 $|\tau| < 0.2$,则表明相关性较弱.

(4) Chatterjee 理论: Chatterjee 相关系数是 Sourav Chatterjee 于 2021 年提出的一种新型非参数相关性度量指标.该方法的核心创新在于能够突破传统单调性关系的限制,探测包括非函数型,非对称关联在内的任意形式依赖关系,其核心思想是通过按 X 排序后观察 Y 秩次的局部波动性来量化依赖强度,若 Y 是 X 的函数,按 X 排序后 Y 的秩应严格递增(或递减),此时 $\epsilon = 1$,但是若 X 与 Y 独立, Y 的秩随机波动, $\epsilon = 0$,计算公式如式(10)所示:

$$\epsilon(X, Y) = \frac{\sum_{i=1}^{n-1} |R(y_{(i+1)}) - R(y_{(i)})|}{n^2 - 1}, \quad (9)$$

其中: $R(y_{(i)})$ 是按 X 排序后 Y 的秩,该系数的取值范围在 0 到 1 之间,当系数值 $\epsilon \geq 0.5$ 时,通常认为变量间存在高度相关性,若 $0.3 \leq \epsilon < 0.5$,则为中度相关,而 $\epsilon < 0.3$,则表明相关性较弱.

3 种传统相关系数 (Pearson/Spearman/Kendall) 的适用性与补充说明: 上述 3 种传统相关系数在不同场景下具有各自优势: Pearson 系数适用于线性关系检测, Spearman 和 Kendall 系数适用于单调关

表 1 相关性系数绝对值
Table 1 Absolute value of the correlation coefficient

算法参数	Pearson	Spearman	Kendall	Chatterjee	算法参数	Pearson	Spearman	Kendall	Chatterjee
井深	0.70	0.74	0.52	0.37	稠度系数	0.12	0.30	0.22	0.47
井径	0.66	0.70	0.57	0.34	氯离子含量	0.05	0.15	0.10	0.19
钻压	0.04	0.03	0.02	0.07	钙离子含量	0.02	0.01	0	0.11
大钩载荷	0.66	0.60	0.42	0.33	膨润土含量	0.46	0.44	0.31	0.30
回转转速	0.30	0.27	0.18	0.24	固相含量	0.70	0.77	0.56	0.59
扭矩	0.69	0.66	0.50	0.36	含砂量	0.17	0.26	0.18	0.13
立管压力	0.13	0.08	0.06	0.11	地震速度	0.39	0.33	0.23	0.33
泵量	0.69	0.75	0.54	0.49	孔隙压力	0.71	0.75	0.53	0.41
钻头工作时间	0.22	0.13	0.11	0.12	破裂压力	0.64	0.71	0.50	0.37
泵工作时间	0.35	0.29	0.21	0.12	上覆压力	0.68	0.73	0.51	0.39
钻井液进口比重	0.71	0.78	0.56	0.55	岩性编号	0.23	0.20	0.14	0.11
钻井液出口比重	0.71	0.78	0.56	0.54	喷嘴数量	0.01	0	0	0.01
钻井液入口温度	0.12	0.06	0.05	0.14	等效直径	0.58	0.56	0.40	0.38
钻井液出口温度	0.19	0.17	0.12	0.12	内排磨损	0	0	0	0
屈服值	0.59	0.56	0.39	0.40	外排磨损	0.20	0.24	0.19	0.14
钻井液密度	0.72	0.79	0.58	0.57	液力提速马达	无	无	无	0
马氏漏斗粘度	0.46	0.42	0.29	0.30	水力脉冲	无	无	无	0
塑性粘度	0.60	0.65	0.46	0.40	钻头齿形	无	无	无	0
3 转读数	0.57	0.63	0.46	0.30	稠度系数	0.12	0.30	0.22	0.47
6 转读数	0.58	0.64	0.46	0.32	氯离子含量	0.05	0.15	0.10	0.19
10 s 静切力	0.50	0.53	0.37	0.28	钙离子含量	0.02	0.01	0	0.11
10 min 静切力	0.55	0.61	0.43	0.36	膨润土含量	0.46	0.44	0.31	0.30
滤失量	0.70	0.66	0.46	0.46	固相含量	0.70	0.77	0.56	0.59
泥饼厚度	0.61	0.55	0.44	0.24	含砂量	0.17	0.26	0.18	0.13
pH	0.62	0.65	0.49	0.34					
流性系数	0.01	0.03	0.02	0.33					

系检测且对异常值不敏感。然而在钻井工程中,钻速(ROP)与某些参数(如钻压、钻井液密度)的关系往往呈现非单调特征。例如,在泥质粉砂岩层段增大钻压可提升钻速,但进入粉砂质泥岩层后继续增大钻压反而导致钻速下降——这种“先增后减”的模式超出了单调关系的范畴。对于此类非单调或非函数型依赖,上述 3 种系数的检测能力有限。相比之下,Chatterjee(2021)提出的相关系数 ϵ 通过按 X 排序后观察 Y 秩次的局部波动来量化依赖强度,能够有效探测包括非单调、非对称、非函数型在内的任意形式关联。因此,Chatterjee 系数并非否定传统方法的有效性,而是提供了一种互补的工具,特别适用于钻井数据中复杂非线性关系的挖掘,为后续特征筛选提供更全面的依据。

2.2 4 种相关性理论应用对比

利用 4 种相关性算法(Pearson/Spearman/Kendall/Chatterjee)对钻井参数与钻速(ROP)进行全局

的相关性分析,可以得到各个特征参数与钻速之间的相关性系数绝对值,并且结合 4 种算法的相关性强度划分,可以得到各算法的高度相关、中度相关以及低度相关的参数所占的比例,具体结果如表 1 和图 3 所示:

对于 Pearson 算法而言,各参数与钻速的相关性系数绝对值分布于 0 至 0.72 之间,中位数为 0.53。根据相关性强度划分,高度相关($|r| \geq 0.7$)的参数占比为 14.3%,中度相关($0.3 \leq |r| < 0.7$)的参数占比为 47.6%,而低度相关($|r| < 0.3$)的参数占比为 35.6%,分析进一步识别出与钻速关联最为显著的关键参数,包括滤失量、井深、钻井液密度、孔隙压力、固相含量、钻井液入口比重以及钻井液出口比重,这些参数的相关性系数绝对值均超过了 0.7。

对于 Spearman 算法而言,各参数与钻速的相关性系数绝对值分布于 0 至 0.79 之间,中位数为 0.54。

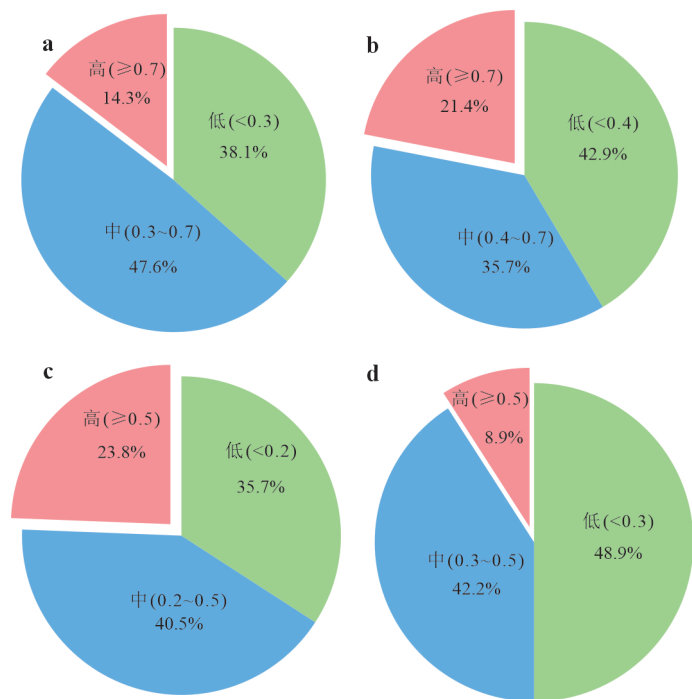


图3 高、中、低相关性参数所占比例

Fig. 3 The proportions of parameters with high, medium, and low correlations
a. Pearson算法; b. Spearman算法; c. Kendall算法 d. Chatterjee算法

根据相关性强度划分,高度相关($|\rho| \geq 0.7$)的参数占比为 21.4%,中度相关($0.4 \leq |\rho| < 0.7$)的参数占比为 35.7%,而低度相关($|\rho| < 0.4$)的参数占比为 42.9%,分析进一步识别出与钻速关联最为显著的关键参数,包括钻井液密度、固相含量、钻井液出口比重、钻井液进口比重、泵量、井深、上覆压力、孔隙压力、井径以及破裂压力,这些参数的相关性系数绝对值均超过了 0.7.

对于 Kendall 算法而言,各参数与钻速的相关性系数绝对值分布于 0 至 0.58 之间,中位数为 0.38. 根据相关性强度划分,高度相关($|\tau| \geq 0.5$)的参数占比为 23.8%,中度相关($0.2 \leq |\tau| < 0.5$)的参数占比为 40.5%,低度相关($|\tau| < 0.2$)的参数占比为 35.7%,分析进一步识别出与钻速关联最显著的关键参数,包括钻井液密度、井径、泵量、固相含量、钻井液进口比重、钻井液出口比重、井深、上覆压力、孔隙压力、破裂压力以及扭矩,这些参数的相关性系数绝对值均超过了 0.5.

对于 Chatterjee 算法而言,各参数与钻速的相关性系数绝对值分布于 0 至 0.59 之间,中位数为 0.30. 根据相关性强度划分,高度相关($\epsilon \geq 0.5$)的参数占比为 8.9%,中度相关($0.3 \leq \epsilon < 0.5$)的参数占比

为 42.2%,低度相关($\epsilon < 0.3$)的参数占比为 48.9%,分析识别出与钻速关联最为显著的关键参数,包括固相含量、钻井液密度、钻井液进口比重,钻井液出口比重,这些参数的 Chatterjee 相关系数均高于 0.5.

在进行全局相关性的分析之后,结合 4 种相关性计算方法各自高,中,低的阈值,可绘制出 4 种相关性确定的高中低相关性参数个数对比如图 4 所示. 对比 4 种相关性的全局分析可知,不同方法的结果存在明显差异: Kendall 算法识别出的强相关参数最多(10 个),而 Chatterjee 算法识别出的强相关参数最少(4 个);在中相关性参数数量上, Pearson 算法(20 个)与 Chatterjee 算法(19 个)方法较为接近;在低相关性参数方面, Chatterjee 算法最多(22 个),而 Kendall 算法则最少(15 个).

综上所述:通过对钻速(ROP)与各类钻井参数进行 4 种相关性分析方法(Pearson/Spearman /Kendall/Chatterjee)的综合评估发现,尽管不同方法在系数大小与参数排序上存在一定差异,但钻井液密度、固相含量、钻井液进出口比重、泵量、井深、上覆压力及孔隙压力等参数在各类方法中均表现出显著的关联性,说明这些因素对钻速具有系统性影响. 其中,钻井液密度与固相含量在 4 种方法中均位列前五,其主要原因在于二者共同参与井壁压力平

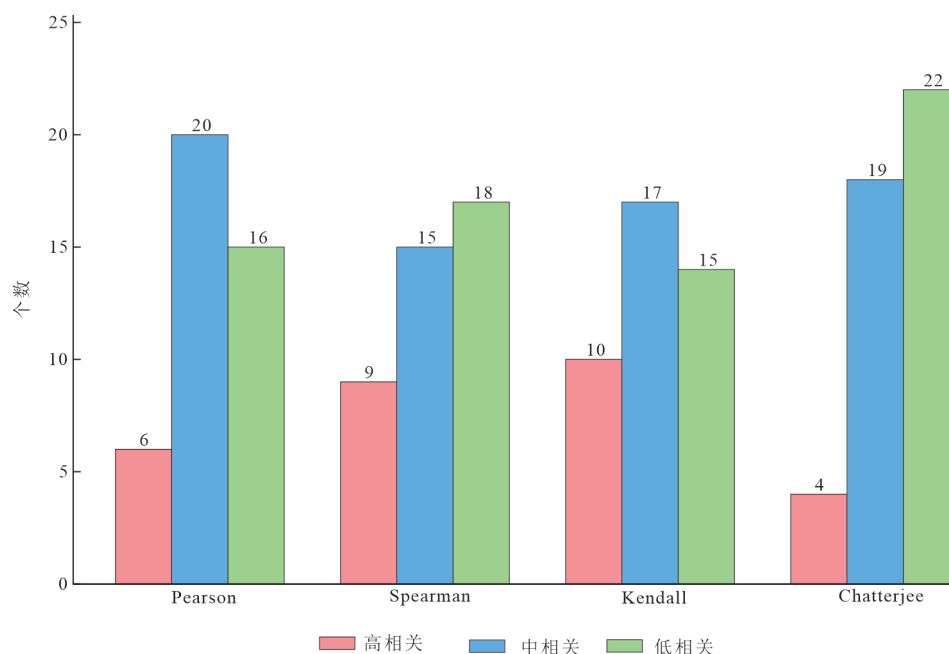


图4 高中低相关性参数个数对比图

Fig. 4 Comparison chart of the number of parameters with high, medium, and low correlations

衡的调控:钻井液密度通过静液柱压力维持井壁稳定,而固相含量过高则会引起滤失量增大和流变性能恶化,进而影响循环密度,最终影响钻速。钻井液进出口比重同时显著,反映出钻井液性能在循环过程中的动态变化对钻井系统具有累积效应。泵量与井深均呈现强相关性,说明二者可能共同调控环空压耗和钻头水力参数,尤其在深井条件下,泵量需随井深合理调整以抑制压力波动。地层力学参数(上覆压力与孔隙压力)的普遍显著性则源于其直接定义井壁力学环境,数值异常易引发井漏或井涌。值得注意的是,在Chatterjee系数分析中,流变参数(如稠度系数、塑性粘度)表现出较高关联性,暗示钻井液流变特性可能通过剪切速率分布和岩屑携带能力以非线性机制作用于钻速。不同方法结果的差异,既反映了参数与钻速之间关系的本质多样性(线性、单调非线性、复杂非参数关联),也提示参数间可能存在交互作用,共同影响钻速。

3 整体建模效果对比

在明确与钻速(ROP)相关的工程参数后,为了系统的评估不同相关性算法(Pearson/Spearman/Kendall/Chatterjee)在构建预测模型时的实际性能,并探究特征数量对预测精度的影响,从而确定适宜建模的特征维度,本文设计了如下验证流程:

(1)特征选择:鉴于4种相关性算法识别出的强相关参数数量均有限(均小于15个),为平衡模型可靠性与计算效率,避免引入噪声,将依据各算法得出的相关系数绝对值,分别选取排名前5至前15的特征参数作为输入,以此评估特征数量对模型性能的影响;

(2)模型选择:选用多层感知机和随机森林两种具有代表性的回归模型进行对比。多层感知机模型能够捕捉复杂的非线性关系,而随机森林模型则擅长处理特征交互的情况并具有较强的抗过拟合能力;

(3)数据划分与可复现性:实验数据来源于1至9号井经过预处理后的数据集。将其按80%的比例随机划分为训练集,剩余20%作为测试集。为确保实验结果的可复现性,设定了固定的随机种子,这意味着每次实验的随机数生成序列一致;

(4)评价指标:采用决定系数(coefficient of determination, R^2)作为核心评价指标。 R^2 的定义为:

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

其中: y_i 为真实值, $\hat{y}_i$ 为预测值, $\bar{y}$ 为真实值的均值。 R^2 的取值范围为 $[0, 1]$ (在拟合极差时可能为负),其物理意义是模型所解释的目标变量方差占总方差的比例。 R^2 越接近1,表明模型对钻速(ROP)变化

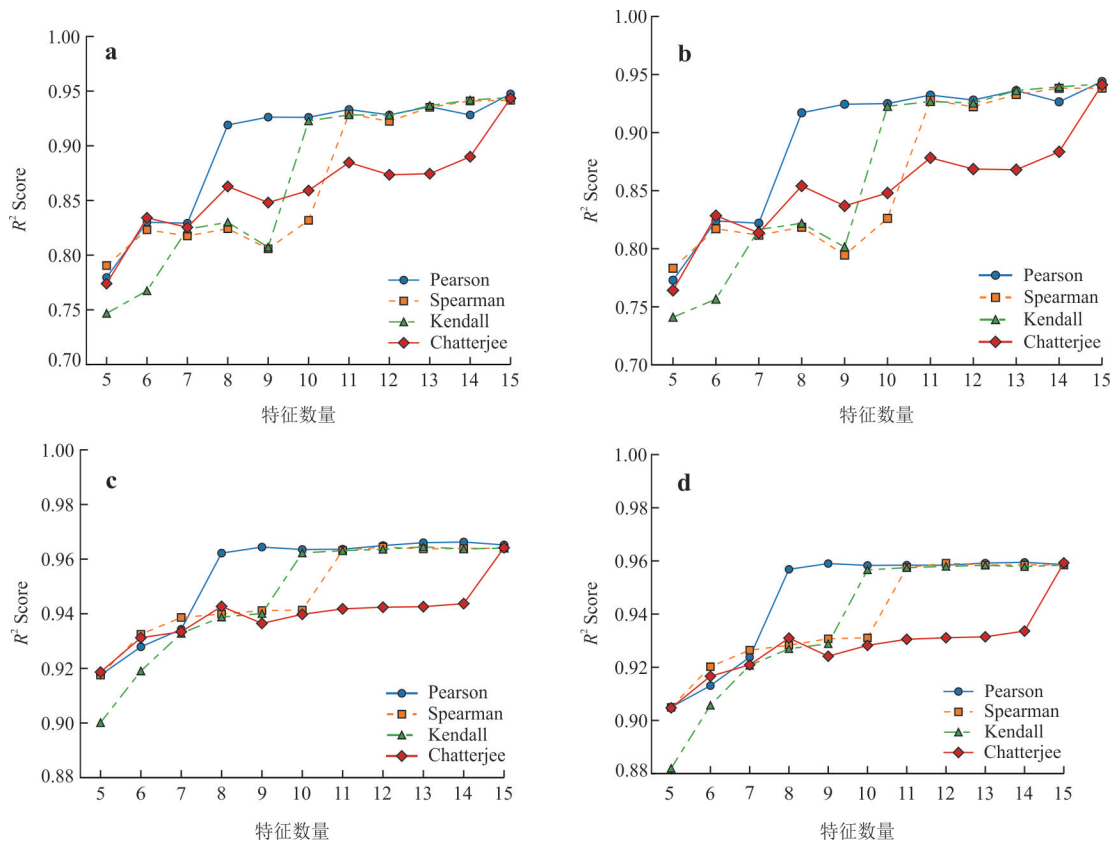


图 5 多层感知机和随机森林模型性能

Fig. 5 Performance of multilayer perceptron and random forest models

a. 多层感知机训练集; b. 多层感知机测试集; c. 随机森林训练集; d. 随机森林测试集

的解释能力越强,预测精度越高.该指标能够直观地比较不同特征数量组合及不同回归模型之间的预测效果,从而辅助确定最优的特征参数数量.

在完成上述步骤后,为清晰评估不同特征选择方法与回归模型的组合性能,本文通过绘制折线图对结果进行可视化分析.图中分别展示各方法在训练集与测试集上 R^2 随特征数量(从排名前5至排名前15)的变化趋势;横轴为特征数量,纵轴为 R^2 ,具体情况如图5所示:

结合图片分析可知,在选取排名前5至排名前15的特征进行建模时,不同相关性筛选方法对多层感知机与随机森林的预测性能影响有所差异.对于多层感知机模型而言,采用Pearson相关系数筛选特征的整体表现更优.当特征参数数量增至8个以上时,其训练集与测试集的 R^2 均可达到0.90以上;特征数达到10个后, R^2 进一步提升至0.93以上. Kendall算法在特征参数数量较少时表现相对较弱($R^2 < 0.80$),但随着特征参数数量的增加,其性能逐步改善. Chatterjee算法的表现则较为平稳, R^2 多集中于0.80与0.90之间.整体上看来,各算法在特征

参数数量达到11个后性能提升均趋于平缓.对于随机森林模型而言,4种特征筛选方法均表现出色且稳定,多数情况下 R^2 介于0.90与0.96之间,其中Pearson算法在特征参数数量较多的条件下 R^2 可维持在0.96左右.

综合比较表明,随机森林模型的整体预测性能优于多层感知机.在整体建模时,在特征筛选的方法上, Pearson相关系数表现最为可靠.此外,当特征参数数量达到10个后,继续增加特征对模型性能的改善并不显著,因此建议选取相关性排名前10的特征参数进行整体建模.

4 动态钻速建模算法

前述第3节基于全井段数据(1~9号井整体)的全局建模虽然取得了较好的预测效果(R^2 可达0.96),但该方法隐含了一个基本假设:全井段范围内钻速与各参数之间的相关性是恒定不变的.然而,钻井工程实践表明,随着地层岩性、压力系统、钻井液性能的沿程变化,参数与钻速之间的关联关

系往往表现出显著的井段差异性(例如,钻压在脆性地层中与钻速正相关,在塑性泥岩中则可能负相关或无关)。因此,全局静态模型难以适应地层突变带来的局部特征变化。为此,本节引入滑动窗口技术,将长井段数据划分为若干局部窗口,在每个窗口内独立进行相关性分析与建模,从而实现动态捕捉参数间局部关联的目标。

4.1 滑动窗口建模原理

钻井过程中,岩性、压力系统与钻井参数间的交互作用具有明显的井段差异性。例如,在灰色泥质粉砂岩层段,钻压增大通常能提升钻速(ROP);然而当岩性变为灰色粉砂质泥岩时,即使继续增加钻压,钻速反而下降。这种地层突变会导致基于全井段的全局相关性分析结果失真,因此,考虑对全井段数据进行分段处理,由此提出了滑动窗口理论,滑动窗口分析方法通过将连续钻井数据按固定长度分段,并沿深度方向滑动窗口进行局部化处理,从而动态捕捉参数间的关联特征,有效克服了全局分析的局限性,滑动窗口原理示意图如图6所示(图中黑色虚框线圈中的是原始窗口,红色实框线圈中的是滑动后的窗口):

4.2 窗口与特征参数优化

在进行滑动窗口分析的过程中,窗口大小的选择会直接影响分析结果的准确性和可靠性,过大的窗口或者是过小的窗口都会带来不同的弊端。当窗口过大时,数据量较大,分析时可能会掩盖局部的特征和短期的波动,例如在对地质数据进行分析时,若窗口大小超过了底层变化周期,可能无法捕捉到薄层(如10 m厚的高渗透层)对钻速(ROP)的影响,同时,计算效率也会降低,大的滑动窗口需要处理更多的数据,内存和计算时间也会陡增。当窗口过小时,数据量较小,模型对噪声会更加敏感,短期的波动(如仪器误差,钻井液脉冲)会被放大,小窗口训练的模型在未见数据上泛化能力较差,存在着过拟合的风险。与此同时,窗口内特征参数数量

的选择也尤为重要,在采用较大间窗口时,本身已包含较多信息量,此时若再纳入大量特征,模型复杂度和过拟合风险会剧增。因此,更需要在特征维度上进行精选,聚焦于最具影响力的核心参数,当窗口较小时,数据量有限,模型更容易受到噪声干扰。此时,精准选择少数关键特征显得尤为重要,有助于模型捕捉到数据中稳定且主要的驱动因素。

综上所述需要进行最优窗口大小和特征参数数量的动态选择,原始数据集是由10口油气井的数据组成,每口井包含近3 000条数据,为了不破坏地层的完整性以及防止数据量过少造成模型过拟合的风险,滑动窗口大小的讨论范围在100到1 000(接近一口井全部数据的三分之一)之间,每次窗口大小增加100,而特征参数数量的讨论范围在排名前5到排名前20(基本包含了4种相关性算法所有的高度相关和中度相关的特征参数)之间,最后通过窗口测试集平均 R^2 的比较来选择出最优滑动窗口大小和特征参数数量的估值。

(1)利用多层感知机建模计算结果如图7所示:

分析图7可知,在利用多层感知机建模时,窗口大小在200到1 000之间,特征参数数量在5到20之间时,窗口测试集平均 R^2 都低于0.80,明显低于之前的整体建模分析,这表明在不破坏地层的完整性以及防止数据量过少造成模型过拟合的风险的情况下,多层感知机模型并不适用于滑动窗口理论。

(2)利用随机森林建模计算结果如图8所示:

分析图8可知,在利用随机森林建模时,模型的整体表现都比较好,窗口大小在200到1 000之间,特征参数数量在5到20之间时,4种算法的窗口测试集平均 R^2 均在0.90之上,这表明在不破坏地层的完整性以及防止数据量过少造成模型过拟合的风险的情况下,随机森林模型更适用于滑动窗口理论,并且表现较为优秀。

对于Pearson算法、Spearman算法、Kendall算法而言,窗口大小在800,特征参数数量为15个时,窗口测试集平均 R^2 较高,能达到0.97之上,对于Chatterjee算法而言,窗口大小在400,特征参数数量在5个时,窗口测试集平均 R^2 较高,接近0.98。值得注意的是,Chatterjee算法识别出的强相关参数仅4个,且最优建模时仅需排名前5的特征即可达到最高预测精度(窗口测试集平均 $R^2=0.979$),而Pearson/Spearman/Kendall算法需要15个特征才能达到相近效果。产生这一差异的根本原因在于不同相

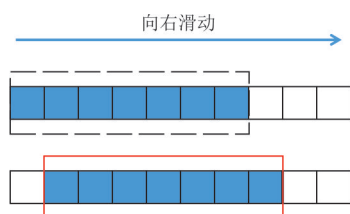


图6 滑动窗口原理示意图

Fig. 6 Sliding window principle

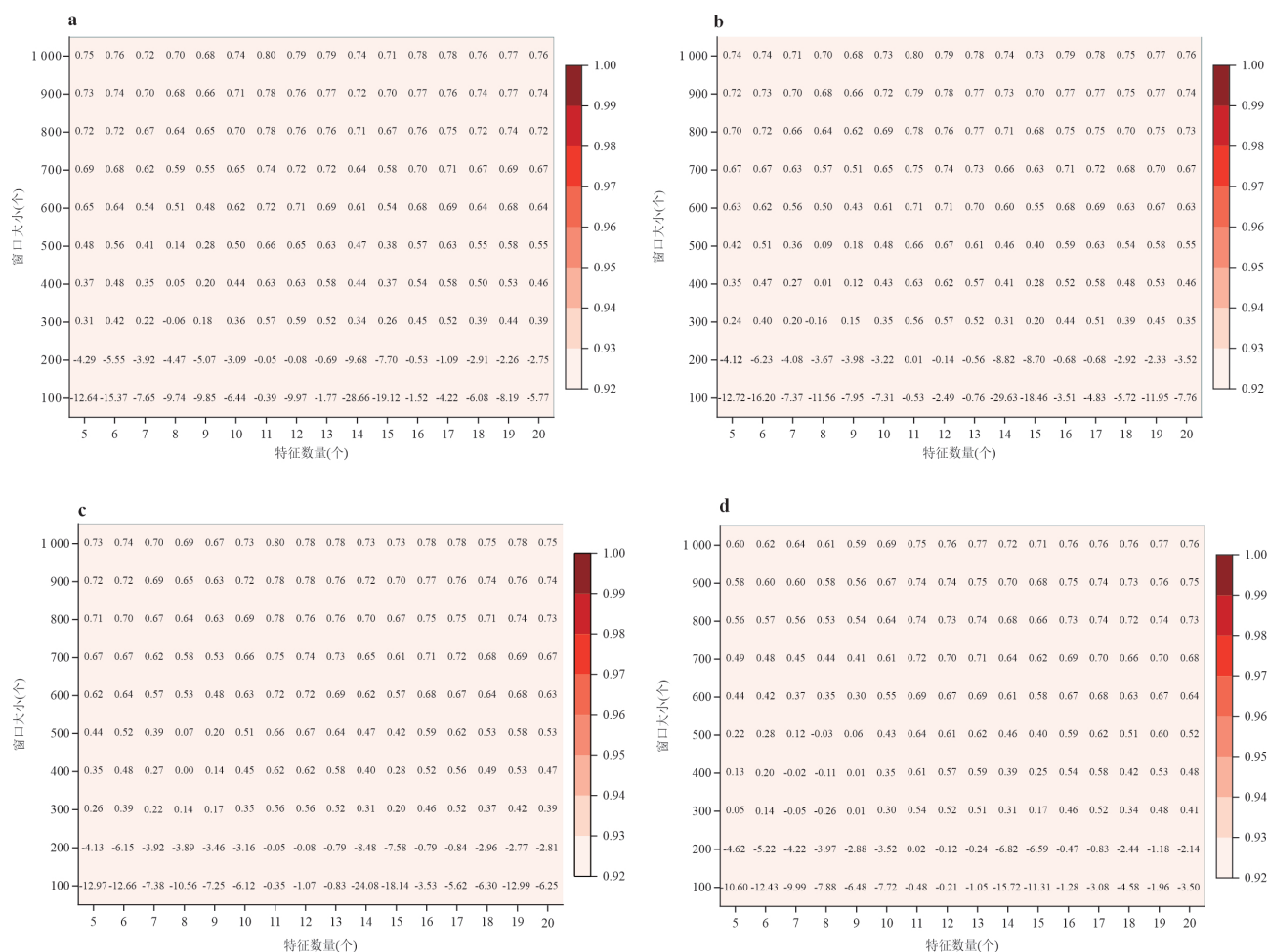


图 7 基于多层感知机的动态选择

Fig. 7 Dynamic selection based on multilayer perceptron

a. Pearson算法; b. Spearman算法; c. Kendall算法; d. Chatterjee算法

关系系数的信息筛选机制不同, Chatterjee系数的核心创新在于, 它通过局部秩次波动来探测 Y 对 X 的依赖程度. 若 X 与 Y 之间存在确定的函数关系 (无论线性或非线性), ϵ 趋近 1; 若 Y 独立于 X , ϵ 趋近 0. 这种机制使得 Chatterjee 系数对非对称、非函数型的伪相关具有天然的抑制作用——它不会因为两个变量同时随第 3 个变量 (如井深) 单调变化而给出虚高的相关性评分. 因此, Chatterjee 筛选出的 Top5 特征 (如固相含量、钻井液密度) 往往是钻速最直接的物理驱动因子, 而非伴随性共变参数.

综上所述: 对于滑动窗口理论而言, 随机森林比多层感知机表现更加优秀, 同时在利用前 3 种算法 (Pearson/Spearman/Kendall) 建立随机森林模型的过程中, 滑动窗口的大小在 800 比较合适, 特征参数数量选取为 15 个比较合适, 利用 Chatterjee 算法建立随机森林模型的过程中, 滑动窗口的大小在 400 比较合适, 特征参数数量选取为 5 个比较合适.

4.3 模型架构

在数据预处理阶段结束之后, 进行滑动窗口的划分, 在之前的分析中可知, 对于滑动窗口理论而言, Pearson、Spearman 和 Kendall 算法的窗口大小选为 800 合适, 步长选为 100, 每个窗口使用 3 种相关性系数算法 (Pearson/Spearman/Kendall) 计算相关性系数, 按照系数绝对值大小筛选特征参数, 前文分析出特征参数个数为 15 比较合适, 之后每个窗口每种算法利用分别得到的 15 个参数和钻速 (ROP) 建立随机森林模型, 而对于 Chatterjee 算法而言, 窗口大小选为 400, 步长选为 100, 每个窗口使用 Chatterjee 算法计算相关性系数, 按照系数大小筛选特征参数, 前文分析出特征参数个数为 5 比较合适, 之后每个窗口利用得到的 5 个参数和钻速 (ROP) 建立随机森林模型. 随机森林模型中决策树的个数为 50, 最后根据这 50 棵决策树的平均值输出最终钻速 (ROP) 预测值, 再进行 R^2 的计算得出模型预测精

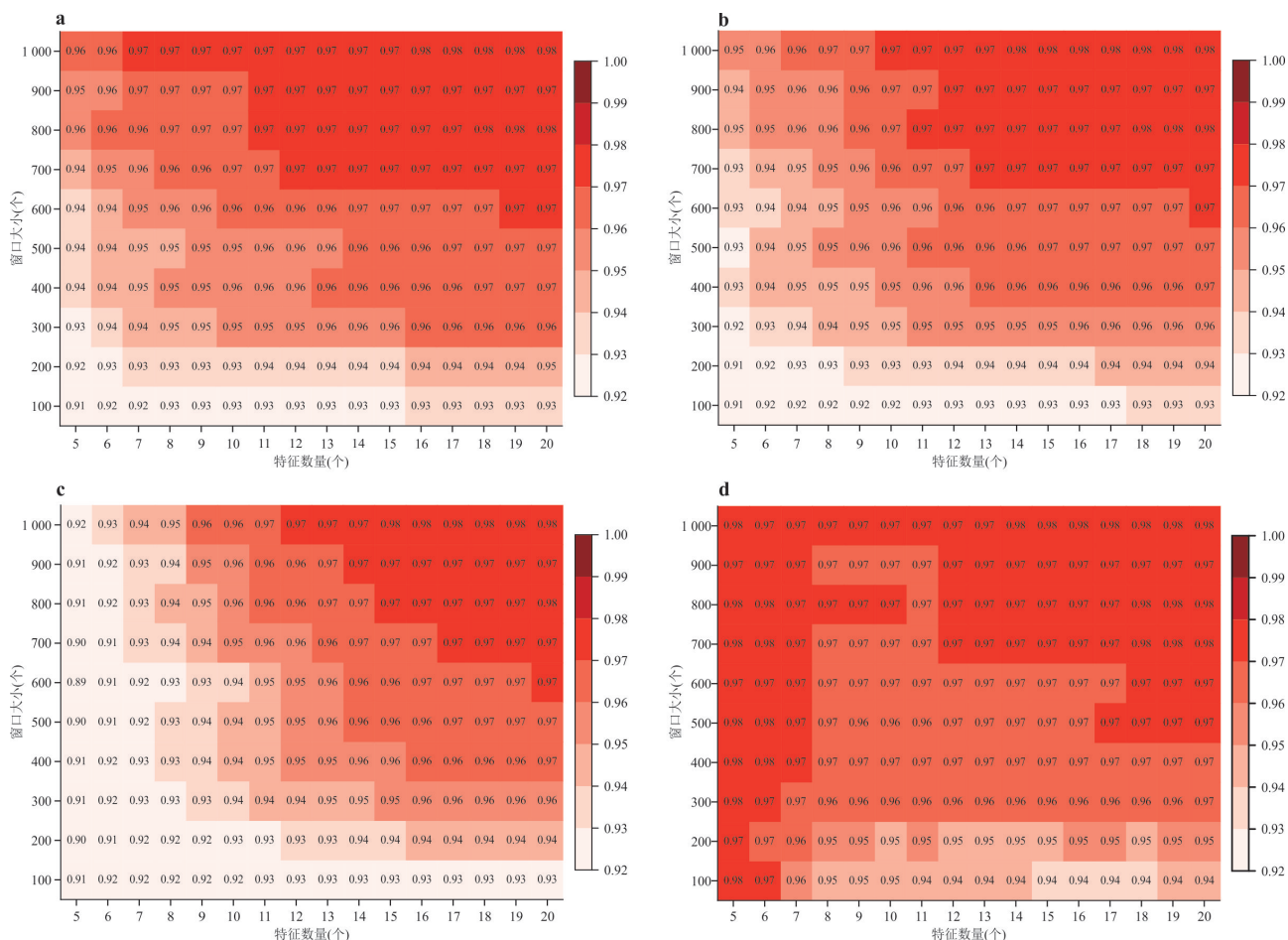


图8 基于随机森林的动态选择

Fig. 8 Dynamic selection based on random forest

a. Pearson算法;b. Spearman算法;c. Kendall算法;d. Chatterjee算法

表2 各方法性能统计汇总

Table 2 Statistical summary of performance of various methods

算法	窗口大小	特征数量	测试集均值	测试集标准差	最佳测试 R^2
Pearson	800	15	0.973 8	0.026 7	0.998 8
Spearman	800	15	0.973 9	0.026 2	0.998 6
Kendall	800	15	0.971 6	0.027 4	0.998 6
Chatterjee	400	5	0.979 0	0.020 8	0.999 1

度. 滑动窗口模型架构如图9所示:

4.4 模型训练与测试

前文第4.3节已给出两个基于相关性的动态分析模型架构,现在依据这两个模型架构对前九口油气井的数据进行训练和测试,然后对两个模型架构的窗口测试集 R^2 进行比较,方便找到基于相关性分析的最优模型架构,具体的运算结果如图10以及表2所示:

分析上面的测试集性能对比图,窗口优于全局

比例图以及各方法性能统计汇总表可知:利用滑动窗口模型分析数据时,模型的窗口测试集平均 R^2 均比全局测试集 R^2 高(全局测试集 R^2 为之前算出的0.96),均能达到0.97以上,表现十分优秀,Chatterjee算法的窗口测试集平均 R^2 甚至接近0.98,在窗口优于全局的比例之中,4种算法均达到了70%之上,最高的为Chatterjee算法,达到了98.4%,在最佳测试 R^2 之中,也是Chatterjee算法最高,达到了0.9991,同时Chatterjee算法的滑动窗口大小为400,单个窗口数据量较小,运算较快,特征参数数量选取的是排名前5的参数,不会引入过多的噪声,证明在这4个相关性算法之中,利用Chatterjee算法进行特征参数的筛选(选取每个窗口排名前5的特征参数),搭配滑动窗口理论(窗口大小为400,步长为100)并建立随机森林模型是比较优秀的方法,能比较准确的预测钻速(ROP).

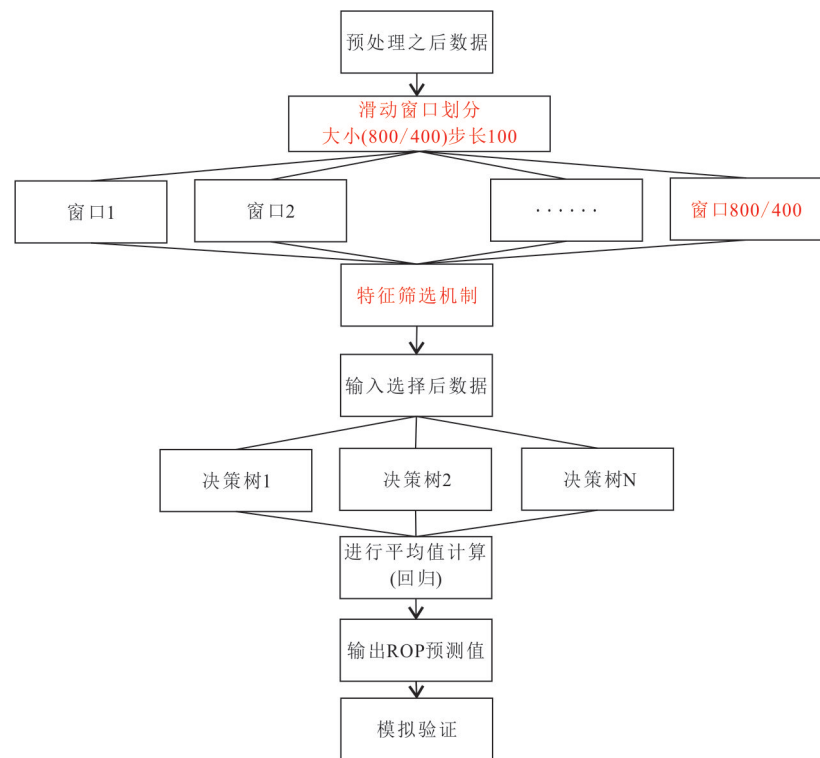


图9 滑动窗口模型架构

Fig. 9 Sliding window model architecture

5 算法验证

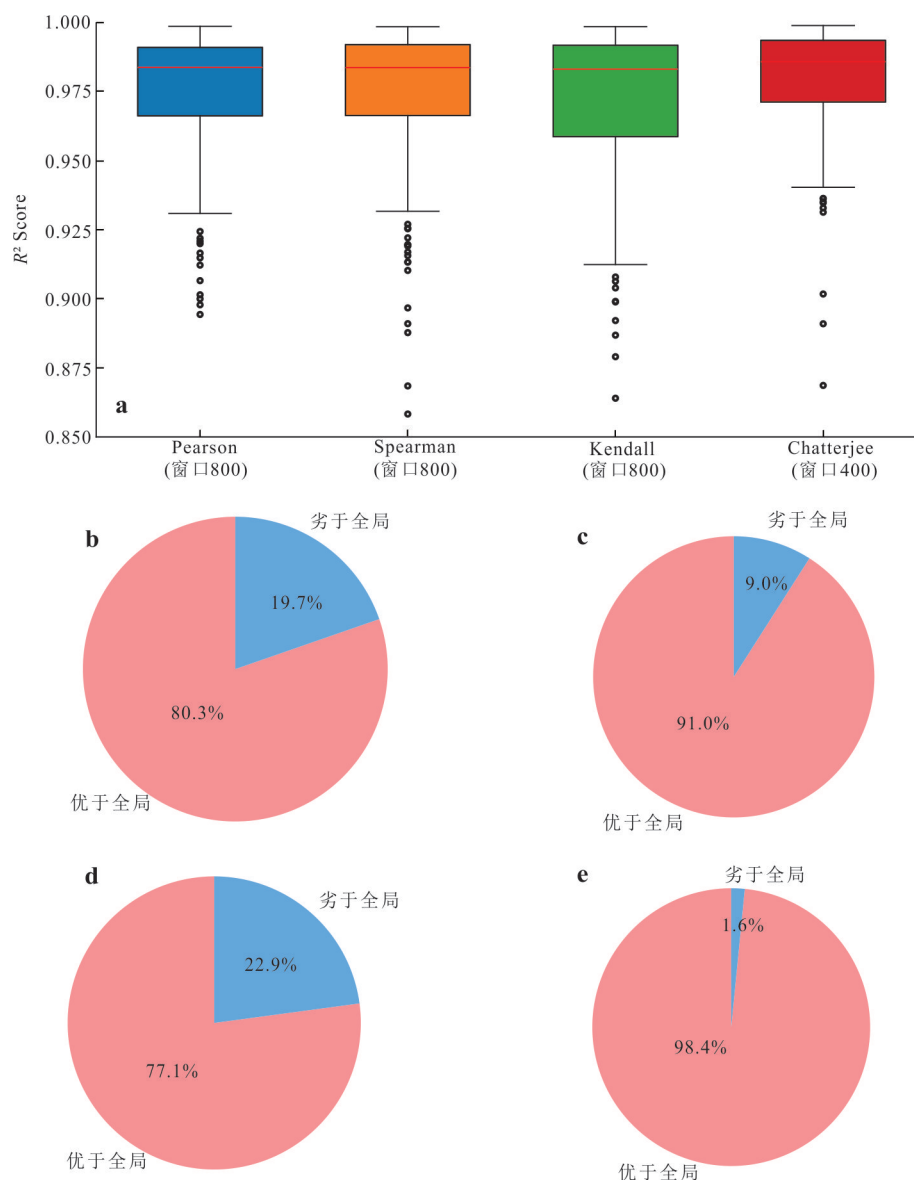
在利用九口油气井的数据进行分析之后,已经找出最优相关性钻速建模算法架构,就是在数据预处理结束之后,利用滑动窗口理论对数据集进行划分,滑动窗口的大小选为400,步长选为100,每个窗口中利用 Chatterjee 算法进行特征参数的筛选,选择系数值排名靠前的5个特征参数,利用这5个特征参数与钻速(ROP)的数据,建立随机森林模型,最后得到钻速(ROP)的预测值,同时得到模型的性能评分 R^2 . 选择对第10口油气井进行测试,十号井测试结果如图11所示:

综上分析可知,利用最优相关性钻速建模算法架构对第10口油气井进行分析,窗口训练集的平均 R^2 能达到0.992,窗口测试集的平均 R^2 能达到0.985,并且观察可知,利用滑动窗口进行动态分析,模型的窗口测试集的平均 R^2 要高于直接进行整体静态建模的测试集 R^2 (全局模型测试集 R^2 为0.981),这说明筛选出来的最优相关性钻速建模算法架构表现地十分优秀,能够比较精准地预测第十口油气井的钻速(ROP).

6 结论

本文针对复杂地层钻井过程中钻速(ROP)预测精度低、参数耦合关系复杂的难题,提出并验证了一套基于“动态最优相关性分析”的钻速建模算法框架,通过对10口实际油气井原始数据的深度挖掘与分析,得出以下主要结论:

(1)构建了完善的数据预处理与特征评价体系:综合运用中位数填充、IQR 离群值处理、Savitzky-Golay 滤波平滑及归一化技术,有效提升了数据质量. 通过对比 Pearson、Spearman、Kendall 及 Chatterjee 4种相关性理论,识别出钻井液密度、固相含量、井深及钻井液进出口比重等是影响钻速的核心共性因素,其工程作用机制如下:钻井液密度直接决定井底静液柱压力,密度过低易引发井壁失稳(坍塌、掉块),过高则导致压持效应增强(岩屑难以脱离井底),二者均会抑制钻速;固相含量升高会导致钻井液流变性能恶化(塑性粘度、屈服值上升),增加环空压耗,同时加剧钻头磨损和钻具泥包风险,从而显著降低机械钻速;井深增加伴随着地层压实程度升高、可钻性变差,同时环空压耗和井底压差增大,导致钻速整体呈下降趋势;钻井液进出口比重的动态差异则反映了井筒内岩屑承载与运移效

图 10 窗口测试集 R^2 运算结果Fig. 10 R^2 calculation results of the window test seta. 窗口测试集 R^2 ; b~e. 窗口优于全局比例, 依次为 Pearson/Spearman/Kendall/Chatterjee

率, 比重差过大表明井眼清洁不良, 易形成岩屑床, 增加钻柱摩阻并限制钻速提升;

(2) 明确了回归算法的适用性差异: 在全局建模环境下, 随机森林模型的预测性能普遍优于多层感知机. 随机森林通过集成多棵决策树, 能更好地处理钻井参数间的复杂非线性交互作用, 且对噪声具有更强的鲁棒性;

(3) 验证了滑动窗口理论在动态建模中的优越性: 不难发现, 钻速与参数的关联性具有明显的井段差异. 利用滑动窗口进行局部建模, 能捕捉到地层突变带来的非平稳特征, 其测试集平均 R^2 均达到 0.97 以上, 显著高于全局静态建模的效果;

(4) 确定了最优相关性钻速建模架构: 实验证明, 采用 Chatterjee 算法进行特征筛选 (Top5 特征), 结合滑动窗口 (窗口大小 400, 步长 100) 与随机森林模型的组合方案表现最为出色. 在独立第十口井的验证测试中, 该架构的窗口测试集平均 R^2 高达 0.985, 展现了极强的泛化能力与预测精度.

致谢: 本论文受到的资助项目为: 四川省自然科学基金项目青年基金, 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); 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:

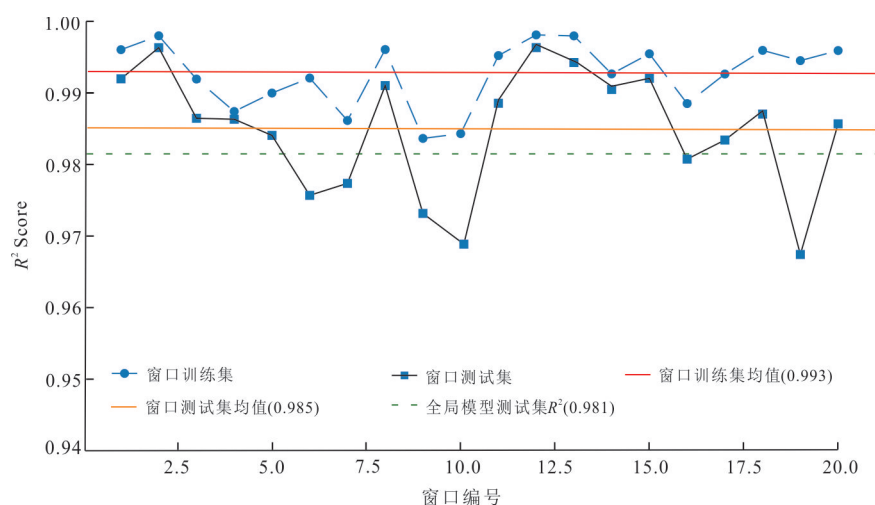


图 11 十号井测试结果

Fig. 11 Test results of Well No.10

CNOOC-KJ135ZDXM38ZJ05ZJ). 感谢成都理工大学朱海燕教授,中海油湛江分公司张超高级工程师提供原始数据.

References

- Bourgoyne, A. T. Jr, Young, F. S. Jr, 1974. A Multiple Regression Approach to Optimal Drilling and Abnormal Pressure Detection. *Society of Petroleum Engineers Journal*, 14(4): 371—384. <https://doi.org/10.2118/4238-pa>
- Cao, Z., 2023. Research and Application of Real-Time Prediction and Optimization Algorithm for Rate of Penetration in Drilling Proces(Dissertation). University of Electronic Science and Technology, Chengdu(in Chinese with English abstract).
- Chatterjee, S., 2021. A New Coefficient of Correlation. *Journal of the American Statistical Association*, 116(536): 2009—2022. <https://doi.org/10.1080/01621459.2020.1758115>
- Dong, L.J., 2024. Changes in Land-Sea Temperature Difference and Its Correlation with Extreme Precipitation in Eastern China Basedon MODIS Data(Dissertation). Zhejiang Ocean University, Zhejiang(in Chinese with English abstract).
- Ehsan, B., Ebrahim, D.B., 2022. Computational Prediction of the Drilling Rate of Penetration (ROP): a Comparison of Various Machine Learning Approaches and Traditional Models. *Journal of Petroleum Science and Engineering*, 210: 110033. <https://doi.org/10.1016/j.petrol.2021.110033>
- Gan, C., 2019. Intelligent Modeling of Formation Drillability Field and Drilling Rate of Penetration Optimization in Complex Conditions(Dissertation). China University of Geosciences, Wuhan(in Chinese with English abstract)..
- Gao, J.X., 2019. Prediction and Optimization of Offshore Drilling ROP Using Deep NeuralNetwork Models(Dissertation). Xi'an Shiyou University, Xi'an(in Chinese with English abstract).
- Hegde, C., Daigle, H., Millwater, H., et al., 2017. Analysis of Rate of Penetration (ROP) Prediction in Drilling Using Physics-Based and Data-Driven Models. *Journal of Petroleum Science and Engineering*, 159: 295—306. <https://doi.org/10.1016/j.petrol.2017.09.020>
- Hou, L. L., Wu, S., Yi, J. Z., et al., 2024. Discriminating Deposit Types Using Chlorite Trace Elements Based on Machine Learning. *Earth Science*, 49(12): 4303—4317(in Chinese with English abstract).
- Huang, X. L., Liu, D. T., Song, J. M., et al., 2021. Real-Time ROP Optimization Technology Based on Big Data and Artificial Intelligence. *Oil Drilling & Production Technology*, 43(4): 442—448(in Chinese with English abstract).
- Jiao, T., 2023. Real-Time Prediction of ROP Based on Deep Learning Method(Dissertation). Xi'an Shiyou University, Xi'an(in Chinese with English abstract).
- Li, B.Z., 2024. Prediction Method for ROP Based on Attention Mechanism of Convolutional Neural Network(Dissertation). Yangtze University, Wuhan(in Chinese with English abstract).
- Pang, L.Y., 2023. Correlation Measurement and Spatial Domain Division Method Basedon Spatial Transcriptome Data (Dissertation). Guangxi University, Nanning(in Chinese with English abstract).
- Peng, S.W., 2023. Correlation and Similarity Analysis of Stock Index Based on Data Mining Methods(Dissertation). Sichuan Normal University, Chengdu(in Chinese with English abstract).

abstract).

- Ren, H., 2023. Research on Multistep Prediction and Optimization of Drilling ROP Based on Cyclic Neural Network (Dissertation). Xi'an Shiyou University, Xi'an (in Chinese with English abstract).
- Song, J. Y., 2024. Research on Feature Selection Method Based on Correlation Analysis for Interval-Valued Data (Dissertation). Taiyuan Normal University, Taiyuan (in Chinese with English abstract).
- Wang, R. Y., 2023. Research on Prediction Method of Drilling Machine Penetration Rate Based on Machine Learning (Dissertation). Xi'an Shiyou University, Xi'an (in Chinese with English abstract).
- Yan, G. X., 2022. Research on Drilling Machinery Drilling Speed Prediction Model and Application Based on Optimization Neural Network (Dissertation). Southwest Petroleum University, Chengdu (in Chinese with English abstract).
- Yang, J. S., 2019. Research on Correlation Knowledge Discovery Method of Spacecraft Telemetry Data (Dissertation). University of Chinese Academy of Sciences (National Space Science Center, Chinese Academy of Sciences), Hefei (in Chinese with English abstract).
- Yin, X. Y., 2023. Modeling and Correlation Estimation of RNA-Seq Data (Dissertation). University of Science and Technology of China, Beijing (in Chinese with English abstract).
- Zeng, X. L., Li, Q., Wei, H. C., et al., 2023. Rate-of-Penetration (ROP) Prediction Model Based on Formation Characteristics of Extremely Thick Plastic Mudstone in South China Sea. *Coal Geology & Exploration*, 51(11): 159—168 (in Chinese with English abstract).
- Zhang, H. J., Zhang, G. F., Wang, G. N., et al., 2022. Classification and Prediction Method for ROP Based on Genetic Algorithm Optimization Random Forest Model. *Science Technology and Engineering*, 22(35): 15572—15578 (in Chinese with English abstract).
- Zhang, Y. H., 2017. The Study of Correlation Analysis and Dimensionality Reduction Methods and Their Applications (Dissertation). Beijing University of Technology, Beijing (in Chinese with English abstract).
- Zhou, C. C., Jiang, J., Li, Q., et al., 2022. Research on Drilling Rate Prediction Model Based on Fusion Feature Selection Algorithm. *Drilling Engineering*, 49(4): 31—40 (in Chinese with English abstract).
- Zhu, J. B., Liu, H. Y., Luan, S. C., et al., 2025. Prediction of On-Site Peak Ground Motion Based on Machine Learning and Transfer Learning. *Earth Science*, 50(5): 1842—1860 (in Chinese with English abstract).

中文参考文献

- 曹祺, 2023. 钻井过程钻速实时预测和优化算法研究与应用(博士学位论文). 成都: 电子科技大学.
- 曾小龙, 李谦, 魏宏超, 等, 2023. 基于南海巨厚塑性泥岩地层特征的钻速预测模型. *煤田地质与勘探*, 51(11): 159—168.
- 董丽洁, 2024. 基于MODIS数据的中国东部地区海陆温差变化及其与极端降水的相关性研究(博士学位论文). 浙江: 浙江海洋大学.
- 甘超, 2019. 复杂地层可钻性场智能建模与钻速优化(博士学位论文). 武汉: 中国地质大学.
- 高家轩, 2022. 基于深度神经网络模型的海上油田钻井提速方法研究(博士学位论文). 西安: 西安石油大学.
- 侯霖莉, 吴松, 易建洲, 等, 2024. 基于机器学习的绿泥石微量元素判别矿床类型. *地球科学*, 49(12): 4303—4317.
- 黄小龙, 刘东涛, 宋吉明, 等, 2021. 基于大数据及人工智能的钻速实时优化技术. *石油钻采工艺*, 43(4): 442—448.
- 焦通, 2023. 基于深度学习方法的钻速实时预测研究(博士学位论文). 西安: 西安石油大学.
- 李博志, 2024. 基于注意力机制的卷积神经网络机械钻速预测方法(博士学位论文). 武汉: 长江大学.
- 庞力源, 2023. 基于空间转录组数据的相关性度量和空间域划分方法(博士学位论文). 南宁: 广西大学.
- 彭绍炜, 2023. 基于数据挖掘方法对股指数据的相关性和相似性分析(博士学位论文). 成都: 成都四川师范大学.
- 任辉, 2023. 基于循环神经网络的钻井ROP多步预测与优化研究(博士学位论文). 西安: 西安石油大学.
- 宋金玉, 2024. 基于相关性分析的区间型数据特征选择方法研究(博士学位论文). 太原: 太原师范学院.
- 王若禹, 2023. 基于机器学习的钻井机械钻速预测方法研究(博士学位论文). 成都: 西安石油大学.
- 严顾鑫, 2022. 基于优化神经网络的钻井机械钻速预测模型及应用研究(博士学位论文). 成都: 西南石油大学.
- 杨甲森, 2019. 卫星遥测数据相关性知识发现方法研究(博士学位论文). 安徽: 中国科学院大学(中国科学院国家空间科学中心).
- 殷辛瑜, 2023. RNA-Seq数据的建模及相关性估计(博士学位论文). 北京: 中国科学技术大学.
- 张海军, 张高峰, 王国娜, 等, 2022. 基于遗传算法优化随机森林模型的机械钻速分类预测方法. *科学技术与工程*, 22(35): 15572—15578.
- 张亚红, 2017. 数据相关性分析和维数约简方法的研究与应用(博士学位论文). 北京: 北京工业大学.
- 周长春, 姜杰, 李谦, 等, 2022. 基于融合特征选择算法的钻速预测模型研究. *钻探工程*, 49(4): 31—40.
- 朱景宝, 刘赫奕, 栾世成, 等, 2025. 基于机器学习和迁移学习的现地地震动峰值预测. *地球科学*, 50(5): 1842—1860.