

基于因子分析的岩性识别智能模型对比

何俊杰¹, 李 谦^{1*}, 李同意¹, 魏思维¹, 姜 杰²

(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: 岩性识别是钻探中十分重要的环节, 传统岩性识别方法耗时较长且准确率较低, 基于机器学习的智能岩性识别可以有效加快岩性识别的效率, 准确率更高, 泛化能力更强。本文以实现基于钻进参数的智能岩性识别预测为目标, 使用逻辑回归、支持向量机、K近邻算法、随机森林、神经网络5种算法, 先对原始数据进行预处理, 然后对地层参数和其他参数进行因子分析、降维处理, 建立了以钻进参数矩阵为输入, 岩性分类识别和地层参数预测为输出的岩性智能识别模型, 实现了钻遇地层的智能预测。实验表明, 5种算法下岩性识别在测试集上的准确率和F1分数大部分达到70%以上, 个别算法准确率较低但也在60%左右, 5个模型地层参数预测的指标MAE和RMSE大部分都在1之下, 识别预测准确。

关键词: 岩性识别; 因子分析; 机器学习; 人工智能建模; 逻辑回归; SVM支持向量机; K近邻算法; 随机森林; 神经网络

中图分类号: P634; TE242 文献标识码: A 文章编号: 2096-9686(2025)S1-0105-07

Comparative study of intelligent lithology identification models based on factor analysis

HE Junjie¹, LI Qian^{1*}, LI Tongyi¹, WEI Siwei¹, JIANG Jie²

(1. College of Environment and 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)

Abstract: Lithology identification is a very critical process in geological exploration and development. With the development of the geological exploration industry, the traditional lithology identification is limited by time and space, and can no longer meet the increasing data scale and dimensionality. Therefore, there is an urgent need for in-depth research on the intelligent prediction of lithology identification, so as to promote the development of lithology identification in the direction of digitalization, intelligence and timeliness. In this paper, with the goal of realizing intelligent lithology identification and prediction based on drilling parameters, five algorithms are used to establish intelligent lithology identification and prediction models, including Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbor (KNN) algorithm, Random Forest (RF), and Neural Network (NN). They are designed to realize the intelligent prediction of drilling encounters. The main research contents are as follows: A regional stratigraphic identification model based on drilling data was designed. Through the data preprocessing of the original data, and then the factor analysis of the formation parameters and other parameters were reduced by factor analysis and modeled by five machine algorithms. An intelligent identification model was established with the drilling parameter matrix as the input, the lithology classification identification and formation parameter prediction as the

收稿日期: 2025-06-15; 修回日期: 2025-07-16 DOI: 10.12143/j.ztgc.2025.S1.016

基金项目: 四川省自然科学基金青年科学基金项目“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: 何俊杰, 男, 汉族, 2002年生, 硕士研究生, 地质资源与地质工程专业, 研究方向为数据分析与人工智能在钻探中的应用, 四川省成都市成华区二仙桥东三路1号, 1281822173@qq.com。

通信作者: 李谦, 男, 汉族, 1987年生, 副教授, 硕士生导师, 地质资源与地质工程专业, 博士, 从事钻井智能化分析与应用研究工作, 四川省成都市成华区二仙桥东三路1号, liqian2014@cdut.edu.cn。

引用格式: 何俊杰, 李谦, 李同意, 等. 基于因子分析的岩性识别智能模型对比[J]. 钻探工程, 2025, 52(S1): 105-111.

HE Junjie, LI Qian, LI Tongyi, et al. Comparative study of intelligent lithology identification models based on factor analysis[J]. Drilling Engineering, 2025, 52(S1): 105-111.

output. Experiments show that for most of the five algorithms, the lithology identification accuracy and F1-score on the test set exceed 70%, and the individual accuracy is low but also about 60%, and most of the MAE and RMSE predicted by the stratigraphic parameters of the five models are below 1, and the identification and prediction are accurate.

Key words: lithology identification; factor analysis; machine learning; artificial intelligence modeling; logistic regression; support vector machine; K-nearest neighbor algorithm; random forest; neural network

0 引言

岩性识别是钻探中十分重要的过程,在地质资源预测和评价过程中发挥了十分关键的作用。随着地质勘探行业的发展,传统的岩性识别耗时较长且准确率较低,急需对岩性识别的智能预测进行深入研究,以推动岩性识别朝着数字化、智能化、时效化的方向发展。基于机器学习的智能岩性识别技术可以加快岩性识别的效率,且准确率更高,泛化能力更强,可有效提高钻进效率,降低钻探成本。

国内外学者开展了大量的岩性识别研究,刘秀娟等^[1]提出用测井数据进行岩性识别的多元统计方法。牟丹等^[2]采用3种经典机器学习算法对火山岩进行岩性识别。苏赋等^[3]改进了支持向量机方法,使其在处理不平衡数据集时能更准确地识别岩性。王淑盛等^[4]提出的加权K近邻法,与普通K近邻算法相比,准确率由51.46%提升至79.14%。于代国等^[5]提出了用支持向量机算法进行岩性识别,把岩性识别率由70%提升到95%。徐晗等^[6]运用非开挖随钻检测系统和随机森林算法的地层识别,利用了PCA主成分分析将地层特征降至三维,识别准确率达到92%。庞志超等^[7]分析不同岩性的测井曲线响应特征,提出了蒙特卡罗数值模拟方法,建立了补偿中子孔隙度曲线校正模型,最终实现了岩性的精细划分。王光宇等^[8]针对不平衡样本集,用录井岩性数据作为岩性样本标签,用岩性弹性参数作为样本特征搭建岩性样本集,然后结合近邻清除算法和合成少数类过采样算法形成NM-SMOTE算法,利用随机森林分类器建立岩性识别模型。罗仁泽等^[9]提出KSMOSEL集成学习模型,使岩性识别准确率提升至94.28%。苏自武等^[10]提出利用人工神经网络原理对钻速进行预测。王振豪^[11]采用MSCNN-GRU神经网络重构测井曲线,使用Optuna-BiGRU-EMHA神经网络进行岩性识别。尹琼^[12]利用两步法的岩性识别策略和生成对抗神经网络进行不平衡样本的岩性识别。张东晓等^[13]提出用循环神经网络(RNN)即长短期记忆神经网络

(LSTM)来建立测井曲线。高铁红^[14]提出基于遗传算法和神经网络的岩性识别方法。王亚飞等^[15]提出了一种融合模型预测岩石类别。模型融合算法的初级学习器为支持向量机、人工神经网络和随机森林,通过次级学习器贝叶斯模型平均算法对每个模型的权重进行计算,融合算法准确率达到96.86%。岳永东等^[16]基于岩心资料和测井数据建立了支持向量机岩性识别模型,该模型识别准确率高、训练样本需求量低,在第四系松散沉积层调查工作中,利用测井数据实现岩性自动识别具有可行性,为建立松散沉积物地层序列提供了有力支撑。Senjoba等^[17]提出基于深度学习,利用钻探振动信号的频谱图像实现实时自动的岩石类型识别。Yaotu等^[18]使用4种人工智能模型(ANN、SVR、RNN、LSTM)预测岩屑床高度。

本文建立了基于钻井数据的区域性岩性分类识别模型。通过对原始钻井数据进行数据预处理,然后对地层因素和其他因素进行因子分析降低参数维数,把多个参数融合为7个主因子,用7个主因子进行模型训练,最终建立了以钻进参数矩阵为输入,岩性分类识别为输出的岩性智能识别模型。本文采用因子分析来辅助建立岩性识别模型,消除了冗余信息,可以有效抑制噪声,提升数据稳定性,提高了算法识别准确率。

本文的技术路线如图1所示,首先对钻井数据进行预处理,然后利用因子分析对处理后的钻井数据进行数据降维,最后通过逻辑回归(Logistic Regression, LR)、支持向量机(Support Vector Machine, SVM)、K近邻算法(K-Nearest Neighbor, KNN)、随机森林(Random Forest, RF)、神经网络(Neural Network, NN)5种机器算法建立区域性地层识别模型。

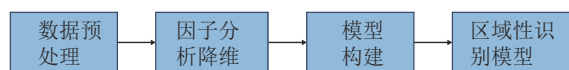


图1 技术路线流程

1 数据采集与预处理

1.1 数据集描述

该区域的岩性有26种,主要的岩性有3种,以灰色泥岩为主,占比65%,其次为灰色粉砂质泥岩,占比16.7%,灰色泥质粉砂岩为9.5%。其余岩性为次要岩性,有部分岩性出现次数太少,如浅灰色泥质细砂岩、灰色泥质灰岩和灰色泥灰岩,都只出现过一次。由此可见,该区域以泥岩为主导地层,砂岩为次要主导地层。在钻进泥岩及粉砂岩等比较软的岩石时,钻速较高,其中钻进粉砂质泥岩时最高,达35 m/h,钻进灰色泥岩和灰色粉砂质泥岩时钻速达到30 m/h,且钻速变化范围比较大,从10 m/h到30 m/h。而钻进砂岩及中砂岩等更硬岩石时钻速显著降低,尤其是钻进灰色中砂岩时,钻速只有5 m/h。钻进大部分砂岩的钻速都普遍较低,保持在10 m/h以下,如灰色泥质细砂岩、灰白色灰质细砂岩等。所以岩性对于钻速参数影响较大,建模时考虑重点分析钻速和岩性的相互关系。不同岩性下钻速分布如图2所示。

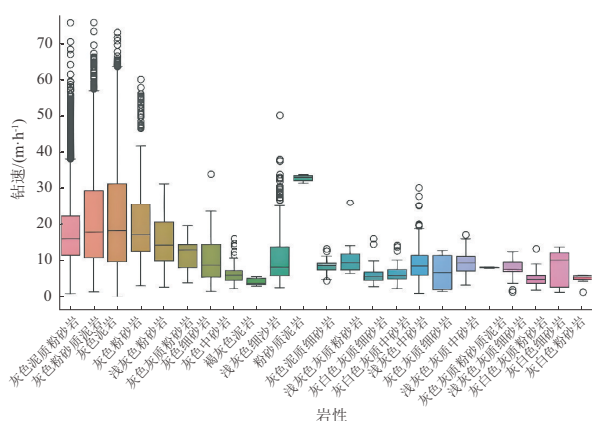


图2 不同岩性下钻速分布

主要钻进参数:机械钻速(ROP)变化范围0~70 m/h,正常范围18~28 m/h,存在高ROP异常值集群(60~70 m/h);钻压(WOB)集中分布在6~12 t,中位数10 t;转速(RPM)主要区间为100~120 r/min;扭矩(Torque)主要分布20~30 kN·m;入口泥浆密度为1.0~2.2 g/cm³,中位数1.8;出口泥浆密度为1.0~2.2 g/cm³,中位数为1.7略低于入口泥浆密度;泥浆黏度中位数为40 s,有部分异常值;pH范围为8.0~11.0;井径中位数为311.15 mm;钻具悬量中位数145 t;泥浆泵流量主分布范围为3000~4000

L/min;屈服值主分布为7.5~10.0 Pa;塑性黏度主分布为10~20 mPa·s;3转读数为3~15;10 s静切力为0~2 Pa;膨润土含量为5~8 kg/m³;滤失量中位值7 mL/30 min。

1.2 数据预处理

本文主要研究地层参数和钻进参数之间的相互关系,所以可以排除一些非相关的参数类型,如钻头参数等,留下的参数为地层参数和钻进参数两大类。采用箱型图的方法检测异常值并用IQR(四分位距)缩尾法处理离群值;采用StandardScaler对所有数据进行全局标准化,消除不同特征在量纲上的差异;采用滚动窗口均值平滑,消除测量噪声。具体流程和部分处理效果如图3、图4所示。



图3 数据处理流程

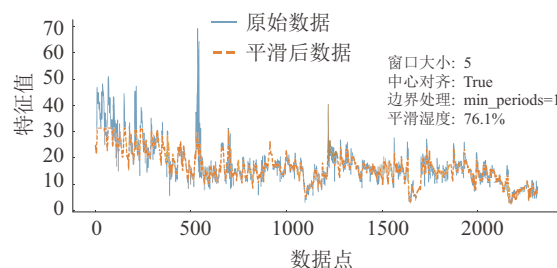


图4 平滑前后钻速曲线

1.3 数据降维

本文研究的参数包括5个地层参数和22个钻进参数,为方便后续建模,需要对这些参数进行相关性分析,本文通过计算皮尔逊相关系数来评价各参数间的相关性程度,地层参数和钻进参数的相关性如图5、图6所示。

由图5、图6可知,地质参数中,三大压力相关系数数值都大于0.94,属于强相关性,三者呈现高度协同变化,地震速度与三个压力参数的相关系数在0.5~0.6,而岩性和其他参数呈现弱相关性;钻进参数中也存在很多强相关性参数,并且还很多中强相关性参数,相关系数大于0.6。由于很多参数之间都存在强相关性,需要对这些强相关参数做降维处理。

因为本文数据中存在较多强相关参数,为了降低参数维数,选择因子分析(Factor Analysis)来进行参数降维。对地层参数进行因子分析的结果如图

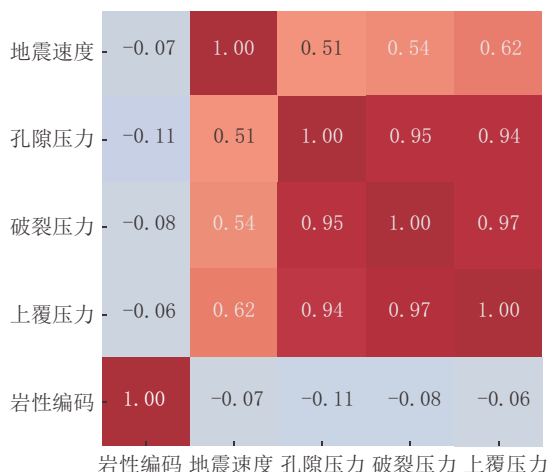


图5 地质参数皮尔逊相关性

7、图8所示,第一主因子与孔隙压力、破裂压力以及上覆压力呈正相关(载荷值0.89~0.94)。第二主因子与地震速度强相关,载荷值为0.96。第一主因子方差解释率53.7%,第二主因子以地震速度和岩性为主导,方差解释率为27.3%。两个因子的累计方差解释率为81%。

对地层参数以外的其他钻进参数进行因子分析以降低参数维数。经因子分析后把钻进参数分为5个因子。泥浆密度、入口泥浆密度、出口泥浆密度、孔径、pH以及固相含量是第一主因子的主要组成部分,载荷值都大于0.9。第二主因子主要以3转

读数、6转读数、10 s静切力和10 min静切力为组成部分。第三主因子主要以钻速、悬重、泥饼厚度以及膨润土含量为主要成分。第四主因子以泥浆黏度和塑性黏度为主要成分。第五主因子主要以钻压和扭矩为主要成分。各因子的载荷矩阵和方差解释率如图9、图10所示。

第一主因子方差解释率为56.2%,表明第一主因子(泥浆参数)的重要性占钻进参数的一半以上,五个主因子的累计方差解释率达到了82.6%,表明因子分析后的五个因子占据所有钻进参数的大部分,因子分析使用合理。

2 区域性地层识别模型构建

2.1 模型架构综述

本文以地层、钻进参数为输入,用包括逻辑回归、支持向量机、K近邻算法、MLP神经网络以及随机森林这5种算法对模型进行训练,训练完成后以测试井的因子分析降维后的钻进参数为输入,以岩性预测和因子分析降维后的地层参数为输出。其模型结构如图11所示。

本文采用4个评价指标来评估模型的好坏,用准确率(Accuracy)、F1分数(F1-Score)评价岩性预测的效果,用MAE和RMSE评价地层参数预测的准确性。MAE和RMSE的值需要结合上文因子分

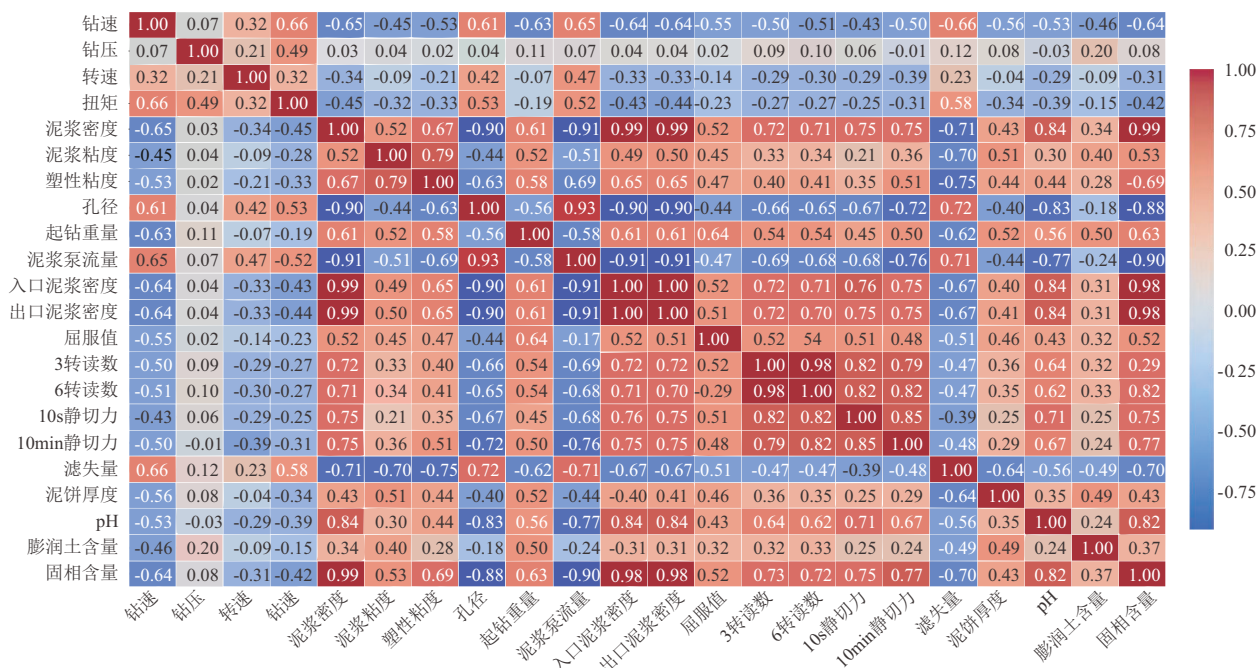


图6 钻进参数相关性

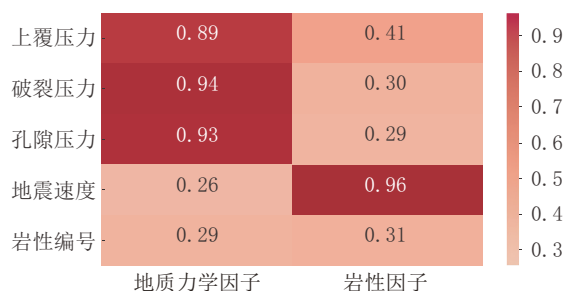


图7 地质因子载荷矩阵

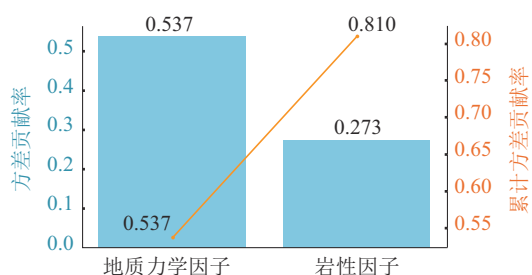


图8 地层因子方差解释率

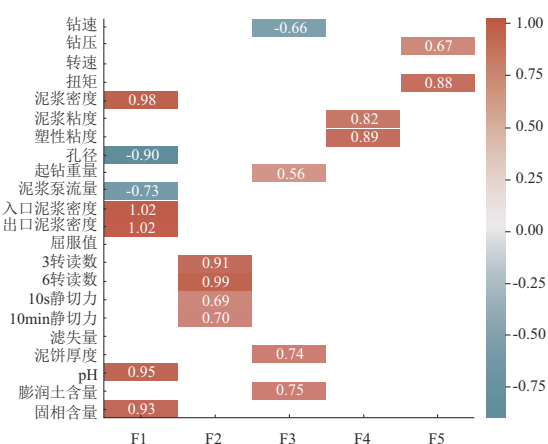


图9 钻进参数载荷矩阵

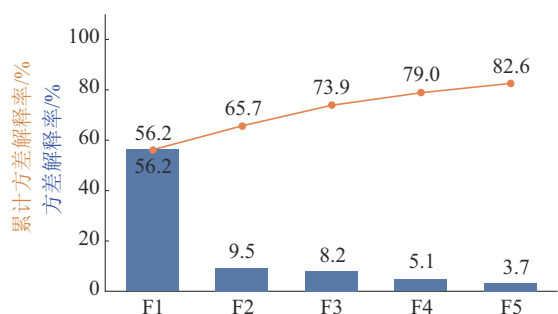


图10 各因子方差及累计方差解释率

析时讨论过的地层参数方差解释率一起评价,上文地层参数因子和钻进参数因子的方差解释率分别为81%和82.6%,所以下文的MAE和RMSE只要

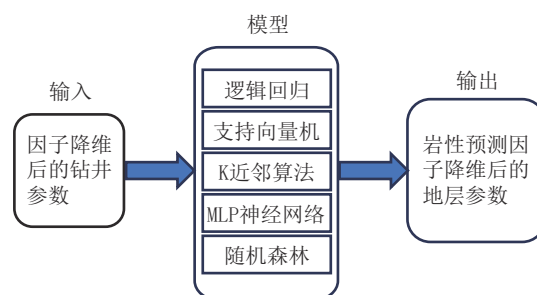


图11 基于地层钻进参数的岩性识别预测结构

在1左右都较为合理,越小模型越准确。

2.2 模型训练与测试

本文的数据集中一共有10口井,依次采用1~9口井的数据对模型进行训练,每次训练后剩下的井都用作该次模型训练后的测试井,并最终给出模型的平均性能表现。

岩性预测方面:不同模型在训练井和测试井中表现具有差异。SVM和神经网络在多数场景都表现较好:当训练井数为1、测试井为10时,SVM和神经网络准确率分别达到84.47%和82.78%,F1分数分别为83.18%和81.91%。在训练井数为3、测试井为10时,神经网络准确率高达82.80%,F1分数80.78%。随机森林表现中等,但稳定性较好:在训练井数为4、测试井为10时,准确率74.86%,F1分数76.43%。逻辑回归和KNN表现较差:逻辑回归在训练井数为1、测试井为6时,准确率仅18.47%,F1分数9.46%;KNN在训练井数为1、测试井为8时,F1分数52.59%。如图12所示。

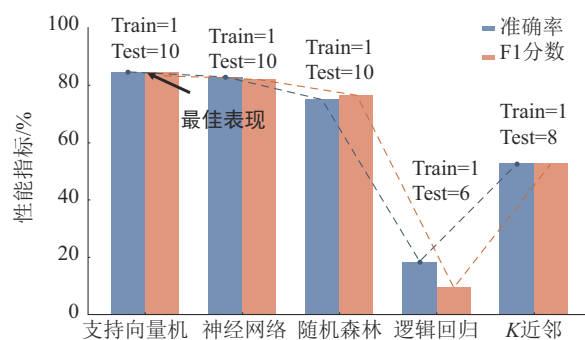


图12 部分测试效果

地层参数预测方面:随机森林和K近邻算法的MAE和RMSE较低(常低于1.0),表现很稳定,神经网络算法的回归性能波动大,支持向量机算法在部分场景下表现优异。

区域性识别模型里的逻辑回归算法在训练井数量为1时的模型表现较差,但是逻辑回归算法的表现随着训练井数量上升而增长最快,并在训练井为9时达到所有算法的最高值,同时逻辑回归算法的F1分数变化趋势也和准确率相似。除此之外,支持向量机算法是五个算法中最稳定的,一直保持高准确率和F1分数。不同数量训练井情况下的平均准确率和平均F1分数如图13所示。

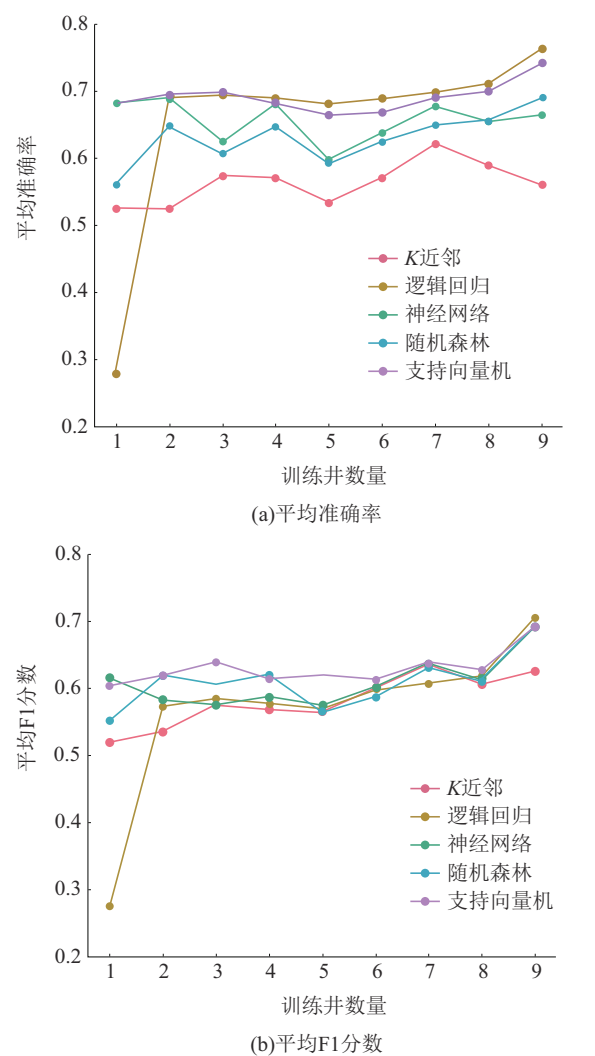


图13 不同训练井下的模型表现

图13展示的各算法的平均表现在训练井数量达到5以后都趋于稳定,且训练井越多,算法越稳定,所以当训练井数量为8和9时模型的综合性能最好,具体表现如表1所示。

分析表1可以更加清晰地发现,在训练井数足够多的情况下,5类模型的准确率相对稳定,尤其是

表1 各模型表现

训练井数	训练井	测试井	模型	准确率	F1分数
8	1,2,3,4,5,6,7,8	9	RF	0.629286	0.545248
8	1,2,3,4,5,6,7,8	9	SVM	0.666821	0.579463
8	1,2,3,4,5,6,7,8	9	NN	0.632994	0.535358
8	1,2,3,4,5,6,7,8	9	LR	0.659407	0.531449
8	1,2,3,4,5,6,7,8	9	KNN	0.610751	0.58105
8	1,2,3,4,5,6,7,8	10	RF	0.684433	0.677191
8	1,2,3,4,5,6,7,8	10	SVM	0.732234	0.677205
8	1,2,3,4,5,6,7,8	10	NN	0.673858	0.688127
8	1,2,3,4,5,6,7,8	10	LR	0.763113	0.705772
8	1,2,3,4,5,6,7,8	10	KNN	0.567682	0.63026
9	1,2,3,4,5,6,7,8,9	10	RF	0.689509	0.694918
9	1,2,3,4,5,6,7,8,9	10	SVM	0.741963	0.691442
9	1,2,3,4,5,6,7,8,9	10	NN	0.665398	0.690941
9	1,2,3,4,5,6,7,8,9	10	LR	0.763113	0.705787
9	1,2,3,4,5,6,7,8,9	10	KNN	0.560068	0.625182

逻辑回归算法和支持向量机算法,准确率都在70%以上,F1分数也在0.7左右,在大数据量条件下模型的性能可以得到保证。

地层参数预测对比维度,不同训练井下的地层参数预测模型表现如图14、图15所示。

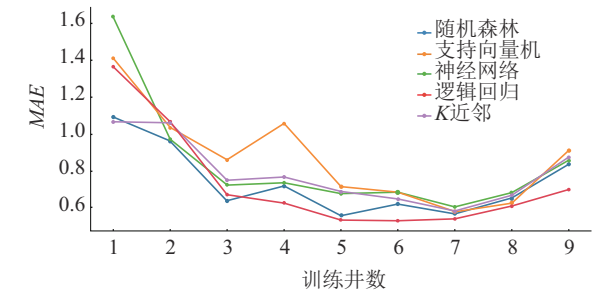


图14 不同训练井下模型的MAE

由图15可知,降维后地层参数的预测模型随着训练井数量增大变得更稳定,随着训练井数量增大,多数模型的MAE指标值都在1之下,而RMSE指标也在1左右,但在训练井数较少的时候,部分模型出现异常状况,总体来看逻辑回归算法、随机森

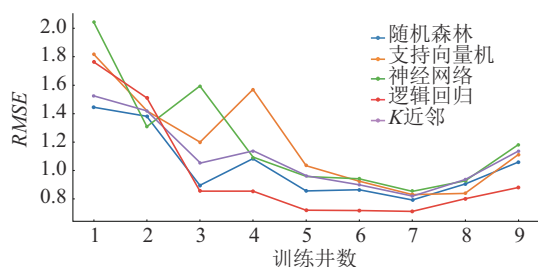


图 15 不同训练井下模型的 RMSE

林算法的地层参数预测更稳定,尤其是逻辑回归算法,在训练井数量增大以后,该算法的两个指标均在1之下,是区域性地层识别模型中地层参数预测表现最好的算法。

3 结论

本文基于20000余组钻井数据,在逻辑回归、支持向量机、K近邻算法、随机森林和神经网络五种算法的支持下建立了区域地层识别模型,5个算法模型在不同条件下表现存在差异。比如岩性预测时,逻辑回归适合大样本数据,平均准确率能达到75%以上;支持向量机算法随样本增加表现变好,且支持向量机算法很稳定,平均准确率和F1分数呈正相关;KNN算法随着样本增加性能变好,但是模型波动较大;随机森林性能随着样本增加而上升;神经网络模型性能随着样本增加而下降,且波动较大,但总体准确率高于60%。地层参数预测时,随机森林算法和支持向量机算法表现最好,且预测稳定。

本研究系统评估了不同训练井数对模型性能的影响,验证了支持向量机算法、随机森林算法以及逻辑回归算法在多数情况下岩性预测方面具有优越性。同时也验证了多数情况下随机森林和支持向量机算法在地层参数预测上的优越性。

参考文献:

[1] 刘秀娟,陈超,曾冲,等.利用测井数据进行岩性识别的多元统

计方法[J].地质科技情报,2007,(3):109-112.

- [2] 牟丹,张丽春,徐长玲.3种经典机器学习算法在火山岩测井岩性识别中的对比[J].吉林大学学报(地球科学版),2021,51(3):951-956.
- [3] 苏赋,马磊,罗仁泽,等.基于改进多分类孪生支持向量机的测井岩性识别方法研究与应用[J].地球物理学进展,2020,35(1):174-180.
- [4] 王淑盛,徐正光,刘黄伟,等.改进的K近邻方法在岩性识别中的应用[J].地球物理学进展,2004,(2):478-480.
- [5] 于代国,孙建孟,王焕增,等.测井识别岩性新方法——支持向量机方法[J].大庆石油地质与开发,2005,(5):93-95,110.
- [6] 徐晗,姚孔轩,程丹仪,等.基于非开挖随钻检测系统与随机森林的地层岩性识别[J].地质科技通报,2021,40(5):272-280.
- [7] 庞志超,肖华,毛晨飞,等.准噶尔盆地南缘地区含膏质地层岩性特征及测井识别方法[J].吉林大学学报(地球科学版),2024,54(4):1419-1431.
- [8] 王光宇,宋建国,徐飞,等.不平衡样本集随机森林岩性预测方法[J].石油地球物理勘探,2021,56(4):679-687,669.
- [9] 罗仁泽,虞娟娟,倪华玲,等.基于改进集成学习的测井岩性识别方法研究[J].石油物探,2023,62(2):212-224.
- [10] 苏自武,杨甘生,陈礼仪.利用神经网络原理对钻井中钻速的预测[J].探矿工程(岩土钻掘工程),2005,(1):48-50.
- [11] 王振豪.基于深度学习的测井岩性识别方法研究[D].大庆:东北石油大学,2024.
- [12] 尹琼.基于生成对抗神经网络的岩性识别方法与应用[J].测井技术,2025,49(1):57-67.
- [13] 张东晓,陈云天,孟晋.基于循环神经网络的测井曲线生成方法[J].石油勘探与开发,2018,45(4):598-607.
- [14] 高铁红.基于遗传算法和神经网络的钻井参数估计和岩性识别的研究[D].天津:河北工业大学,2000.
- [15] 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.
- [16] 岳永东,渠洪杰,谭春亮,等.基于支持向量机的测井岩性识别在松散沉积物调查中的应用研究[J].钻探工程,2021,48(4):29-36.
- [17] Senjoba L, Ikeda H, Toriya H, et al. Deep learning-based rock type identification using drill vibration frequency spectrum images[J]. International Journal of Mining, Reclamation and Environment, 2025, 39(1): 40-55.
- [18] Yaotu H, Xiaocheng Z, Zhengming X, et al. Cuttings Bed Height Prediction in Microhole Horizontal Wells with Artificial Intelligence Models[J]. Energies, 2022, 15(22), 8389.

(编辑 王跃伟)