文章编号：2095-2163(2023)03-0210-05 中图分类号：TE24 文献标志码：A

基于遗传算法优化的随机森林钻井机械钻速预测模型研究

徐英卓[1]，王若禹[1]，王六鹏[2]
（1 西安石油大学 计算机学院，西安 710065；2 西安石油大学 石油工程学院，西安 710065）

摘 要：钻井机械钻速的提高是降低钻井周期，减少作业成本的重要措施。目前，采用传统的改进工具工艺手段来提高机械钻速不仅投资成本高，而且在应用效果上差异性大。针对这一难题，本文提出基于遗传算法优化的随机森林机械钻速预测模型。首先对数据进行处理、筛选，并为提高数据在算法模型的拟合程度，使用卡尔曼滤波对其进行降噪处理。然后，对输入特征参数进行相关性分析，筛选出最终适合的特征参数，降低模型冗余。最后通过实验验证，结果表明本文提出的遗传算法-随机森林机械钻速模型具有较高的精度，同时具有良好的收敛性.

关键词：机械钻速预测；卡尔曼滤波；随机森林算法；遗传算法

Research on ROP prediction model of random forest drilling machinery based on genetic algorithm optimization

XU Yingzhuo[1], WANG Ruoyu[1], WANG Liupeng[2]
(1 School of Computer Science, Xi′an Shiyou University, Xi′an 710065, China;
2 College of Petroleum Engineering, Xi′an Shiyou University, Xi′an 710065, China)

【Abstract】 The improvement of ROP of drilling machinery is an important measure to reduce drilling cycle and operation cost. At present, the traditional means of improving tool technology to improve the ROP has not only high investment cost, but also has great differences in application effects. To solve this problem, this paper proposes a random forest machinery ROP prediction model based on genetic algorithm optimization. First of all, the data is processed and screened, and in order to improve the fitting degree of the data in the algorithm model, Kalman filter is used for noise reduction. Then, after the correlation analysis of input characteristic parameters, the final suitable characteristic parameters are selected to reduce the redundancy of the model. Finally, the characteristic parameters are combined with the model through experiments. The results show that the genetic algorithm-random forest machinery ROP model proposed in this paper has high accuracy and good convergence.

【Key words】 prediction of ROP; Kalman filter; random forest algorithm; genetic algorithm

0 引 言

随着经济的快速发展，石油天然气等自然资源的消耗量也在不断增加。机械钻速(ROP)是影响钻井效率的关键因素之一，是石油工程钻井作业的重要经济指标。传统工艺技术实现“硬”提速，但由于各井之间地质条件不同导致提速效果差异大，从而陷入提速瓶颈。所以快速、准确地提高机械钻速，得到主要影响因素，进而优化钻井参数，该课题已成为钻井工程领域亟需解决的研究热点。

2007 年，范翔宇等学者[1]利用地震资料提出以数理统计方法对钻速进行预测，符合率达到 70%，然而由于地震资料的精度导致准确率难以进一步提升。2019 年，刘胜娃等学者[2]建立基于误差反向传播神经网络设计的机械钻速预测模型，但因为数据有限、特征较少导致对机械钻速影响规律未能进行有效探索。2021 年，许明泽等学者[3]研究多模型集成学习应用于机械钻速预测中，预测效果优于单一模型。但并未对单一模型进行调参，并不能解释集成模型优劣。

综上所述，目前学界对机械钻速影响因素的研究并不全面，导致机械钻速模型的精确度也不高。

基金项目：陕西省自然科学基础研究计划项目(2019JM-383)。

作者简介：徐英卓(1964-)，女，教授，主要研究方向：石油勘探开发领域应用；王若禹(1997-)，男，硕士研究生，主要研究方向：智能计算与可视化；王六鹏(1980-)，男，博士，副教授，主要研究方向：钻井信息化应用技术。

通讯作者：王若禹 Email：924709207@qq.com

收稿日期：2022-04-18

本文提出遗传算法-随机森林(GA-RandomForest)机械钻速预测模型，仿真实验结果表明所建预测模型具有更高精度。

1　GA-Random Forest 算法模型

(1)随机森林算法。该方法是一种通过集成学习思想将多个决策树集成在一起的算法。随机地从数据集中抽取数据用作决策树[4]的训练集，并随机地从特征数据中选取特征节点建立决策树，重复操作后形成森林。在此基础上，对所有树得出的值进行选择，被选择最多的即是最终的输出结果。

(2)遗传算法。该方法是解决复杂优化问题最常用的方法[5]。遗传算法模拟生物遗传进化的过程。首先，初始化总体，每个染色体代表一个解决方案。其次，适应度函数决定了种群进化的方向，适应度函数的值决定了解的质量。适应度函数定义为：

$$F = 1 - \frac{1}{M}\sum_{k=1}^{M}\left[P(x_k) - y_k\right]^2 \quad k = 1,\cdots,M \tag{1}$$

然后，按照适者生存的自然选择原则，优秀的个体更有可能保留自己的基因，因此具有高适应值的个体更有可能被选为下一代的父母。本研究用轮盘赌法进行选择操作，使个体被选择概率与其适应度值成正比，个体 α 被选择的概率 p^{α} 可表示为：

$$p^{\alpha} = \frac{F^{\alpha}}{\sum\limits_{\alpha' F^{\alpha}}} \tag{2}$$

其中，F^{α} 为个体 α 的适应度值，$F^{\alpha'}$ 为个体 α' 的适应度值。

最后，通过交叉和变异生成下一代种群，当得到满意解或达到定义代数时，则结束进化过程。

(3) GA-Random Forest 算法[6]。GA-Random Forest 机械钻速预测模型的建模过程如图 1 所示。由图 1 可看到，首先，将随机森林中的每一个决策树作为染色体对其进行编码，规定决策树的数量就是染色体的长度。然后，设置条件函数来计算该树的准确率，用来评价决策树组合的优缺点。每个决策树组合的分类正确率作为对应染色体的适应度。其次，用轮盘赌法进行选择操作，规定其中每一代优秀率高的组合具有更高的被选择遗传下来的概率。最后，通过交叉产生子代，变异可为决策树的组合提高随机性，从而避免陷入局部最优。通过上述步骤，得到了更加优秀的个体，如此即可以加快进化速度。

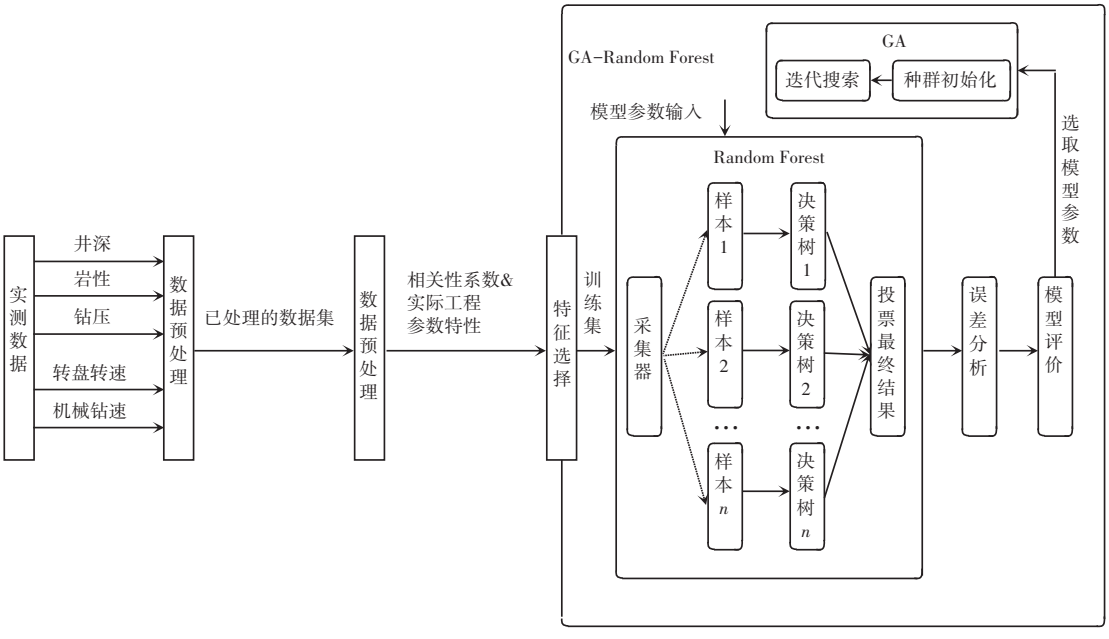


图 1　模型框架结构图

Fig. 1　Model frame structure

2　基于 GA-Random Forest 的机械钻速预测建模

2.1　机械钻速影响因素分析

本次实例数据选用某油田特定区块下的井史数据。表 1 列举了部分数据。文中对此进行初步筛选后，拟以表 1 中的特征参数作为影响因素。

2.2　输入参数预处理

2.2.1　CatBoost 对类别变量的处理

CatBoost 编码器可以避免均值编码对 y 变量敏感的弊端，并减少过拟合且不改变数据集的大小。

其基本思想也是计算某一行数据的特征编码时，避免使用到该行的目标值(Target)。首先，将相同类别的元素分组，求出每一组 *target* 的平均值作为其对应的编码。然后，引入“前缀和”的思想，即对于某一类别的某一个值，其对应的编码值等于其之前行的所有该类别值的对应 *target* 的平均值。前缀和定义如下：

$$S_k = \sum_{i=1}^{k} target_i \tag{3}$$

本文中，岩性作为有 11 种类别的变量，将采用 CatBoost编码器对类别特征无序且对类别数量较多的目标变量编码方式进行处理。编码结果见表 2。

表 1 机械钻速预测模型输入数据表

Tab. 1 Partial data of ROP prediction

井深/m	岩性	钻压/kN	转盘转速/rpm	排量 *L*/min	泵压/MPa	钻井液密度/(g·cm^{-3})	入口流量/(L·min^{-1})	立压/MPa	扭矩/kNm	机械钻速/(m·h^{-1})
4526	玄武岩	134.1	61	35.3	15.2	1.24	35.33	15.2	4.1	2.78
4616	凝灰岩	76.9	61	36.5	17.8	1.24	36.50	17.8	13.7	6.98
4932	英安岩	115.5	50	26.7	10.8	1.24	26.74	10.8	11.9	4.48
4957	泥岩	74.8	50	29.3	12.1	1.23	29.29	12.1	8.7	5.94
…	…	…	…	…	…	…	…	…	…	…
5998	泥岩	46.8	50	25.4	23.7	1.38	25.80	22.4	13.4	5.31

表 2 类别变量编码结果表

Tab. 2 Category variable coding results

岩性	细砂岩	凝灰岩	粉砂岩	泥岩	泥灰岩	砂岩	灰岩	英安岩	绿岩	白云岩	玄武岩
编码	6.36	6.32	5.45	4.22	3.63	3.66	3.52	3.16	2.63	2.06	1.57

2.2.2 卡尔曼滤波数据降噪处理

卡尔曼滤波是一种借助线性算法的方程，通过系统输入输出观测数据，对系统状态进行最优估计的算法。

卡尔曼滤波分为 2 个步骤。第一步，基于上一时刻状态数据预测当前时刻状态。第二步，是综合第一步预测出的当前时刻状态和实际观测状态，估计出最优的状态作为滤波的结果。对此数学方法，可用如下公式进行描述：

$$\boldsymbol{x}_k = \boldsymbol{A}\boldsymbol{x}_{k-1} + \boldsymbol{B}\boldsymbol{u}_{k-1} \tag{4}$$

$$\boldsymbol{P}_k = \boldsymbol{A}\boldsymbol{P}_{k-1}\boldsymbol{A}^{\mathrm{T}} + \boldsymbol{Q} \tag{5}$$

$$\boldsymbol{K}_k = \boldsymbol{P}_k\boldsymbol{H}^{\mathrm{T}}(\boldsymbol{H}\boldsymbol{P}_k\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{R})^{-1} \tag{6}$$

$$\boldsymbol{x}_k = \boldsymbol{x}_k + \boldsymbol{K}_k(\boldsymbol{z}_k - \boldsymbol{H}\boldsymbol{x}_k) \tag{7}$$

$$\boldsymbol{P}_k = (\boldsymbol{I} - \boldsymbol{K}_k\boldsymbol{H})\boldsymbol{P}_k \tag{8}$$

这里，式(4)是状态预测；式(5)是误差矩阵预测；式(6)是卡尔曼增益计算；式(7)是状态校正，运算输出的就是最终的卡尔曼滤波结果；式(8)是误差矩阵更新。

卡尔曼滤波对其中机械钻速数据的降噪前后对比如图 2 所示。分析图 2 可知，经过卡尔曼滤波处理，本来包含许多尖峰和突变的原始数据相较于之前变得轮廓更加清晰，峰值不再尖锐。所以卡尔曼滤波有效去除了原始数据中明显的信号干扰，在处理过后并未改变原数据的变化特性。

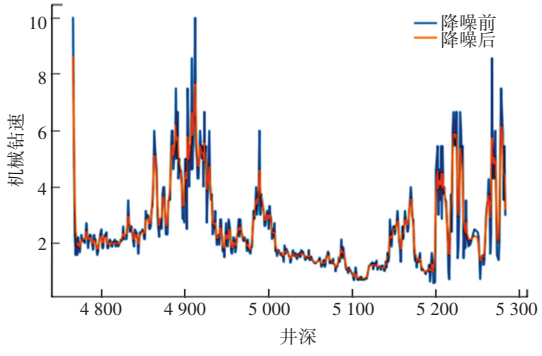


图 2 卡尔曼滤波降噪处理对比图

Fig. 2 Comparison of Kalman filter denoising

2.3 特征选择

在工程实践中获得的钻井数据类别繁多，将收集到的所有特征参数输入机器学习模型进行训练，会导致模型维度过多，也就无法有效提升拟合程度。为此，利用最大互信息系数 (*MIC*)，最大程度地根据信息寻找参数之间线性或者非线性的关系。

最大互信息系数计算公式如下：

$$mic(x;y) = \max_{a*b<B} \frac{I(x;y)}{\log_2 \min(a,b)} \tag{9}$$

其中，a、b 分别表示在 x，y 方向上的区域分割个数；B 表示可设置参数；$I(x;y)$ 表示 *MIC* 值。式(9)为在不同规定范围下得到各自的 *MIC* 值，并在归一化处理后来求得最大值。

钻井特征参数最大互信息相关分析图如图 3 所

示。由图 3 可见，立压与泵压、相关性极强（0.98），排量和入口流量、相关性极强（0.98）。因此，通过 *MIC* 计算值与实际工程原理结合筛选井深、岩性、钻压、转盘转速、钻井液密度、入口流量、立压、扭矩等 8 项参数筛选作为机械钻速预测模型的输入变量。

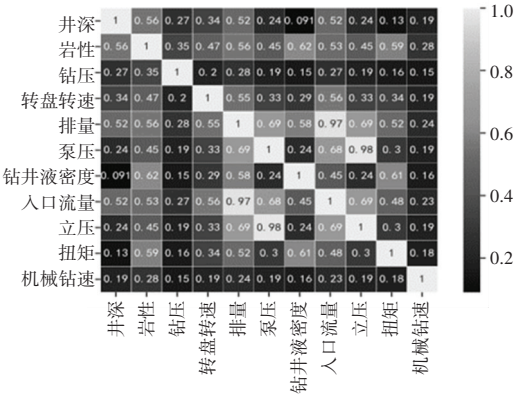


图 3　钻井特征参数最大互信息相关分析图

Fig. 3　MIC of drilling characteristic parameters

2.4　机械钻速预测模型的建立与实验验证

这里，研发建立了 GA-Random Forest 机械钻速预测模型。随机森林模型中涉及到的 2 个主要参数是树的深度和决策树的数量，所以利用遗传算法对其进行优化。首先，根据经验设定树的深度和决策树的数量，并在遗传算法中设定繁殖的代数为 100，种群的数量为 500，同时设定交配的概率为 0.6，变异概率为 0.01。当代数达到设定的 100 代时算法停止，给出最优的一代和其中解码后的参数。研究中得到的繁殖迭代过程参数见表 3。

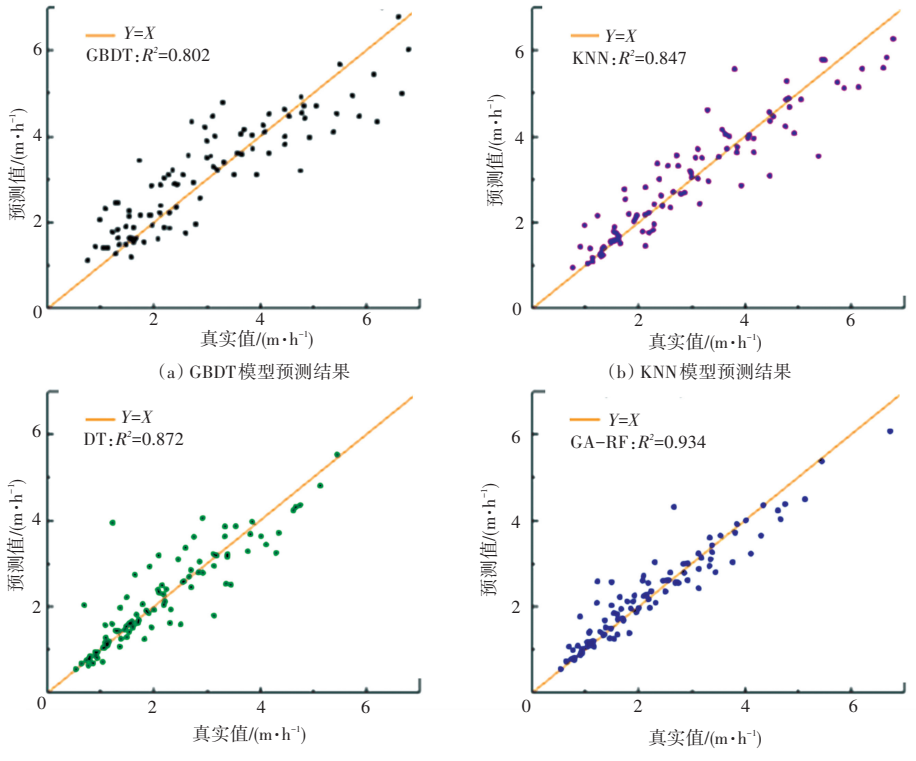
表 3　每一代繁殖参数表

Tab. 3　Parameters value of each generation

繁殖代数	*n_estimators*	*max_depth*	*r2_score*
0	120	14	0.864 3
1	110	16	0.921 4
…	…	…	…
99	140	8	0.926 0
100	120	6	0.904 2

最终，确定最优代为第 76 代，*n_estimators* 为 120，*max_depth* 为 16，R^2*_score* 为0.937 4。

为了证明 GA-Random Forest 机械钻速预测模型在本次实验中与其他模型相比具有更高精度，故选取决策树回归模型、KNN 回归模型、SVR 回归模型进行对比分析，实验结果如图 4 所示。



（a）GBDT模型预测结果　（b）KNN模型预测结果

（c）DT模型预测结果　（d）GA-RF模型预测结果

图 4　多模型预测结果对比图

Fig. 4　Prediction results of multiple models

为比较模型的优劣,用拟合优度 R^2 作为区别的标准。R^2 越大,模型的解释程度越高,预测点在回归直线附近越密集。由图 4 可见,GA - Random Forest 模型的预测值与实测数据曲线变化一致、对应数值点相近,并且该模型的 R^2 值优于其他 3 种算法模型。因而可知,本文研究的机械钻速预测模型精度更高。

3 结束语

(1)使用 CatBoost encoder 得到更直接表示分类变量和目标变量之间的关系的目标编码,并且有效降低模型过拟合。

(2)去除多余的干扰获得真实有用的数据,使用卡尔曼滤波降噪处理后达到信噪分离的效果,进一步提高算法模型的拟合程度。

(3)本次研究提出的方法在随机森林的基础上又提高了计算准确度和适应能力,并通过简化模型的结构,有效提高了计算速度。

参考文献

[1] 范翔宇,夏宏泉,郑雷清,等. 利用地震层速度预测地层可钻性和钻速的新方法[J]. 钻采工艺,2007(01):4-6,143.
[2] 刘胜娃,孙俊明,高翔,等. 基于人工神经网络的钻井机械钻速预测模型的分析与建立[J]. 计算机科学,2019,46(S1):605-608.
[3] 许明泽,韦明辉,邓霜,等. 多模型集成学习在机械钻速预测中的新应用[J]. 计算机科学,2021,48(S1):619-622,657.
[4] 胡金涛. 基于 C4.5 决策树的学生成绩预测教学系统的研究与实现[D]. 成都:西南交通大学,2017.
[5] LOUISEB, MARCO P, PATRICK A, et al. Unraveling the motivational secrets of honey bee foraging during the COVID pandemic[J]. Science, 2022, 25(4):104116.
[6] VALINGER D, LONGIN L, GRBEŠ F, et al. Detection of honey adulteration—The potential of UV - VIS and NIR spectroscopy coupled with multivariate analysis[J]. LWT, 2021, 145:111316.

(上接第 209 页)

图 7 报警信息图片

Fig. 7 Alarm information picture

3 结束语

当今,老年公寓的数量也在逐渐增加。只是,老年公寓现如今大多仍是采取人工管理模式,管理起来相对繁杂。而对老年公寓的管理质量将直接影着老年公寓中老人的健康以及心理,因此亟需建立更加舒适的管理模式。本文设计研发系统经过仿真得知,该系统集多功能于一体,具有一定的创新性,更加有利于掌握老人的心理及健康状态,同时也为护工的管理提供了便利条件。现已表明,本文研发成果具有较强的市场前景以及可行性。

参考文献

[1] 王茹. 互联网+居家养老服务.养老服务模式的创新[D]. 长春:吉林大学,2017.
[2] 温海红,王怡欢. “互联网+居家养老”服务平台构建及其实现路径[J]. 河北大学学报(哲学社会科学版),2017,42(06):138-146.
[3] 同春芬,汪连杰. “互联网+”时代居家养老服务的转型难点及优化路径[J]. 广西社会科学,2016(02): 160-166.