文章编号:1009-9603(2023)01-0076-10 DOI:10.13673/j.cnki.cn37-1359/te.202205031

基于FL-XGBoost算法的砂泥岩识别方法
——以胜利油田牛庄地区为例

彭 英[1],李克文[2],朱应科[1],徐志峰[2],杨澎涛[1],孙秀玲[3]
(1.中国石化胜利油田分公司 物探研究院,山东 东营 257000; 2.中国石油大学(华东)计算机科学与技术学院,山东 青岛 266580; 3.山东胜软科技股份有限公司,山东 东营 257000)

摘要:砂泥岩识别任务通常基于测井曲线,依据经验公式、实地岩心取样、交会图和聚类分析等传统方法实现,但这些方法难以充分利用测井曲线所包含的砂泥岩特征,且精度低、效率低,人为影响因素大。为此,以测井和录井资料为基础,综合砂泥岩识别的关键技术难点,对测井参数进行敏感性分析,以选取适当的影响因素,通过多项预处理操作构建完整的训练数据集,并根据测井标签稀疏性的特点,引入Focal Loss函数,提出FL-XGBoost模型,进而开展胜利油田牛庄地区砂泥岩识别。研究结果表明,采用FL-XGBoost算法的砂泥岩识别模型对研究区砂泥岩识别的准确率达到了0.827。通过5种公开分类数据集设计对比实验,证明FL-XGBoost算法在识别分类领域上具有强泛化能力。

关键词:FL-XGBoost算法;迭代决策树;机器学习;砂泥岩识别;测井资料
中图分类号:TE319 **文献标识码**:A

FL-XGBoost algorithm-based method for identifying sandstone and mudstone: A case study of Niuzhuang area in Shengli Oilfield

PENG Ying[1], LI Kewen[2], ZHU Yingke[1], XU Zhifeng[2], YANG Pengtao[1], SUN Xiuling[3]
(1. *Geophysical Exploration Research Institute of SINOPEC Shengli Oilfield Company, Dongying City, Shandong Province*, 257000, *China*; 2. *College of Computer Science and Technology, China University of Petroleum (East China), Qingdao City, Shandong Province*, 266580, *China*; 3. *Shandong Shengruan Technology Co., Ltd., Dongying City, Shandong Province*, 257000, *China*)

Abstract: sandstone and mudstone identification tasks are usually based on logging curves and rely on traditional methods such as empirical formulas, field core sampling, cross plots, and cluster analysis, but these methods fail to make full use of the sandstone and mudstone features contained in the logging curves. At the same time, these traditional methods have low accuracy and slow efficiency and are greatly affected by human factors. To address the above problems, this paper uses logging data as the basis, combines the key technical difficulties of sandstone and mudstone identification, and conducts sensitivity analysis on logging parameters, so as to select appropriate influencing factors and construct a complete training data set through several pre-processing operations. In addition, the paper introduces the Focal Loss function and proposes the FL-XGBoost model according to the sparsity of logging labels and carries out sandstone and mudstone identification in Niuzhuang area of Shengli Oilfield. The experimental results show that the sandstone and mudstone identification model using the FL-XGBoost algorithm achieves an accuracy of 0.827 in identifying the sandstone and mudstone in the study area. Finally, the strong generalization ability of the FL-XGBoost algorithm in the identification classification field is verified through five publicly classified dataset design comparison experiments.

Key words: FL-XGBoost algorithm; iterative decision tree; machine learning; sandstone and mudstone identification; logging data

收稿日期:2022-05-15。
作者简介:彭英(1970—),男,河北迁安人,高级工程师,博士,从事油气勘探数据分析及勘探信息系统开发管理工作。E-mail:pengy.slyt@sinopec.com。
基金项目:国家自然科学基金项目"储层天然气水合物相变和渗流多场时空演化规律"(51991365),山东省自然科学基金项目"基于多源数据融合的浊积岩有效储层预测方法"(ZR2021MF082)。

岩性识别对石油勘探开发具有重要意义,已成为众多学者关注的焦点。砂泥岩识别是储层预测工作中非常重要的环节[1],也是诸多研究的基础,其所需的测井资料通常由专家按经验解释完成,因此识别结果存在一定的主观性。在常规的砂泥岩识别方法中[2-6],地震反演作为砂体预测的常规技术已得到广泛应用,但不论是叠后反演还是叠前反演,均受限于地震的纵向分辨率,井间预测结果分辨率较低、可靠性较弱,准确率有待进一步提高。对于岩性信息的获取多依靠实地岩心取样、交会图和聚类分析[7]等传统方法和数理统计方法,但这些方法仍存在人力和时间成本较高等局限,因此有必要提出更可靠、稳定的学习算法以解决地质应用中砂泥岩自动识别分类的问题。

近年来,随着计算机硬件性能的高速提升以及大数据技术的不断发展,对石油工业的发展产生了巨大的推动作用[8]。因此,将迅速发展的大数据技术与测井曲线相结合识别砂泥岩,已成为目前储层研究砂泥岩识别的重要手段[9-14]。机器学习算法从井点出发,充分挖掘地震属性与测井岩性敏感曲线之间的数据关系,最大限度地发挥地震属性的利用价值,其预测结果的纵向分辨率高于确定性反演,井间可靠性优于地质统计学反演。随机森林算法[15-17]的训练速度快、准确率较高,能够有效地运行于大型数据集,且引入随机性,不易过拟合;该算法对于不平衡的数据集可以平衡误差,但对于小型数据或低维数据(测井数据),则难以产生较好的分类,易出现很多相似的决策树,导致真实的预测结果被掩盖。深度神经网络算法[18-22]可以较好地解决非线性问题,进而实现面向相关专业领域的迁移学习,这是建立在充足训练数据量的基础之上,但若在岩性识别任务的训练过程中,面对较为稀少的测井数据,神经网络在推理过程中无法提取足够的测井特征,易导致过拟合问题[23],使得模型无法获得较高的准确率。XGBoost是一种基于迭代决策树模型的集成学习算法[24-26],是基于利用一阶导数相关信息的迭代决策树(Gradient Boosting Decision Tree,简称GBDT)的改进算法,在很大程度上提高了模型的训练速度和预测的准确度。对于深度学习算法而言,XGBoost算法只适用于处理结构化的特征数据,而直接对测井、录井曲线等数据进行处理则较为困难,且XGBoost算法的参数过多,调参复杂。

由以上分析可以得出,诸如随机森林、深度神经网络等机器学习算法可以较好的解决相关地质问题,已经获得了显著的效果,为提升地质工作效率提供了新的思路和方法,然而在砂泥岩识别领域仍存在关键技术难点:①样本集的选取以及预处理对于机器学习算法的性能具有较大影响。②砂泥岩岩性数据复杂多样,根据测井参数与岩性的分析,选取合适的测井曲线参数是影响砂泥岩识别准确性的关键之一。因此,需基于特定样本数据设计相关人工智能算法与超参数调优策略,充分发挥智能算法的优势,以满足砂泥岩识别准确性的需求。

为此,笔者以测井和录井资料为基础,考虑砂泥岩识别的关键技术难点,对测井参数进行敏感性分析,以明确影响因素;通过多项预处理操作构建完整的训练数据集,根据测井标签稀疏性的特点,将Focal Loss函数引入XGBoost算法(FL-XGBoost算法),对胜利油田牛庄地区构建砂泥岩识别模型;并将随机森林、深度神经网络算法的训练结果作为对照,以最终砂泥岩识别分类结果的准确率作为评价标准,验证FL-XGBoost算法应用于测井砂泥岩识别的可行性;最后通过5种公开分类数据集设计对比实验,验证FL-XGBoost算法在识别分类领域上的强泛化能力。研究成果可以为FL-XGBoost算法对砂泥岩识别的可行性提供理论依据,为传统的测井岩性识别提供新的思路。

1 相关理论

GBDT算法是一个树结构(可以是二叉树或非二叉树)[27],由多棵决策树组成,以所有决策树的结论累加起来作为最终答案,具体原理为:每个非叶子节点表示一个特征属性的测试,每个分支代表这个特征属性在某个值域的输出,而每个叶子节点存放一个类别,迭代决策的过程是从根节点开始,测试待分类项中相应的特征属性,并按照其值选择输出分支,直到到达叶子节点,将叶子节点存放的类别作为决策结果[27]。GBDT算法的思路是不断地添加决策树,进行特征分裂以生长一棵决策树,且每次添加一个决策树,为学习一个新函数,进而拟合上次预测的残差。当训练完成得到k棵决策树,则要预测一个样本的分数,其实就是根据这个样本的特征,在每棵决策树中落到对应的一个叶子节点,每个叶子节点即对应一个分数,最后只需将每棵决策树对应的分数相加即为该样本的预测值。

XGBoost算法是基于二阶泰勒展开式将损失函数展开,并且将正则项置于目标函数之外,这降低了模型的复杂度,更易于获得最优解,通过控制目标函数的不断下降,使得模型能够更好地收敛,有

效避免过拟合，从而提高了预测准确率。该算法在训练前对数据进行预处理，将其结果保存，在后面的迭代中可以重复使用，从而降低计算复杂度，实现并行化，提高整体计算效率。

2　基于FL-XGBoost算法的砂泥岩识别模型构建

基于GBDT与XGBoost算法，将不平衡样本分类思想引入训练损失函数，构建基于FL-XGBoost算法的砂泥岩识别模型。结合砂泥岩识别存在的关键技术难点，首先对测井参数进行敏感性分析，以明确影响因素，通过多项预处理操作构建完整的训练数据集并将其输送至FL-XGBoost模型中进行训练，迭代计算FL损失并判断是否继续收敛，期间进行超参数调优，最终获得训练完备的砂泥岩识别模型。基于FL-XGBoost算法的砂泥岩识别流程如图1所示。

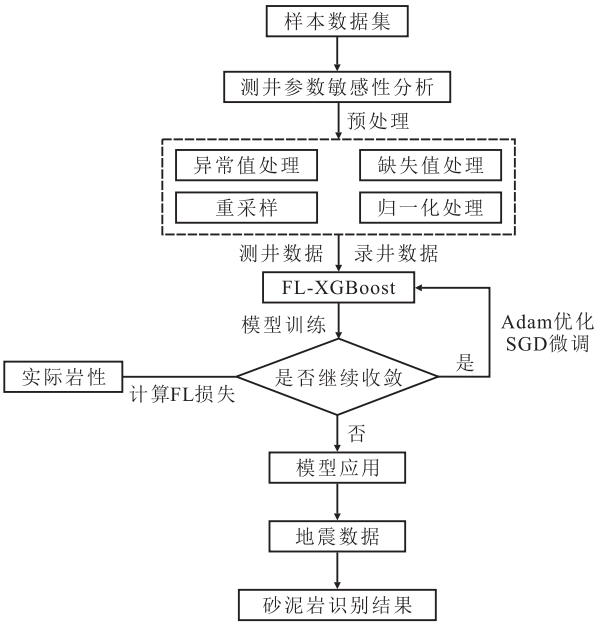


图1　基于FL-XGBoost算法的砂泥岩识别流程

Fig.1　Flow chart of sandstone and mudstone identification based on FL-XGBoost algorithm

Focal Loss是LIN等于2017年专门为解决不平衡分类问题提出的损失函数[28]。其从2个方面解决数据类别不平衡的问题：①损失函数更加倾向于关注少数类样本。②避免易分类样本主导模型训练过程而导致的性能降低。相对于庞大的地震数据体，测井与录井标签数据存在稀疏性，基于机器学习算法的砂泥岩识别可视为非平衡样本训练问题。

FL-XGBoost算法的思路与集成学习中的GBDT算法的类似。FL-XGBoost算法训练时每一次迭代会增加一棵决策树来拟合上一次迭代过程中的真实值与预测值之间的FL残差，进而逐渐逼近真实值，其训练过程中的目标函数为：

$$obj = L_{\mathrm{FL}} + \sum_{i=1}^{n}\Omega\left(f_i\right) \tag{1}$$

$\sum_{i=1}^{n}\Omega\left(f_i\right)$为复杂度函数项，也称为正则化项，将$L_{\mathrm{FL}}$展开得到：

$$L_{\mathrm{FL}} = \begin{cases} -\alpha\left(1-\hat{y}_c\right)^{\beta}\lg\hat{y}_c, & y_c = 1 \\ -(1-\alpha)\left(\hat{y}_c{}^{\beta}\right)\lg\left(1-\hat{y}_c\right) & y_c = 0 \end{cases} \tag{2}$$

在(2)式中，通过引入系数α来调整测井标签中不同参数在损失函数中的权重，引入聚焦稀疏系数β来调整易分类样本和难分类样本的损失权重。

将$\Omega\left(f_i\right)$展开得到：

$$\Omega(f_i) = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T}w_j{}^2 \tag{3}$$

新生成的决策树需拟合上一迭代预测的残差，即第t次迭代目标函数，其砂泥岩识别结果可以表示为：

$$obj^{(t)} = \sum_{i=1}^{n}l\left[y_c,\hat{y}_c^{(t-1)} + f_t(x_c)\right] + \Omega(f_k) + \sum_{k=1}^{t-1}\Omega(f_i) \tag{4}$$

将损失函数使用泰勒二阶展开，引入正则项并去除常数项后得到：

$$obj^{(t)} = \sum_{i=1}^{n}\left[g_i w_{q(x_c)} + \frac{1}{2}h_i w_{q(x_c)}{}^2\right] + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T}w_j{}^2 \tag{5}$$

将(5)式中表示的所有训练样本按照叶子节点进行分组得到：

$$obj^{(t)} = \sum_{j=1}^{T}\left[\left(\sum_{i\in n}g_i\right)w_j + \frac{1}{2}\left(\sum_{i\in n}h_i + \lambda\right)w_j{}^2\right] + \gamma T \tag{6}$$

FL-XGBoost算法中经过k次迭代后，形成的决策树模型对第c个样本的输出结果为：

$$\hat{y}_c = \sum_{k=1}^{K}f_i(x_c) \qquad f_k \in F \tag{7}$$

$$f_k = w_{q(x)} \qquad w \in R^T \qquad q{:}R^d \rightarrow \{1,2,...,T\} \tag{8}$$

3　应用实例分析

3.1　研究区概况

牛庄洼陷为济阳坳陷东营凹陷中南部的次级

洼陷，为渤海湾盆地油气最丰富的地区之一。其南北两侧均受断层控制，构造活动较为频繁，沉积岩性主要为深灰色的厚层泥岩、灰质砂岩和泥质粉砂岩等。牛庄洼陷发育多种类型的油气藏，对其地层岩性的准确识别可为后期的油气预测奠定基础。

3.2　数据获取及预处理

本次研究数据来源于牛庄洼陷220口井的测井及录井数据，其中200口井的测井曲线为las文件格式，20口井的测井曲线为文本文档，采样间隔均为0.125 m。目标任务为完成纯泥岩、砂岩、其他泥岩（除纯泥岩之外的泥岩）、其他岩层（除纯泥岩、砂岩、其他泥岩三者之外的岩层）4类岩性的识别。利用实际采集到的测井和录井数据，检查标签数据，建立样本库，并对样本数据进行预处理，包括：①异常值处理。根据业务专家制定的不同特征的合理取值范围，对数据中的特征值设置阈值并进行过滤，对超过阈值的不合理值依据临近数据或单井平均数据进行修正。②缺失值处理。对于测井曲线中的缺失数据，利用贝叶斯估计插补缺失值。③重采样。将测井数据采样间隔为0.1 m的井应用插值进行重采样，采样间隔为0.125 m；对标签数据进行上采样，以保证标签类别均衡。④数据归一化。在机器学习领域中，不同特征向量往往具有不同的量纲和单位，这样会影响数据分析的结果。为了消除特征向量的量纲影响，需进行数据标准化处理，以解决数据指标之间的可比性。而原始数据经过数据归一化处理后，各指标处于同一数量级，适合进行综合对比评价（图2）。

最终将整个数据划分为训练集、测试集和验证集。训练集和测试集数据是利用岩屑录井资料确定，为避免岩屑录井资料的错误，在岩屑录井图上，显示4条岩性曲线，即自然电位曲线（*SP*）、自然伽马曲线（*GR*）、井径曲线（*CAL*）和声波时差曲线（*AC*）。业务专家现场对岩性分类进行审定，去除不可靠的岩性分类，最终完成纯泥岩、砂岩、其他泥岩、其他岩层4类样本的标定工作，按点构建1 048 575条样本数据。4个点构建1个深度段，按深度段构建28 619条样本数据（表1）。

3.3　特征参数提取

针对测井曲线数据进行多维度表征，测井曲线

DEPTH/m	*AC*/（μs·m^{-1}）	*SP*/mV	*CNL*/%	*DEN*/（g·cm^{-3}）	*GR*/API
1 800.000	351.755 0	136.404 0	26.988 0	2.196 0	55.290 0
1 800.125	351.032 0	136.269 0	27.153 0	2.190 0	53.292 0
1 800.250	351.508 0	136.084 0	27.828 0	2.185 0	51.212 0
1 800.375	352.059 0	135.872 0	28.542 0	2.185 0	50.669 0
1 800.500	351.496 0	135.591 0	29.815 0	2.183 0	51.181 0
1 800.625	349.598 0	135.390 0	30.552 0	2.183 0	52.149 0
1 800.750	345.678 0	135.314 0	31.119 0	2.184 0	53.355 0
1 800.875	343.309 0	135.282 0	31.170 0	2.185 0	53.831 0
1 801.000	341.647 0	135.284 0	30.887 0	2.185 0	54.465 0
1 801.125	342.945 0	135.351 0	30.590 0	2.185 0	55.259 0
1 801.250	348.015 0	135.461 0	30.321 0	2.187 0	57.202 0
1 801.375	352.259 0	135.628 0	30.327 0	2.197 0	58.808 0
1 801.500	358.608 0	135.984 0	30.570 0	2.208 0	61.481 0
1 801.625	361.819 0	136.397 0	30.745 0	2.226 0	63.232 0
1 801.750	364.199 0	136.659 0	30.599 0	2.237 0	65.250 0
1 801.875	364.672 0	136.875 0	30.210 0	2.246 0	65.911 0
1 802.000	364.052 0	137.023 0	29.393 0	2.248 0	65.629 0
1 802.125	362.492 0	136.966 0	28.914 0	2.245 0	64.840 0
1 802.250	358.579 0	136.808 0	28.510 0	2.240 0	63.747 0
1 802.375	355.594 0	136.580 0	28.338 0	2.230 0	63.224 0
1 802.500	351.436 0	136.157 0	28.005 0	2.222 0	62.552 0

重采样 → 归一化

DEPTH/m	*AC*	*SP*	*CNL*	*DEN*	*GR*
1 800.000	0.484 0	0.786 0	0.404 0	0.382 0	0.190 0
1 800.125	0.481 0	0.784 0	0.408 0	0.375 0	0.169 0
1 800.250	0.483 0	0.782 0	0.421 0	0.370 0	0.147 0
1 800.375	0.484 0	0.779 0	0.435 0	0.370 0	0.141 0
1 800.500	0.483 0	0.774 0	0.460 0	0.367 0	0.147 0
1 800.625	0.477 0	0.772 0	0.475 0	0.367 0	0.157 0
1 800.750	0.465 0	0.771 0	0.486 0	0.368 0	0.170 0
1 800.875	0.457 0	0.770 0	0.487 0	0.370 0	0.175 0
1 801.000	0.452 0	0.770 0	0.482 0	0.370 0	0.182 0
1 801.125	0.456 0	0.771 0	0.476 0	0.370 0	0.190 0
1 801.250	0.472 0	0.773 0	0.470 0	0.372 0	0.211 0
1 801.375	0.485 0	0.775 0	0.470 0	0.384 0	0.228 0
1 801.500	0.505 0	0.780 0	0.475 0	0.396 0	0.256 0
1 801.625	0.515 0	0.786 0	0.479 0	0.417 0	0.275 0
1 801.750	0.522 0	0.790 0	0.476 0	0.430 0	0.297 0
1 801.875	0.523 0	0.794 0	0.468 0	0.440 0	0.304 0
1 802.000	0.522 0	0.796 0	0.452 0	0.443 0	0.301 0
1 802.125	0.517 0	0.795 0	0.442 0	0.439 0	0.292 0
1 802.250	0.505 0	0.793 0	0.434 0	0.433 0	0.281 0
1 802.375	0.495 0	0.789 0	0.431 0	0.422 0	0.275 0
1 802.500	0.483 0	0.783 0	0.424 0	0.413 0	0.268 0

图2　数据重采样和归一化示例

Fig.2　Diagram of data resampling and normalization

表1　样本数据的样式

Tabel1　Sample data style

井名	层顶深/m	层底深/m	GR_{mean}	SP_{mean}	AC_{mean}	GR_{max}	SP_{max}	AC_{max}	GR_{min}	SP_{min}	AC_{min}	GR_{med}	SP_{med}	AC_{med}	岩性分类
A6	1 000.0	1 000.5	0.314 0	0.330 9	0.641 0	0.320 5	0.331 6	0.642 5	0.309 1	0.330 6	0.639 4	0.313 1	0.330 7	0.641 0	B
A6	1 000.5	1 001.0	0.351 8	0.332 0	0.620 7	0.361 3	0.332 3	0.636 2	0.334 9	0.331 5	0.599 2	0.355 6	0.332 2	0.623 7	B
A6	1 001.0	1 001.5	0.320 0	0.330 7	0.562 8	0.351 8	0.331 0	0.578 9	0.276 5	0.330 4	0.548 2	0.325 9	0.330 7	0.562 0	B
A6	1 001.5	1 002.0	0.297 0	0.331 4	0.615 6	0.335 5	0.331 8	0.637 2	0.271 4	0.330 6	0.592 0	0.290 5	0.331 6	0.616 7	B
A6	1 002.0	1 002.5	0.312 0	0.329 9	0.433 6	0.337 8	0.330 5	0.567 5	0.270 6	0.329 4	0.304 5	0.319 8	0.329 7	0.431 3	B
A6	1 002.5	1 003.0	0.252 2	0.331 8	0.312 6	0.269 7	0.334 0	0.371 8	0.241 7	0.331 4	0.272 9	0.248 7	0.332 8	0.303 0	B
A6	1 003.0	1 003.5	0.318 8	0.329 0	0.479 7	0.335 6	0.333 5	0.548 9	0.293 4	0.329 7	0.410 3	0.323 2	0.332 1	0.479 8	B
A6	1 003.5	1 004.0	0.313 5	0.331 0	0.622 2	0.327 9	0.329 5	0.652 9	0.299 3	0.328 7	0.581 7	0.313 5	0.328 9	0.627 0	B
A6	1 004.0	1 004.5	0.286 6	0.332 0	0.662 4	0.296 3	0.331 7	0.665 7	0.273 2	0.330 1	0.658 3	0.288 4	0.331 0	0.663 0	B
A6	1 004.5	1 005.0	0.296 7	0.332 1	0.653 0	0.329 5	0.332 2	0.660 7	0.274 8	0.331 8	0.645 6	0.291 2	0.332 0	0.652 9	B
A6	1 005.0	1 005.5	0.342 5	0.334 1	0.639 1	0.345 4	0.332 5	0.642 4	0.339 8	0.332 0	0.636 9	0.342 5	0.332 1	0.638 5	B
A6	1 005.5	1 006.0	0.342 2	0.333 0	0.642 9	0.351 9	0.334 9	0.645 6	0.337 9	0.333 1	0.639 0	0.339 5	0.334 2	0.643 5	B
A6	1 006.0	1 006.5	0.378 9	0.331 7	0.632 6	0.400 6	0.334 4	0.643 3	0.359 8	0.331 4	0.621 3	0.377 6	0.333 1	0.632 9	B
A6	1 006.5	1 007.0	0.378 7	0.332 0	0.614 4	0.397 7	0.332 3	0.616 9	0.383 0	0.331 3	0.622 9	0.386 9	0.331 6	0.613 8	B
A6	1 007.0	1 007.5	0.366 6	0.331 6	0.614 3	0.386 7	0.332 4	0.641 5	0.328 6	0.331 4	0.614 0	0.375 6	0.332 1	0.614 3	B
A6	1 007.5	1 008.0	0.294 2	0.332 0	0.614 4	0.312 5	0.332 6	0.627 9	0.269 6	0.331 3	0.615 5	0.297 3	0.331 5	0.620 7	B
A6	1 008.0	1 008.5	0.250 4	0.331 9	0.614 3	0.257 4	0.332 1	0.643 4	0.242 6	0.331 6	0.632 0	0.250 8	0.332 0	0.639 0	B
A6	1 008.5	1 009.0	0.305 9	0.333 8	0.621 2	0.341 6	0.332 5	0.643 4	0.274 3	0.331 6	0.637 8	0.308 0	0.331 7	0.640 4	B
A6	1 009.0	1 009.5	0.331 9	0.333 2	0.638 4	0.347 0	0.334 4	0.646 2	0.303 0	0.333 0	0.638 2	0.338 8	0.333 8	0.642 1	B

按点构建以及按0.5 m每段提取特征参数。牛庄洼陷主要为砂泥岩沉积，且该区测井资料大多是2010年以前测得，9条基础测井曲线齐全，其他测井曲线较少，其中与岩性相关的测井曲线有*GR*，*SP*，*AC*和*CAL*曲线，而*CAL*曲线受钻井和裂缝的影响较大，因此选取*AC*，*GR*和*SP*这3条测井曲线作为岩性识别的基础数据。录井资料的采样间隔为0.5 m，测井资料的采样间隔为0.125 m，为了匹配录井数据，将测井资料按照0.5 m进行特征参数提取，特征参数有最大值、最小值、平均值、标准差、中位数、累加值、数值排序的百分比；经过特征参数与岩性参数交汇分析，优选最大值、最小值、中位数、平均值作为测井曲线特征，分别提取每条测井曲线同一时窗内的最大值、最小值、中位数和平均值作为曲线的特征。

将处理后的特征数据与录井数据按深度进行匹配构建样本数据，并将样本数据划分为训练集和验证集，其中训练集样本占样本总数的80%，验证集样本占样本总数的20%。标签共包含4类，分别为纯泥岩、砂岩、其他泥岩和其他岩层。

3.4　砂泥岩识别结果对比

分别使用FL-XGBoost和XGBoost、随机森林、深度神经网络算法学习胜利油田牛庄洼陷的砂泥岩样本数据，并进行超参数设置、模型性能以及应用效果的对比分析。

3.4.1　FL-XGBoost算法

为契合砂泥岩识别，改进目标函数的计算方式，进一步提高模型的精确度，并将目标函数的优化问题转化为求二次函数的最小值问题，利用损失函数的二阶导数信息训练决策树模型，同时将树复杂度作为正则化项加入到目标函数中，以提升模型的泛化能力。XGBoost模型中有多个超参数，选出对模型影响较大的超参数作为网格搜索法遍历寻优的参数，其余超参数为默认值。在本次应用实例中，分别对以FL-XGBoost算法和XGBoost算法为基础设计的30棵决策树构建对比实验，即初始迭代30次。初始学习率采用0.01，控制每次迭代更新权重时的步长，设置每棵决策树的初始深度为3，最大值为20，并且设置早停轮数，防止模型过拟合。

由表2可知模型学习率、决策树的最大深度和迭代产生决策树超参数的数量分别为10，10和5，将以上参数进行组合得到500条超参数组合。运用网格搜索法，遍历网格中的500条超参数组合，寻找最优超参数组合。随机取80%的训练集数据分批输

入到XGBoost模型中，用剩余20%的数据集对模型的精度进行评估。根据评估结果的精确度调整模型所用样本和超参数。利用训练好的XGBoost模型，按照0.5 m为一段对新井的测井数据进行预测，并输出预测结果，将预测结果与标签值进行比较，只统计纯泥岩和砂岩预测正确的数量，其他泥岩和其他岩性不参与统计。其中，预测准确率=(纯泥岩预测为泥岩+砂岩预测为砂岩)/(泥岩样本总数+砂岩样本总数)。

表2　FL-XGBoost算法参数设置

Tabel2　Parameter settings of FL-XGBoost algorithm

超参数	取值范围	步长
模型学习率	[0.01,0.1]	0.01
决策树的最大深度	[10,20]	1
迭代产生决策树的数量	[100,500]	100

表3显示在1 000条超参数组合中具有代表性的组合与预测准确率，当决策树的最大深度为20、最优迭代次数为487次，FL-XGBoost模型的预测准确率达到最高值，为0.827，其在测试集下的推理速度为0.192 0 s，在迭代超过487次以后，预测准确率出现持续的下降，推测模型出现过拟合现象。由此得到，当FL-XGBoost模型在更加侧重于测井资料方面训练，而非无关(负)样本训练时，模型的预测精度将会得到显著提高。

表3　XGBoost模型与FL-XGBoost模型迭代及识别结果

Tabel3　Iteration and identification results of XGBoost and FL-XGBoost models

算法	决策树的最大深度	最优迭代次数	学习率	准确率
XGBoost	10	312	0.03	0.721
	15	334	0.05	0.724
	20	451	0.01	0.813
FL-XGBoost	10	278	0.04	0.714
	15	329	0.05	0.756
	20	487	0.01	0.827

3.4.2　随机森林算法

随机森林算法通过集成学习的方法集成多棵决策树，每一棵都是一个分类器，对于每一个输入样本，每棵决策树与分类结果是一一对应的，通过集成分类投票结果，将投票次数最多的类别指定为最终的输出。笔者将测井数据集作为输入，在基尼指数与交叉熵2种标准下，对随机森林算法进行训练，并展示了迭代产生数、评判标准、决策树的最大深度等超参数对砂泥岩识别结果的影响(图3)。

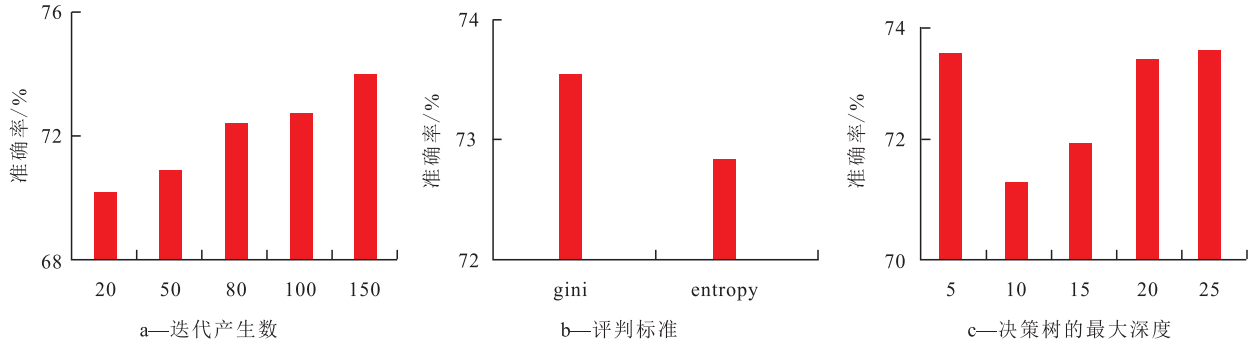
由图3可知，随着迭代次数的增加，随机森林算法对砂泥岩的识别精度也在提高，但对于诸如测井数据的小样本数据，识别效果并不是最优的。在多参数设置最优的情况下，测试集的识别精度仅为74.13%，其在测试集下的推理速度为0.214 6 s。

3.4.3　深度神经网络算法

深度神经网络算法是机器学习的分支，是一种试图使用包含复杂结构或由多重非线性变换构成的多个处理层对数据进行高层抽象的算法。笔者将测井数据集作为输入，设计对应的深度神经网络模型，通过控制不同的隐含层数目与迭代次数进行训练，最终得到不同的砂泥岩识别结果(表4)，在多参数设置最优的情况下，测试集的识别精度仅为0.745，其在测试集下的推理速度为1.453 1 s。

深度神经网络算法虽然具有强大的非线性拟合能力，但这是建立在充足训练数据量基础之上的。面对较为稀少的测井数据量，该模型在推理过程中无法提取足够的测井特征，导致模型无法获得较高的准确率。

综合来看，采用FL-XGBoost算法的砂泥岩识别结果与采用随机森林、深度神经网络算法所得到的识别结果进行比较(图4)，结果表明在使用交叉验证测试模型精度及相同训练数据下，使用FL-XGBoost模型的训练速度最快，识别准确率有明显提



a—迭代产生数　　b—评判标准　　c—决策树的最大深度

图3　随机森林算法结果分析

Fig.3　Result analysis of random forest algorithm

表4　深度神经网络模型及识别结果

Tabel4　Deep neural network model and identification results

迭代次数	隐含层数目	准确率
20	2层	0.655
50	2层	0.690
20	3层	0.677
100	3层	0.704
20	4层	0.736
50	4层	0.745

升，同时计算复杂度更低，为砂泥岩的测井识别提供了新的思路。

3.5　公共数据集及实验对比分析

在通用的分类识别问题中，业内常采用准确率、$F1$值、AUC等作为评估指标，其计算所需的混淆矩阵如表5所示。

利用混淆矩阵可计算相应的准确率、召回率、$F1$值和AUC等评估指标，其计算式如下：

$$Pre = \frac{TP}{TP + FP} \tag{9}$$

$$Rec = \frac{TP}{TP + FN} \tag{10}$$

$$F1 = \frac{2 \times Pre \times Rec}{Pre + Rec} \tag{11}$$

$$AUC = \frac{\sum^{i \in P_c} rank_i - \dfrac{NP(NP+1)}{2}}{NP \times NN} \tag{12}$$

KEEL是一种集成海量标准分类数据集的综合库，为验证FL-XBoost算法的有效性以及不同智能算法之间的性能差异，采用KEEL中的mushroom（蘑菇是否有毒的分类数据集）、magic（魔法射线望远镜数据集）、spambase（电子邮件分类数据集）、titanic

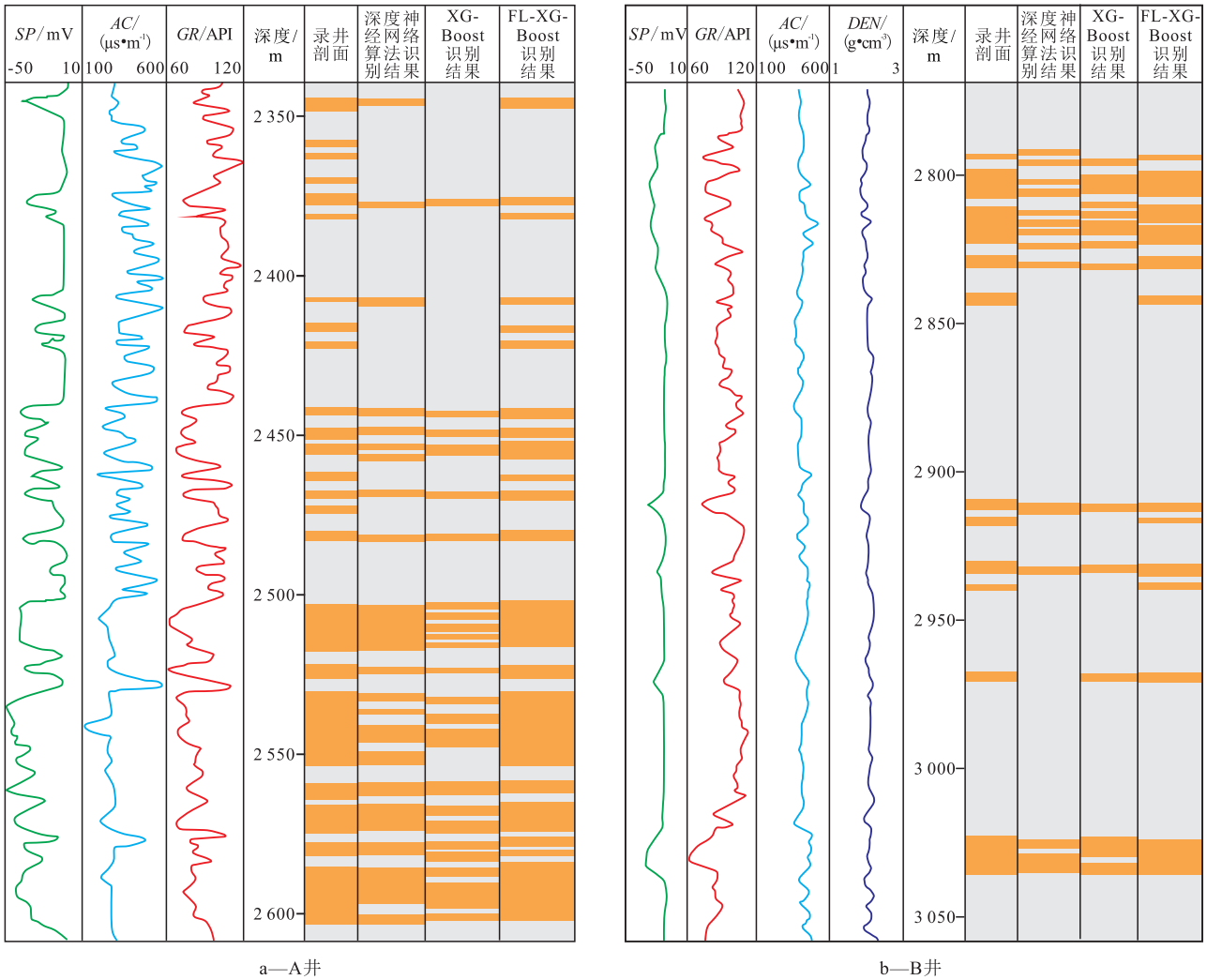


图4　不同算法的识别结果对比

Fig.4　Comparison of identification effects of different algorithms

表5　混淆矩阵
Tabel5　Confusion matrix

真实类别	预测类别	
	正类	负类
正类	*TP*	*FN*
负类	*FP*	*TN*

（泰坦尼克轮船乘客的幸存分类数据集）、phoneme（声音分类数据集）等5种公共数据集，其分别为特征数不同、样本量不同的代表性数据集。利用训练完备的随机森林、深度神经网络、XGBoost和FL-XGBoost算法分别对这5个数据集进行预测，并以准确率、*F*1值和*AUC*作为评价指标，其数据集信息与预测结果如表6所示。

由表6可知，FL-XGBoost算法在5种公开数据集中的预测结果均优于随机森林、深度神经网络和XGBoost算法，由于XGBoost算法中的正则化项可在一定程度上解决稀疏测井数据过拟合问题，不仅使用一阶导数，还推理二阶导数，使得损失函数更加精确。在此基础之上，FL-XGBoost算法的损失函数相比于均方根误差、交叉熵等损失，Focal Loss更加适用于难训练的样本，如测井曲线、录井数据等。因此，FL-XGBoost算法的预测准确率相对于随机森林、深度神经网络、XGBoost算法具有明显提升，具备更好的泛化能力。

4　结论

以测井、录井资料为基础，结合砂泥岩识别任务存在的关键技术难点，对测井参数进行敏感性分析，以选取适当的影响因素。通过多项预处理操作构建完整的训练数据集，根据测井标签稀疏性的特点，将Focal Loss函数引入XGBoost算法，并对胜利油田牛庄洼陷构建砂泥岩识别模型，相比于随机森林和深度神经网络算法，FL-XGBoost算法可以解决常规测井砂泥岩识别过拟合和准确率较低的问题。

FL-XGBoost算法应用于砂泥岩识别任务的准确率达到0.827，构建智能化工作流程，同时形成测井岩性识别样本库，具有一定的泛化能力，可以在砂岩油藏中推广应用。FL-XGBoost算法在KEEL

表6　5种公共数据集及预测结果
Tabel6　Five public datasets and prediction results

数据集	特征数	样本量	算法	准确率	*F*1值	*AUC*
mushroom	22	8 124	随机森林	0.801	0.839	0.896
			深度神经网络	0.848	0.854	0.908
			XGBoost	0.852	0.873	0.924
			FL-XGBoost	0.864	0.880	0.932
magic	10	19 020	随机森林	0.866	0.806	0.847
			深度神经网络	0.874	0.813	0.850
			XGBoost	0.880	0.821	0.856
			FL-XGBoost	0.887	0.828	0.858
spambase	57	4 597	随机森林	0.941	0.953	0.930
			深度神经网络	0.948	0.961	0.947
			XGBoost	0.954	0.963	0.952
			FL-XGBoost	0.962	0.966	0.963
titanic	3	2 201	随机森林	0.764	0.837	0.653
			深度神经网络	0.782	0.841	0.661
			XGBoost	0.794	0.868	0.676
			FL-XGBoost	0.816	0.873	0.719
phoneme	5	5 404	随机森林	0.883	0.883	0.917
			深度神经网络	0.861	0.857	0.894
			XGBoost	0.894	0.894	0.925
			FL-XGBoost	0.907	0.912	0.938

中5种公开数据集的预测效果均优于随机森林、深度神经网络和XGBoost算法，证明该算法的有效性与泛化性。

符号解释

AUC——数据集中任取一个正样本和负样本，预测正例排在负例前面的概率，%；
c——地震道的道数，个；
f——*F*中的某棵决策树，棵；
f_i——*F*中的第*i*棵决策树，棵；
f_k——*F*中的第*k*棵决策树，棵；
f_t——*F*中的第*t*棵决策树，棵；
F——特征空间；
*F*1——准确率与召回率的调和平均值，%；
FN——错误的负例，即错误的将样本中的正例识别为负例，%；
FP——错误的正例，即错误的将样本中的负例识别为正例，%；
g_i——损失函数的一阶导数；
h_i——损失函数的二阶导数；
i——当前决策树的棵树，棵；
j——当前的叶子节点，个；
k——当前的迭代次数，次；
K——总的迭代次数，次；
l——损失函数；
L_{FL}——Focal Loss误差项，%，
n——FL-XGBoost算法的决策树数量，棵；
N——负样本，%；
NN——负样本（多数类）总数，%；
NP——正样本（少数类）总数，%；
obj——目标函数；
P——正样本，%；
Pre——准确率，%；
P_c——正样本的类别，无单位；
q——表示样本x_c被预测后落入在对应节点上的概率，%；
$rank_i$——正样本的置信度排序，%；
R——每个节点的分值集合；
R^d——每个节点的集合；
Rec——召回率，%；
t——当前迭代次数，次；
T——叶子节点的总个数，个；
TN——被预测为负类的负样本，%；
TP——被预测为正类的正样本，%；
w——叶子节点的分值，%；
w_j——第*j*个叶子节点的分值，%；
w_q——第*q*个叶子节点的分值，%；
x_c——多地震道训练数据，个；
y_c——与x_c对应的测井和录井曲线标签数据；%；
$\hat{y}_c$——训练数据x_c经所有预测后得到的估计值，%；
α——系数，%；
β——聚焦稀疏系数，%；
γ——可以控制叶子节点的个数，个；
λ——分数控制系数，可以控制叶子节点的分数不会过大，防止过拟合，%；
$\Omega(f_t)$——决策树的正则化项。

参考文献

[1] 梁久红，张丽艳，韩冰冰，等.松辽盆地古龙页岩油储层岩性识别与流体评价技术[J]大庆石油地质与开发，2020，39(3)：163-169.
LIANG Jiuhong，ZHANG Liyan，HAN Bingbing，et al.Lithology identification and fluid evaluation techniques for the Gulong shale oil reservoirs in Songliao Basin[J].Petroleum Geology & Oilfield Development in Daqing，2020，39(3)：163-169.

[2] 马峥，张春雷，高世臣.主成分分析与模糊识别在岩性识别中的应用[J].岩性油气藏，2017，29(5)：127-133.
MA Zheng，ZHANG Chunlei，GAO Shichen.Application of principal component analysis and fuzzy recognition in lithologic identification[J].Lithologic Reservoir，2017，29(5)：127-133.

[3] 宋增强.叠后地震反演技术预测河道砂体[J].大庆石油地质与开发，2018，37(4)：151-156.
SONG Zengqiang.Prediction of the channel sandbody by post-stack seismic inverting technique[J].Petroleum Geology & Oilfield Development in Daqing，2018，37(4)：151-156.

[4] 王瑞，朱筱敏，王礼常.用数据挖掘方法识别碳酸盐岩岩性[J].测井技术，2012，36(2)：197-201.
WANG Rui，ZHU Xiaomin，WANG Lichang.Using data mining to identify carbonate lithology[J].Well Logging Technology，2012，36(2)：197-201.

[5] 刘君毅，王清辉，冯进，等.基于岩石物理相的复杂砂岩储层分类评价——以珠江口盆地惠州凹陷为例[J].中国石油勘探，2021，26(2)：92-102.
LIU Junyi，WANG Qinghui，FENG Jin，et al.Classification and evaluation of complex sandstone reservoirs based on petrophysical phases-an example of Huizhou depression in the Pearl River mouth basin[J].China Petroleum Exploration，2021，26(2)：92-102.

[6] 周萍.王府断陷火石岭组火山岩岩性及岩相识别[J].断块油气田，2020，27(2)：188-192.
ZHOU Ping.Identification of volcanics reservoir lithology and lithofacies in Huoshiling Formation of Wangfu fault depression[J].Fault-Block Oil and Gas Field，2020，27(2)：188-192.

[7] 张涛，莫修文.基于交会图与模糊聚类算法的复杂岩性识别[J].吉林大学学报：地球科学版，2007，52(增刊1)：109-113.
ZHANG Tao，MO Xiuwen.Complex lithologic identification based on cross plot and fuzzy clustering algorithm[J].Journal of Jilin University：Earth Science Edition，2007，52(S1)：109-113.

[8] 周景润.深度学习在油气储层岩性识别中的应用研究[D].兰

州:兰州理工大学,2021.
ZHOU Jingrun.Research on application of deep learning in lithology recognition of oil and gas reservoir[D]. Lanzhou: Lanzhou University of Technology,2021.

[9] 牟丹,王祝文,黄玉龙,等.基于SVM测井数据的火山岩岩性识别:以辽河盆地东部坳陷为例[J].地球物理学报,2015,58(5):1 785-1 793.
MOU Dan, WANG Zhuwen, HUANG Yulong, et al. Lithological identification of volcanic rocks from SVM well logging data: A case study in the eastern depression of Liao heBasin[J].Chinese Journal of Geophysics,2015,58(5):1 785-1 793.

[10] 江凯,王守东,胡永静,等.基于Boosting Tree算法的测井岩性识别模型[J].测井技术,2018,42(4):396.
JIANG Kai,WANG Shoudong,HU Yongjing,et al.Logging lithology identification model based on boosting tree algorithm[J].Logging Technology,2018,42(4):396.

[11] XU Ting, CHANG Ji, FENG Deyong, et al. Evaluation of active learning algorithms for formation lithology identification[J].Journal of Petroleum Science and Engineering,2021,206(4):108999.

[12] REN Xiaoxu,HOU Jiagen,SONG Suihong,et al.Lithology identification using well logs: A method by integrating artificial neural networks and sedimentary patterns[J].Journal of Petroleum Science and Engineering,2019,182(5):106336.

[13] XIANG Min, QIN Pengbo, ZHANG Fengwei.Research and application of logging lithology identification for igneous reservoirs based on deep learning[J].Journal of Applied Geophysics,2020,173(1):103929.

[14] 牟丹.辽河盆地中基性火成岩测井岩性识别方法研究[D].长春:吉林大学,2015.
MOU Dan.Methods research on logging lithology identification for intermediatel/basaltic rocks in Liaohe Basin:doctor's degree thesis[D].Changchun:Jilin University,2015.

[15] 王啟,杨添微,刘永震,等.基于随机森林算法的复杂碳酸盐岩岩性识别[J].工程地球物理学报,2020,17(5):550-558.
WANG Qi,YANG Tianwei,LIU Yongzhen,et al.Lithology identification of complex carbonate rocks based on random forest algorithm[J].Journal of Engineering Geophysics,2020,17(5):550-558.

[16] 康乾坤,路来君.随机森林算法在测井岩性分类中的应用[J].世界地质,2020,39(2):398-405.
KANG Qiankun, LU Laijun. Application of random forest algorithm in logging lithology classification[J].World Geology,2020,39(2):398-405.

[17] 王志宏,韩璐,戚磊.随机森林分类方法在储层岩性识别中的应用[J].辽宁工程技术大学学报:自然科学版,2015,34(9):1 083-1 088.
WANG Zhihong, HAN Lu, QI Lei. Application of random forest classification method in reservoir lithology identification[J].Journal of Liaoning University of Engineering Technology:Natural Science Edition,2015,34(9):1 083-1 088.

[18] 李克文,周广悦,路慎强,等.一种基于机器学习的有利区评价新方法[J].特种油气藏,2019,26(3):7-11.
LI Kewen,ZHOU Guangyue,LU Shenqiang,et al.A new method of favorable zone evaluation based on machine learning[J].Special Oil and Gas Reservoirs,2019,26(3):7-11.

[19] 隋微波,程思.基于卷积神经网络的砂岩数字岩心绝对渗透率计算方法[J].油气地质与采收率,2022,29(1):128-136.
SUI Weibo,CHENG Si.Calculation methods for absolute permeability of sandstone digital cores based on convolutional neural networks[J].Petroleum Geology and Recovery Efficiency,2022,29(1):128-136.

[20] 安鹏,曹丹平.基于深度学习的测井岩性识别方法研究与应用[J].地球物理学进展,2018,33(3):1 029-1 034.
AN Peng,CAO Danping.Research and application of logging lithology identification method based on deep learning[J].Progress in Geophysics,2018,33(3):1 029-1 034.

[21] 李建国,张卫东,刘冠男.深度学习在测井岩性识别中的应用[J].科技创新与应用,2015,4(14):21-22.
LI Jianguo,ZHANG Weidong,LIU Guannan.Application of deep learning in logging lithology identification[J].Innovation and Application of Science and Technology,2015,4(14):21-22.

[22] 陈钢花,梁莎莎,王军,等.卷积神经网络在岩性识别中的应用[J].测井技术,2019,43(2):129-134.
CHEN Ganghua,LIANG Shasha,WANG Jun,et al.Application of convolutional neural network in lithology identification[J].Well Logging Technology,2019,43(2):129-134.

[23] 蔡中超.自组织竞争神经网络在砂岩型铀矿测井数据解释中的应用研究[D].南昌:东华理工大学,2015.
CAI Zhongchao. The research and application of selforganized competitive neural network in interpretation of well logging data in sandstone type uranium deposits[D]. Nanchang: East China University of Technology,2015.

[24] 刘宇,乔木.基于聚类和XGboost算法的心脏病预测[J].计算机系统应用,2019,28(1):228-232.
LIU Yu, QIAO Mu.Heart disease prediction based on clustering and XGboost[J]. Computer System Application, 2019, 28(1): 228-232.

[25] PAO Lin Kuo, LIM Bee Yen, YI Chun Du, et al.Combination of XGBoost analysis and rule-based method for intrapartum cardiotocograph classification[J].Journal of Medical and Biological Engineering,2021,41:534-542.

[26] NOH Byungjoo, YOUM Changhong, GOH Eunkyoung, et al.XGBoost based machine learning approach to predict the risk of fall in older adults using gait outcomes[J].Scientific Reports,2021,11(1):1-9.

[27] FRIEDMAN J H.Greedy function approximation:a gradient boosting machine[J].The Annals of Statistics, 2001, 29(5): 1 189-1 232.

[28] LIN T Y,GOYAL P,GIRSHICK R,et al.Focal loss for dense object detection[J].IEEE Transactions on Pattern Analysis & Machine Intelligence,2017,99(1):2 980-2 988.

编辑 邹潋滟