https://doi.org/10.3799/dqkx.2022.181

基于随机森林算法的泥页岩岩相测井识别

王 民[1,2],杨金路[1,2],王 鑫[1,2],李进步[1,2],徐 亮[1,2],言 语[1,2]

1. 中国石油大学深层油气重点实验室,山东青岛 266580
2. 中国石油大学地球科学与技术学院,山东青岛 266580

摘 要: 泥页岩岩相识别是页岩油空间分布及勘探目标预测的一项重要工作,受地层非均质性及测井信息冗余的制约,基于测井响应方程的岩相识别十分困难.本文建立了一种基于随机森林算法的岩相识别模型,使用SHAP方法量化测井参数重要性.结果表明:随机森林算法可以很好地识别泥页岩岩相,其准确率高于支持向量机、KNN和XGBoost,并且对数据集中岩相类别不均衡的分类问题更加有效;对模型识别岩相最重要的前3项测井参数是自然电位、井径和声波时差;该模型可快速识别单井岩相,再根据总孔隙度、游离烃S_1、TOC等参数可确定有利岩相类型,进而确定研究区有利岩相分布,为后续"甜点"预测提供依据.

关键词: 随机森林;机器学习;测井;岩相识别;泥页岩.

中图分类号: P618.13 **文章编号:** 1000-2383(2023)01-130-13 **收稿日期:** 2022-02-11

Identification of Shale Lithofacies by Well Logs Based on Random Forest Algorithm

Wang Min[1,2], Yang Jinlu[1,2], Wang Xin[1,2], Li Jinbu[1,2], Xu Liang[1,2], Yan Yu[1,2]

1. *Key Laboratory of Deep Oil and Gas, China University of Petroleum, Qingdao* 266580, *China*
2. *School of Geosciences, China University of Petroleum, Qingdao* 266580, *China*

Abstract: Shale lithofacies identification is an important task in the spatial distribution of shale oil and exploration target prediction, but it is difficult to identify lithofacies based on logging response equations due to the formation heterogeneity and redundancy of logging information. In this paper, a lithofacies identification model based on random forest algorithm is proposed, which uses the SHAP method to quantify the contribution of logging parameters. The results show that the random forest algorithm can identify shale lithofacies well, and its accuracy is higher than support vector machine, k-nearest neighbors and XGBoost; *SP*, *CAL* and *AC* contribute the most to the model's identification of lithofacies. The model can quickly identify the lithofacies of a single well, and determine the favorable lithofacies by combining total porosity, free hydrocarbon S_1, TOC, etc., and then determine the distribution of favorable lithofacies in the whole area, providing a basis for subsequent "sweet spot" prediction.

Key words: random forest; machine learning; logging; lithofacies identification; shale.

基金项目: 国家自然科学基金项目(Nos.42072147,41922015).

作者简介: 王民(1981-),男,教授,博导,主要从事非常规油气地质研究.ORCID: 0000-0003-4611-2684. E-mail: wangm@upc.edu.cn

引用格式: 王民,杨金路,王鑫,李进步,徐亮,言语,2023.基于随机森林算法的泥页岩岩相测井识别.地球科学,48(1):130-142.

Citation: Wang Min, Yang Jinlu, Wang Xin, Li Jinbu, Xu Liang, Yan Yu, 2023. Identification of Shale Lithofacies by Well Logs Based on Random Forest Algorithm. *Earth Science*, 48(1): 130-142.

0　引言

北美页岩油勘探的成功掀起了全球范围内的页岩油勘探热潮(Hackley *et al.*, 2016).我国页岩油可采资源量为(74～372)×10^8 t(胡素云等,2020),展现了广阔的前景.目前页岩油的研究主要集中在页岩有机-无机化学特征、页岩储层定量表征、页岩油赋存机理等,如结合岩心手标本、岩石薄片、全岩X射线衍射(XRD)、主微量元素分析、镜质体反射率、总有机碳含量(TOC)、常规岩石热解、流体包裹体、电子探针、傅里叶红外光谱、拉曼光谱等研究页岩岩石矿物特征(Li *et al.*, 2019a)、页岩形成环境(张斌等, 2021)、页岩成岩作用(Lin *et al.*, 2021)、元素地球化学特征(李琪琪等, 2021);联合场发射扫描电镜(FE-SEM)、微/纳米CT、聚焦离子束-场发射扫描电镜等高分辨率成像实验和低温CO_2/N_2吸附、He注入、核磁共振(NMR)、高压压汞等流体注入法或X射线成像法等手段分析页岩全孔径信息(曾宏斌等, 2021; 王子萌等, 2022)、有机-无机孔形成演化(吴松涛等, 2015)、页岩储层分级(卢双舫等, 2018)、孔裂隙空间分布(Goral *et al.*, 2019)、页岩油富集主控因素(聂海宽等, 2016);通过离心/驱替的NMR、洗油前后的CO_2吸附-N_2吸附-高压压汞、高分辨率环境扫描电镜观察-EDS能谱检测、岩石热解(常规/分步)以及分子动力学模拟等方法揭示赋存孔径与状态(王民等, 2019)、吸附烃与游离烃含量(Li *et al.*, 2020)、页岩油可动量(李士祥等, 2021)、总产油量(Li *et al.*, 2019b).上述研究大多聚焦于单个页岩样品的研究,导致在识别、预测、推广、应用过程中难度较大,无法更快速且精准地明确全区有利层位及页岩油空间分布特征;而岩相作为反映岩石物理、化学性质的基本单元,具备可推广的条件,开展泥页岩岩相划分成为了页岩油勘探评价中的一项重要工作.

考虑页岩含油气性影响因素及与产出有关的可压裂性,普遍将有机质丰度、岩石构造、矿物组分作为页岩岩相划分依据(柳波等, 2018;刘忠宝等, 2019).一般通过岩心观察和薄片分析可以获得岩相信息,但受限于取心数量和费用,无法做到全井、全区评价.测井资料蕴含着丰富的岩石物理信息,可间接实现岩相分类,但由于储层复杂性及非均质性强,测井曲线间存在着大量的信息冗余,数据集类间分布不平衡,采用线性方程和经验统计公式无法较好地描述页岩岩相.对此,不少学者尝试使用机器学习算法解决岩相识别问题(Gifford and Agah, 2010; Wang *et al.*, 2014; Al-Mudhafar, 2015),既降低解释成本,又提高了分析效果.Bhattacharya *et al.*(2016)使用支持向量机成功识别了巴肯页岩岩相,正确率高于人工神经网络、自组织映射算法和多分辨率聚类算法.Al-Mudhafar *et al.*(2019)基于自然伽马、自然电位、中子孔隙度等11种测井资料,利用K-means聚类算法成功识别了伊拉克南部西库尔纳油田Mishrif层的5种碳酸盐岩岩相.但以上算法多为单一机器学习算法,准确率普遍不够高,有待进一步提升.随机森林算法通过集成决策树,具有更高的精度和泛化能力,在分类和回归两方面都有相当好的表现(Biau and Scornet, 2016),在岩性识别领域得到广泛运用(Wang *et al.*, 2020; Feng, 2021).

本文以松辽盆地某凹陷A段泥页岩岩相预测为目标,基于“有机质丰度＋岩石构造＋矿物组成”厘定出页岩岩相类型,通过随机森林算法构建岩相识别模型,建立岩相的快速识别方法,进而探索该方法在全井段及全区有利岩相分布预测中的应用,为后续“甜点”优选提供依据.

1　岩相类型划分

有机质含量的不同往往导致页岩物性、孔隙结构和含油量的差异(卢双舫等, 2018).多数学者认

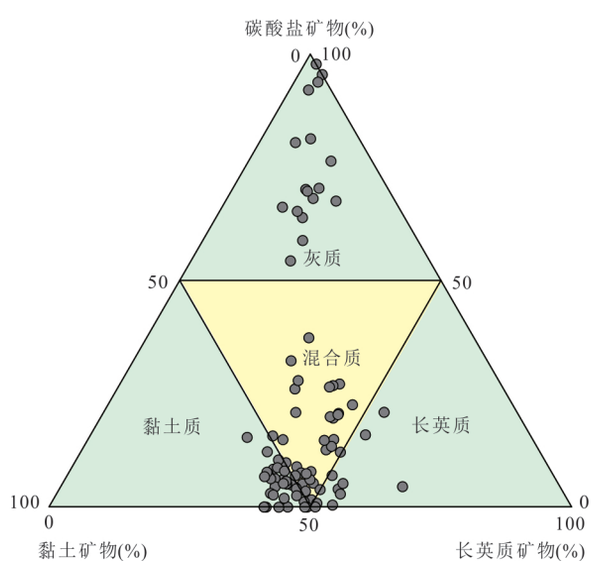
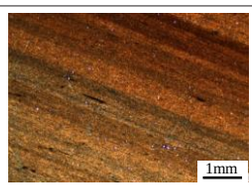
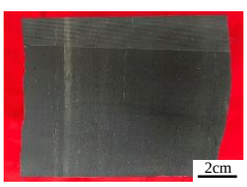
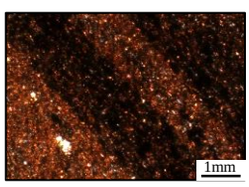
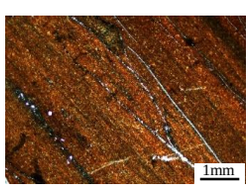
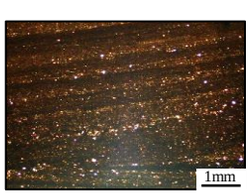
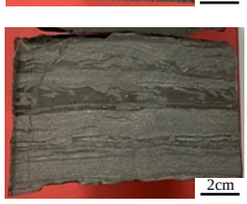
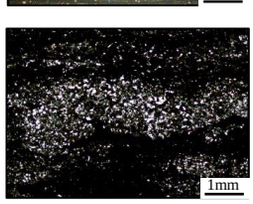
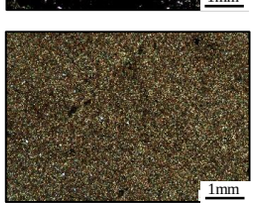


图1　松辽盆地某凹陷A段泥页岩矿物组成分布

Fig.1　Mineral composition distribution of shale in Formation A of a depression in Songliao Basin

表 1 松辽盆地某凹陷 A 段岩相类型发育特征

Table 1 Characteristics of lithofacies type development in Formation A of a depression in the Songliao Basin

序号	岩相类型	岩心观察	薄片观察	TOC(%)	岩石构造	矿物组成(%)
1	富有机质纹层状黏土质页岩			＞2.0	页理发育,黏土纹层夹长英质/碳酸盐纹层	黏土＞50,长英质＜50,碳酸盐＜50
2	富有机质层状黏土质页岩			＞2.0	页理发育,黏土层夹长英质/碳酸盐层	黏土＞50,长英质＜50,碳酸盐＜50
3	中等有机质纹层状黏土质页岩			1.0～2.0	页理发育,黏土纹层夹长英质/碳酸盐纹层	黏土＞50,长英质＜50,碳酸盐＜50
4	富有机质纹层状混合质页岩			＞2.0	页理发育,黏土质、长英质、碳酸盐纹层互层	黏土＜50,长英质＜50,碳酸盐＜50
5	富有机质层状混合质页岩			＞2.0	页理发育,黏土质、长英质、碳酸盐交互层	黏土＜50,长英质＜50,碳酸盐＜50
6	低有机质块状灰质泥岩			＜1.0	无页理,碳酸盐矿物颗粒均匀分布	黏土＜50,长英质＜50,碳酸盐＞50

为当页岩的 TOC 大于 2%(富有机质)时具备工业开采价值(Su *et al.*, 2019; Atchley *et al.*, 2021),同时参考前人对松辽盆地某凹陷 A 段页岩 TOC 分级的界定依据(柳波等, 2018;卢双舫等, 2018),本文将 TOC 大于 2.0% 的页岩命名为富有机质页岩,TOC 介于 1.0%～2.0% 为中等有机质,TOC 小于 1.0% 时为低有机质.岩石构造作为划分岩相类型的重要因素,不仅可以直接反映页岩的形成环境,还因不同构造形式提供的储集空间存在差异,进而对页岩油气储集或渗流产生影响(Atchley *et al.*, 2021).通常利用岩心手标本及镜下薄片观察,依据单个层理发育厚度划分出纹层状、层状、块状构造,其中单层厚度小于 1 mm 时对应纹层状,单层厚度介于 1～10 mm 为层状,层理不发育时对应块状构造(Su *et al.*, 2019;张斌等, 2021).页岩中矿物组分丰富,通常将主要发育的黏土矿物、碳酸盐矿物和长英质矿物(长石＋石英)作为三端元并考虑将矿物含量大于 50% 的部分和小于 50% 的部分作为界限,区分出 4 种岩石类型:黏土质、灰质、长英质、混合质页(泥)岩(刘忠宝等, 2019),矿物组成分布特征如图 1 所示.

通过大量自测和单井连续取心观察统计,对松辽盆地某凹陷 A 段页岩层系共厘定出 12 种岩相类型,发育 6 种主要岩相(占比超过 78%),即

富有机质纹层状黏土质页岩、富有机质层状黏土质页岩、中等有机质纹层状黏土质页岩、富有机质纹层状混合质页岩、富有机质层状混合质页岩、低有机质块状灰质泥岩(表 1).

2　随机森林算法

随机森林(Random Forest,RF)算法是由 Breiman(2001)提出,采用决策树构成的一种集成学习算法.RF 算法利用多棵决策树对样本进行训练,并以投票的方式汇总预测结果,从而提高模型的预测精度;其实现简单,对多特征数据及部分特征缺失数据表现优异,故在机器学习领域有广泛应用.

2.1　决策树算法

决策树(Decision Tree,DT)是机器学习中一种用于分类和回归的算法,它的理论结构是一个树状图,由根节点、内部节点和叶子节点三部分构成.树状图中每个非叶子节点代表一种决策,每个叶子节点代表一种类别,每一棵决策树只有一个根节点,数据的训练和预测均从根节点开始逐级向下延伸.对于新输入的一个数据,从根节点开始对数据进行决策,每一个非叶子节点都会按照当前节点的划分特征将数据划分到下一级的节点处理,直到该数据被划分到最下层的一个叶子节点,此叶子节点所代表的类别即该输入数据的类别.

CART 算法是常用的决策树生成算法之一.CART决策树从上而下地生成,即首先从根节点的特征集合中取出某一特征,根据此特征将数据集划分为2个子节点,每个子节点中包含一定数量的样本,之后分别对每个子节点进行同样的处理,直到达到决策树不再生长的条件完成决策树的构建.决策树构建的关键是如何选择非叶子节点的划分特征,使得决策树尽可能准确地将同一类样本划分到同一个节点中.当某一节点中的样本都属于同一类别时,信息纯度最高,分类效果最好.为了选出最佳划分特征,CART决策树使用Gini指数作为判断准则.Gini指数定义如下:

$$\mathrm{Gini}(\boldsymbol{S})=1-\sum_{i=1}^{c}P_i^{\,2}\ , \tag{1}$$

其中 $\boldsymbol{S}$ 代表样本集,c 为 $\boldsymbol{S}$ 中含有的样本类别,P_i 表示第 i 个样本类别出现的概率.

Gini指数为0表示样本集中的样本属于同一类别.当按照某特征属性f进行划分后,样本集$\boldsymbol{S}$被划分为$\boldsymbol{S}_1$和$\boldsymbol{S}_2$两类,则此时子样本集的Gini指数为:

$$\mathrm{Gini}_f(\boldsymbol{S})=\frac{n_{s_1}}{n}\mathrm{Gini}(\boldsymbol{S}_1)+\frac{n_{s_2}}{n}\mathrm{Gini}(\boldsymbol{S}_2)\ , \tag{2}$$

其中 $\boldsymbol{S}$ 代表样本集,n_{S_1}、n_{S_2} 为数据集 $\boldsymbol{S}_1$ 和 $\boldsymbol{S}_2$ 的样本数量,n 为总样本数量.

若划分后可使$\mathrm{Gini}_f(\boldsymbol{S})$取得最小值,$f$即可视为最佳划分特征.在CART算法构建决策树的过程中,从根节点开始,每次迭代都会寻找当前节点的最佳划分特征f生成新的子节点,反复进行,直到到达决策树停止生长的条件完成整棵决策树的构建.

2.2　随机森林算法

因为单棵决策树在构建时极易出现过拟合的情况,而且性能具有一定的局限性,所以Breiman引入集成学习的思想对其改进.随机森林算法是一种组合分类器算法,采用CART决策树作为基分类器,其通过随机有放回的抽取样本集和随机无放回的抽取特征集构建决策树,最终生成的多棵决策树组成随机森林模型,并通过基分类器投票的方式决定模型最终的预测结果.当用于分类时,通过随机抽取数据集中的样本及样本特征构建决策树,重复多次,生成的决策树间互不相关,并统计所有决策树的结果作为最终结果.预测新样本时,统计森林中的每个决策树的分类结果,选择最多的类别作为新样本预测结果.

随机森林算法主要步骤及流程如图2所示.步

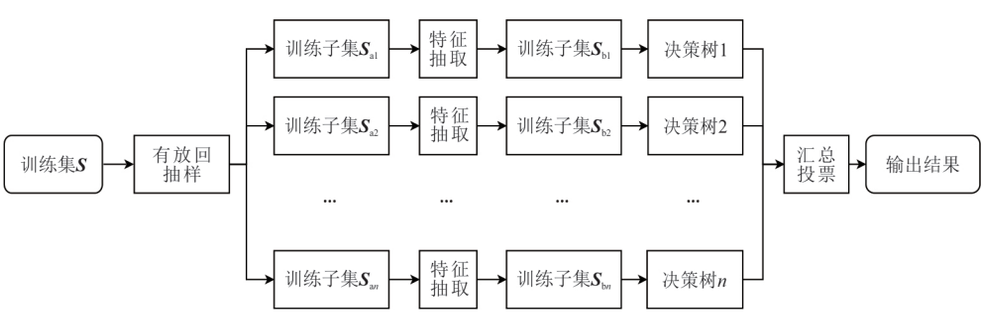


图2　随机森林算法工作流程图

Fig.2　Flow chart of the random forest algorithm

骤1:确定随机森林训练参数,如训练样本集S,特征属性集F,决策树个数n,随机特征个数m.步骤2:从训练样本集S中有放回的抽取与训练样本集同样个数的样本,形成训练样本集S_a.步骤3:从特征属性集F中无放回地抽取出m个特征,在训练样本集S_a中只保留该m个特征的数据形成训练样本集S_b.步骤4:利用训练样本S_b训练一棵CART决策树.步骤5:重复n次步骤2至步骤4,得到n棵决策树.步骤6:通过汇总投票得到最终结果并输出.

随机森林作为一种组合分类器,和单分类器相比,其算法效果稳定,泛化能力更强,且不易出现过拟合.

3 实验结果与分析

3.1 数据预处理

3.1.1 数据准备 本研究采用的岩心数据和测井资料来自松辽盆地某凹陷两口井,目的层是A段,共获取347个样本数据.样本数据按照7:3随机划分为训练集和测试集.使用1、2、3、4、5和6的岩相识别标签来分别对应富有机质纹层状混合质页岩、富有机质纹层状黏土质页岩、富有机质层状黏土质页岩、中等有机质纹层状黏土质页岩、富有机质层状混合质页岩、低有机质块状灰质泥岩6种主要岩相,使用岩相标签10表示其他岩相.表2显示了不同岩相的样本分布.

选取以下8种常规测井参数作为样本属性值,分别为自然伽马(GR)、自然电位(SP)、井径测井(CAL)、深侧向电阻率(LLD)、浅侧向电阻率(LLS)、声波时差(AC)、补偿中子(CNL)和密度(DEN).每一样本数据均由9维向量组成,包括8维不同参数值及1维对应岩相标签.

表2 数据集中不同岩相样本分布

Table 2 The distribution of the different lithofacies samples in the dataset

岩相标签	岩相	样本数	占比(%)
1	富有机质纹层状混合质页岩	46	13.26
2	富有机质纹层状黏土质页岩	78	22.48
3	富有机质层状黏土质页岩	30	8.65
4	中等有机质纹层状黏土质页岩	26	7.49
5	富有机质层状混合质页岩	48	13.83
6	低有机质块状灰质泥岩	76	21.90
10	其他	43	12.39
	统计	347	

3.1.2 数据归一化 数据归一化处理是机器学习分类的一项基础工作.由于各类测井曲线的量纲不同且差异较大,如果直接将测井数据作为输入训练模型,会影响岩相分类的结果,为了消除量纲对分类效果的影响,需要对数据进行归一化.通过最大最小归一化函数将输入曲线值映射到[0,1],即该组曲线值中最大值为1,最小值为0.定义如下:

$$x^*=\frac{x-x_{\min}}{x_{\max}-x_{\min}}, \tag{3}$$

其中:x^*代表归一化后数据,$x_{\min}$为样本数据最小值,$x_{\max}$为样本数据最大值.归一化后,所有测井数据值均在[0,1]区间内.

3.1.3 探索性数据分析 通过绘制不同岩相与测井参数的相关矩阵图进行探索性数据分析.图3为不同岩相与测井参数的相关矩阵图,横轴和纵轴均为8项测井参数,右上区为测井参数的两两关系散点图,左下区为对应的散点图以等值线图显示,中间斜对角为对应横轴的核密度估计图,不同颜色代表不同岩相.

从图3右上区可知,多数测井参数间相关性不明显,少数测井参数间具有相关性,如浅侧向电阻率与深侧向电阻率呈正相关、声波时差与补偿中子呈正相关、声波时差与密度呈负相关、补偿中子与密度呈负相关.从图3左下区可知,多数岩相间的测井参数重叠,并无明显分界线,说明模型的分类难度大.

3.2 实验过程

3.2.1 模型参数寻优 为了得到最佳的模型分类性能,一般使用网格搜索和k-折交叉验证法为模型选取合适的参数.在基于随机森林算法建立的岩相识别模型中,对模型性能影响较大的参数有:迭代次数,即决策树的个数,若迭代次数太大则容易使得模型过拟合,若迭代次数太小则容易使得模型欠拟合,二者均会导致模型的泛化能力下降;最大树深度,该参数用于控制模型的复杂度,最大树深度太小会使得模型欠拟合,降低模型的准确率,最大树深度太大则会使得模型过拟合,降低模型的泛化能力;内部节点再划分所需最小样本数和叶子结点最小样本数,当节点内样本数量不满足条件时,节点停止分裂.随机森林算法的重要参数、搜索范围及步长如表3所示.

由于样本数量较少,故使用5-折交叉验证对模型的参数进行调优.5-折交叉验证会将训练集随机分成5份大小相同的子集,将其中4份子集用作训

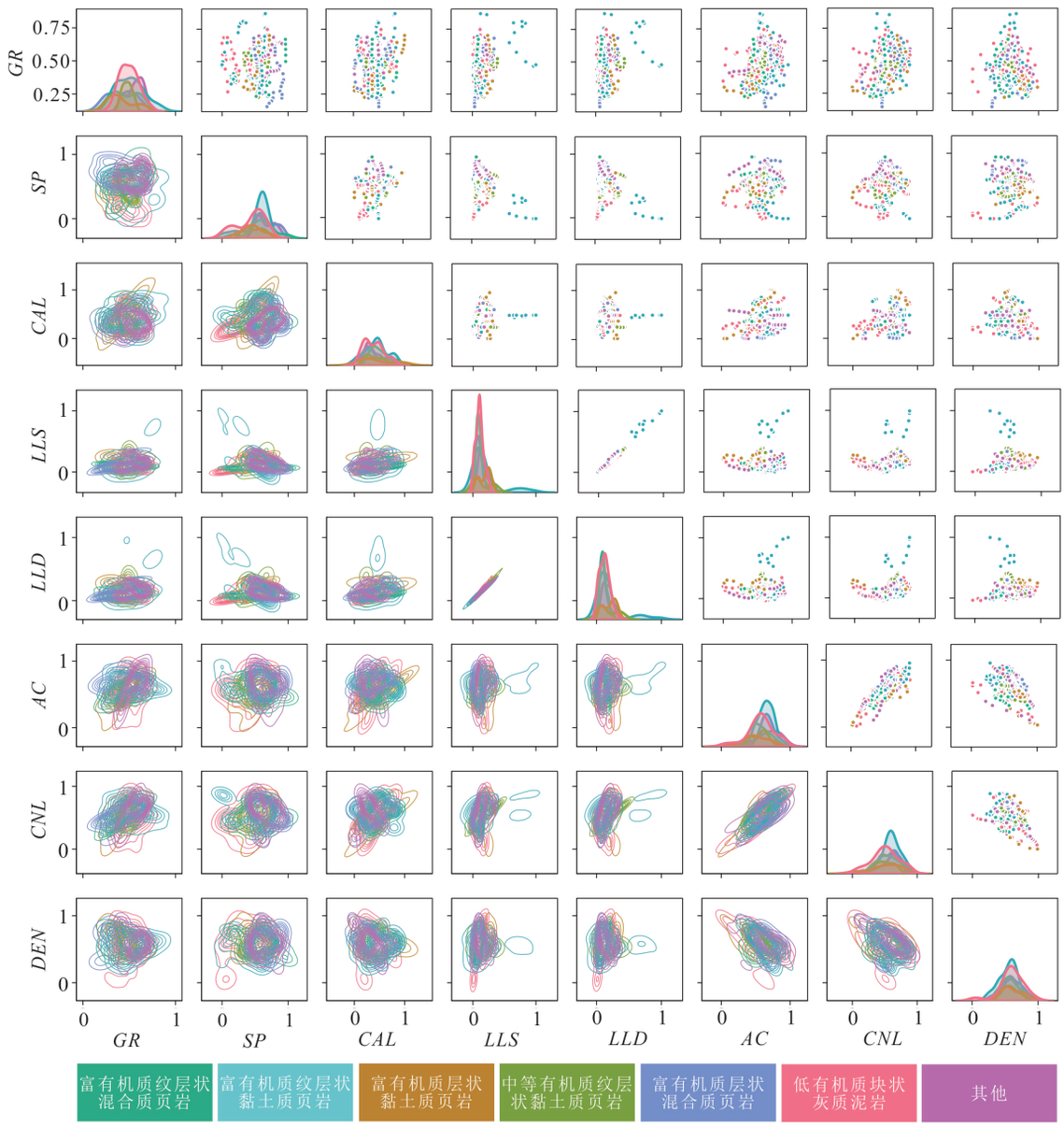


图 3　不同岩相与测井参数相关矩阵图

Fig.3　Correlation matrix of different lithofacies and logging parameters

表 3　随机森林算法参数调优

Table 3　Parameters tuning for random forest

参数	搜索范围	步长	最优值
迭代次数	100～500	2	120
最大树深度	1～15	1	10
内部节点再划分所需最小样本数	1～50	1	4
叶子结点最小样本数	1～20	1	1

练模型,剩余 1 份子集对该模型进行验证,循环 5 次.模型的交叉验证评分由 5 份验证评分平均得到.假设模型有多组参数可供调参,一组参数可计算一个交叉验证评分,最后选择使得交叉验证评分最大的那一组参数.完成参数调整后,确定随机森林算法最优参数组合(表 3).

将随机森林算法应用于页岩岩相识别时,需注意数据质量,数据集中缺失值、异常值、量纲不同均会影响模型的质量.若样本数量不充足,建议按照 7∶3 划分训练集和测试集,并使用 k-折交叉验证和训练集对模型进行训练和调参,测试集则用来评估模型.

3.2.2　评价标准　在本研究中,采用查准率、查全率和 F1-score 评价分类模型的优劣.查准率定义为预测为正例中的真正例的比例.查全率定义为所有

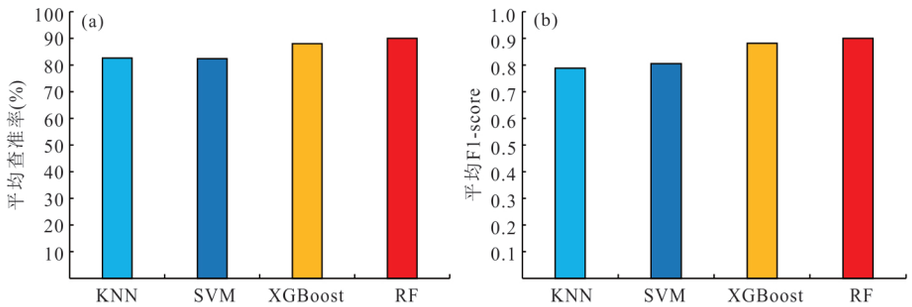


图 4 不同模型岩相识别结果对比

Fig.4 Comparison of predicted results from different lithofacies identification models

表 4 不同模型的岩相识别 F1-score

Table 4 F1-score of different lithofacies identification models

岩相标签	岩相	KNN	SVM	XGBoost	RF
1	富有机质纹层状混合质页岩	0.88	0.85	0.91	0.88
2	富有机质纹层状黏土质页岩	0.91	0.80	0.93	0.95
3	富有机质层状黏土质页岩	0.78	0.84	0.82	0.91
4	中等有机质纹层状黏土质页岩	0.63	0.82	0.78	0.82
5	富有机质层状混合质页岩	0.86	0.70	1.00	1.00
6	低有机质块状灰质泥岩	0.79	0.82	0.87	0.89
10	其他	0.67	0.81	0.86	0.86

真正例中被预测为正例的比例.F1-score为查准率和查全率的调和平均数.以上评价指标公式如下:

$$\text{查准率}=\frac{\text{真正例}}{\text{真正例}+\text{假正例}}, \tag{4}$$

$$\text{查全率}=\frac{\text{真正例}}{\text{真正例}+\text{假负例}}, \tag{5}$$

$$\text{F1}-\text{score}=2\times\frac{\text{查准率}\times\text{查全率}}{\text{查准率}+\text{查全率}}, \tag{6}$$

其中:真正例指预测为正的正样本,假正例指预测为正的负样本,真负例指预测为负的负样本,假负例指预测为负的正样本.

3.3 结果分析

3.3.1 单点预测结果分析 完成模型的参数调优后,采用测试集对建立的岩相识别模型预测效果进行验证.为评价随机森林算法的预测性能,将其与支持向量机(SVM)、K-近邻算法(KNN)、XGBoost比较.不同模型预测结果对比如图4所示.由图4a、4b可知,KNN的岩相识别平均F1-score最低,为0.79;采用SVM的识别平均F1-score为0.81,高于KNN,但KNN的平均查准率略高于SVM;随机森林算法的平均F1-score为0.90,平均查准率为90%,均高于XGBoost算法.由上可知,随机森林算法的岩相识别效果强于KNN、SVM和XGBoost.

由表4及图5可知,不同模型对各岩相的识别能力不同.RF的平均F1-score最高,达到了0.90,其中对富有机质纹层状黏土质页岩、富有机质层状黏土质页岩和富有机质层状混合质页岩的F1-score均在0.91以上,并主要将低有机质块状灰质泥岩错分为富有机质纹层状混合质页岩、富有机质纹层状黏土质页岩及富有机质层状黏土质页岩;XGBoost的平均F1-score仅次于RF,其对富有机质纹层状混合质页岩、富有机质纹层状黏土质页岩和富有机质层状混合质页岩的F1-score均在0.91以上,其主要是将富有机质纹层状黏土质页岩错分为富有机质纹层状混合质页岩和低有机质块状灰质泥岩,以及错分发生在富有机质层状黏土质页岩和低有机质块状灰质泥岩之间;SVM对富有机质纹层状混合质页岩的F1-score为0.85,其余岩相的F1-score均较低;KNN对富有机质纹层状黏土质页岩有较高的F1-score,达到了0.91,其余岩相的F1-score均不高于0.88.

当各个岩相样本数量分布不均衡时,模型的分类效果将会受到影响.由表2可知,富有机质层状黏土质页岩和中等有机质纹层状黏土质页岩的样本数量仅占总体的8.65%和7.49%.如表4所示,针对富有机质层状黏土质页岩,KNN、SVM、XGBoost

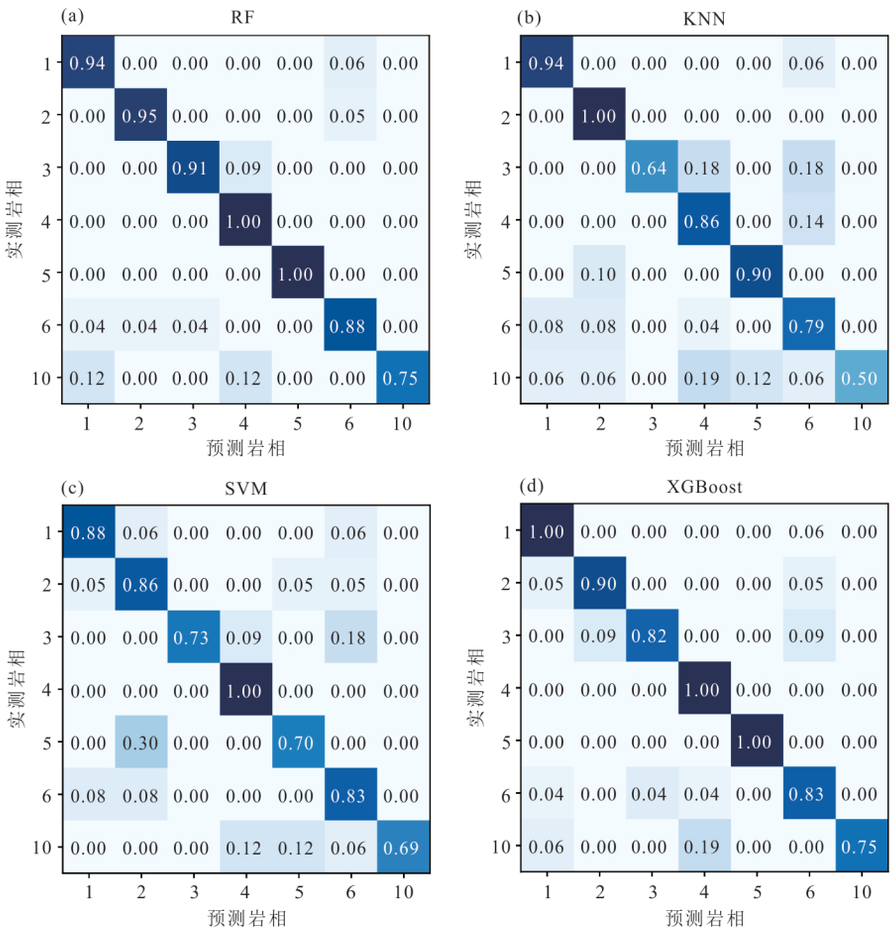


图 5　不同岩相识别模型混淆矩阵

Fig.5　Comparison of confusion matrices for different lithofacies identification models

混淆矩阵中行代表真实岩相,每一行中的数据表示真实岩相被预测的岩相类别分布,如 a 图第二行中的 0.95,表示 95% 的实测岩相为 2 的样品被预测为类别 2;列代表预测岩相,每一列中的数据表示实测岩相被预测为该类的比例分布

和 RF 的 F1-score 分别为 0.78、0.84、0.82 和 0.91;针对中等有机质纹层状黏土质页岩,KNN 的 F1-score 为 0.63,XGBoost 的 F1-score 为 0.78,SVM 和 RF 的 F1-score 均为 0.82. 由上可知,随机森林算法受岩相类别分布不均衡数据的影响最小. 随机森林算法比 XGBoost 的得分更高,这可能是因为测井数据中存在的噪声,XGBoost 这类 boosting 算法容易受到噪声影响,而随机森林算法受影响较少,因此随机森林算法表现更好.

3.3.2　测井参数重要性分析　在完成随机森林分类模型训练后,使用 Shapley additive explanations (SHAP)模型量化不同测井参数对模型识别岩相的重要性. SHAP 模型是由 Lundberg and Lee(2017)提出的,使用博弈论中的 Shapley 值来计算每个特征对预测的贡献,可以用于解释各种分类和回归模型中不同参数对模型预测结果的重要性. 与随机森林中基于基尼不纯度计算的参数重要性相比,SHAP 模型采用边际贡献计算参数重要性更加合理,它还可以提供更加丰富的信息,如不同测井参数对各岩相的重要性. SHAP 模型计算公式如下:

$$\phi_j=\sum_{S\subseteq\{x_1,\cdots,x_p\}\setminus\{x_j\}}\frac{|S|!(p-|S|-1)!}{p!}\left(f_x(S\cup\{x_j\})-f_x(S)\right),\tag{7}$$

其中:ϕ_j 是特征 j 的贡献度,S 是输入特征的可能的子集,$\{x_1,\cdots,x_p\}$ 是所有输入特征的集合,p 是输入特征的个数,$\{x_1,\cdots,x_p\}\setminus\{x_j\}$ 为不包括 $\{x_j\}$ 的所有输入特征可能的集合,$f_x(S)$ 为特征子集 S 的预测,$\frac{|S|!(p-|S|-1)!}{p!}$ 是子集 S 的特征组合情况占比.

图 6 为测井参数重要性汇总图,它展示了不同测井参数的重要程度. 该图横轴为 SHAP 平均绝对值,可划分为对不同岩相的 SHAP 值,使用不同颜

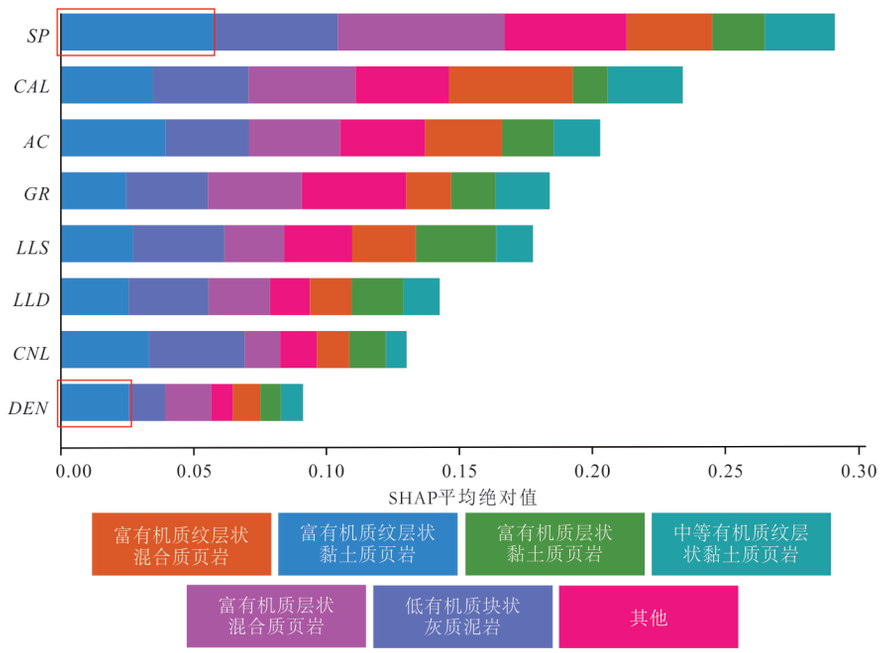


图 6　岩相识别中测井参数重要性分析

Fig.6　Analysis of the importance of logging parameters in lithofacies identification

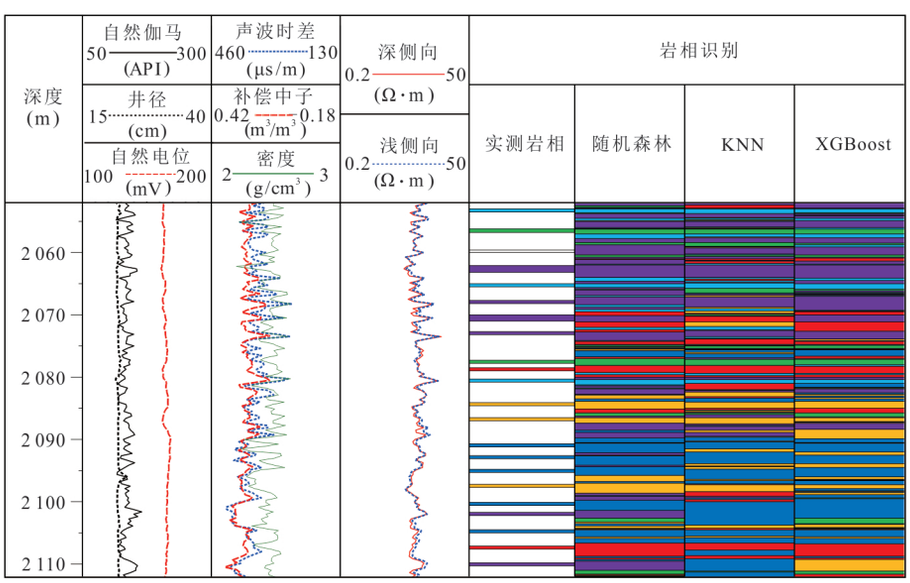


图 7　X 井不同页岩岩相识别模型预测结果对比

Fig.7　Comparison of predicted results from different lithofacies identification models for Well X

色代表对不同岩相的重要性，纵轴为各测井参数按重要性排序，从下至上测井参数重要性增大．由图 6 可知，自然电位是划分不同岩相的最关键因素，井径、声波时差、自然伽马、浅侧向电阻率、深侧向电阻率、补偿中子的重要性逐渐降低，密度对岩相分类影响最小．对于每类岩相，测井参数重要性各有不同，如划分富有机质纹层状黏土质页岩，自然电位是影响最大的因素，密度是影响最小的因素．

3.3.3　单井连续预测评价　单井测井岩相识别结果表明，随机森林、KNN、XGBoost 均能较好地识别页岩岩相，但随机森林算法预测的岩相的准确度最高（图 7），说明随机森林算法在泥页岩岩相识别及预测中具有更高的可靠性，不仅可弥补因无法连续取心而难以获取完整岩相分布特征的空白，还能大大提高岩相识别效率．X 井页岩岩相预测结果显示，垂向上岩相具有类型变化快、沉积厚度薄、有机质

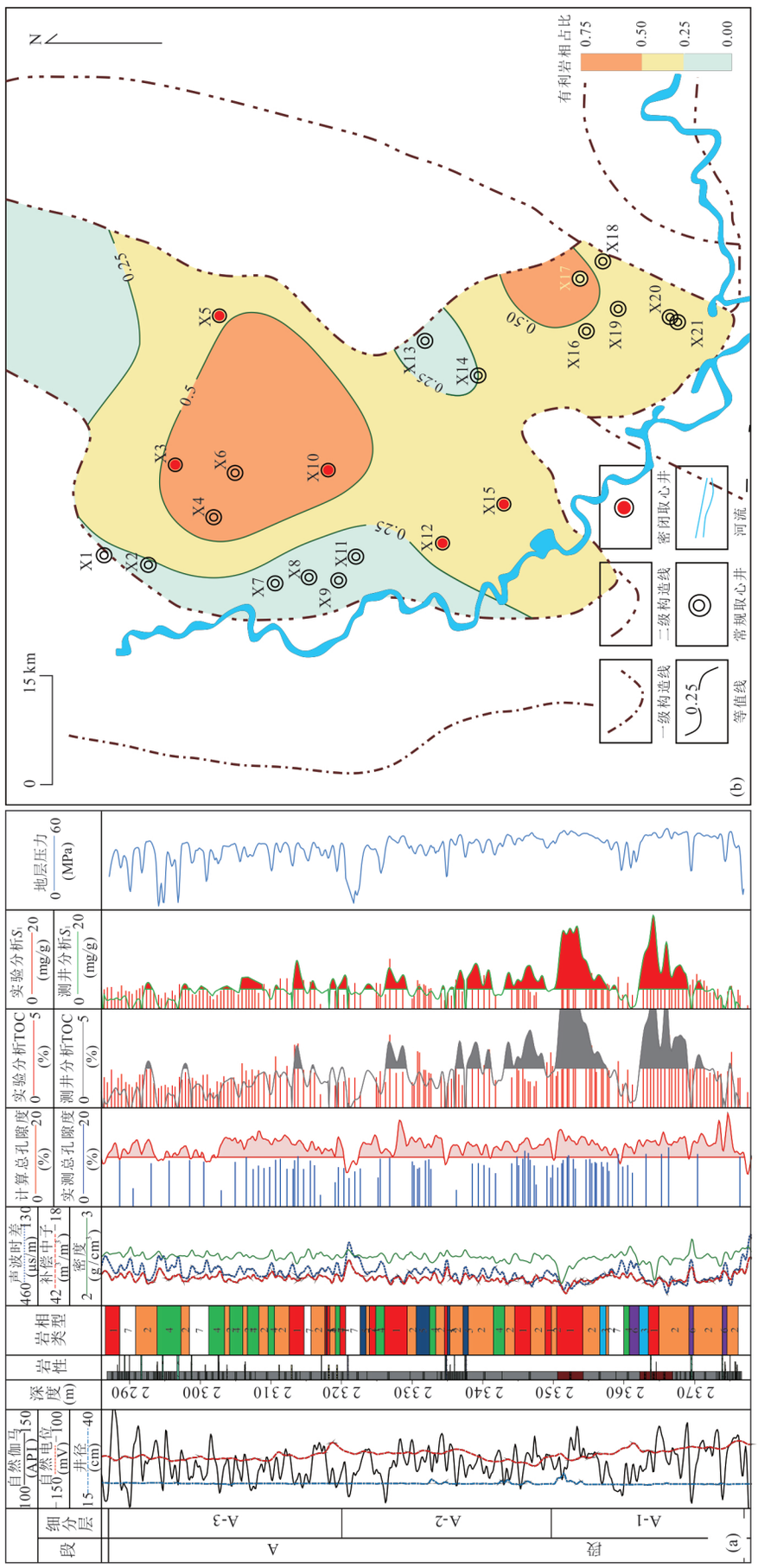


图8　松辽盆地某凹陷A段有利岩相分布

Fig.8　Distribution of favourable lithofacies in Formation A of a depression in the Songliao Basin

含量变化快等特点,表现出强非均质性,主要发育富有机质层状混合质页岩、低有机质块状灰质泥岩.

3.3.4 平面有利岩相分布评价 基于以上研究,将基于随机森林算法的岩相预测模型在研究区21口井进行推广应用.以关键井X6为例,通过对比岩相预测结果与总孔隙度、游离烃S_1、TOC可知,不同岩相对应的储集性、含油性差异明显,其中总孔隙度>9%、游离烃S_1>4 mg/g、TOC>2%对应的岩相主要为富有机质纹层状黏土质页岩和富有机质纹层状混合质页岩(图8a).通过统计研究区单井对应的目的层段有利岩相类型(富有机质纹层状黏土质页岩、富有机质纹层状混合质页岩)发育占比,绘制有利岩相等值线图(图8b).有利岩相主要分布于两个区块,其中一个区块分布于研究区的西北部且分布面积较广(覆盖X3、X4、X6、X10井),有利岩相类型发育占比超过50%,随着向四周不断延伸,有利岩相占比逐渐降低;另一区块分布于研究区的西南部,分布面积较小且有利岩相类型发育占比超过50%,靠近二级构造线(覆盖X17井).该两个区块可作为有利"甜点"区重点勘探开发.

4 结论

随机森林算法能够很好的拟合测井数据与岩相间的非线性映射关系,在泥页岩岩相识别中有很好的应用前景,结论如下:

(1)随机森林算法可以很好地识别泥页岩岩相,其准确率要高于支持向量机、KNN和XGBoost,并且对数据集中岩相类别不均衡的分类问题更加有效.

(2)运用SHAP模型量化不同测井参数对模型识别岩相的重要性,重要性由高到低为自然电位、井径、声波时差、自然伽马、浅侧向电阻率、深侧向电阻率、补偿中子和密度.

(3)运用随机森林算法岩相识别模型开展研究区岩相测井预测,进一步结合总孔隙度、游离烃S_1、TOC等确定出有利岩相类型,进而确定研究区有利岩相分布,为有利区"甜点"预测提供依据.

References

Al-Mudhafar, W. J., 2015. Integrating Component Analysis & Classification Techniques for Comparative Prediction of Continuous & Discrete Lithofacies Distributions. In: Offshore Technology Conference. SPE, Houston.

Al-Mudhafar, W. J., Al Lawe, E. M., Noshi, C. I., 2019. Clustering Analysis for Improved Characterization of Carbonate Reservoirs in a Southern Iraqi Oil Field. In: Offshore Technology Conference. SPE, Houston.

Atchley, S. C., Crass, B. T., Prince, K. C., 2021. The Prediction of Organic-Rich Reservoir Facies within the Late Pennsylvanian Cline Shale (Also Known as Wolfcamp D), Midland Basin, Texas. *AAPG Bulletin*, 105(1): 29−52. https://doi.org/10.1306/07272020010

Bhattacharya, S., Carr, T. R., Pal, M., 2016. Comparison of Supervised and Unsupervised Approaches for Mudstone Lithofacies Classification: Case Studies from the Bakken and Mahantango-Marcellus Shale, USA. *Journal of Natural Gas Science and Engineering*, 33: 1119−1133. https://doi.org/10.1016/j.jngse.2016.04.055

Biau, G., Scornet, E., 2016. A Random Forest Guided Tour. *Test*, 25(2): 197−227. https://doi.org/10.1007/s11749-016-0481-7

Breiman, L., 2001. Random Forests. *Machine Learning*, 45(1): 5−32. https://doi.org/10.1023/A:1010933404324

Feng, R. H., 2021. Improving Uncertainty Analysis in Well Log Classification by Machine Learning with a Scaling Algorithm. *Journal of Petroleum Science and Engineering*, 196: 107995. https://doi. org/10.1016/j. petrol.2020.107995

Gifford, C. M., Agah, A., 2010. Collaborative Multi-Agent Rock Facies Classification from Wireline Well Log Data. *Engineering Applications of Artificial Intelligence*, 23(7): 1158−1172. https://doi. org/10.1016/j. engappai.2010.02.004

Goral, J., Walton, I., Andrew, M., et al., 2019. Pore System Characterization of Organic-Rich Shales Using Nanoscale-Resolution 3D Imaging. *Fuel*, 258: 116049. https://doi.org/10.1016/j.fuel.2019.116049

Hackley, P. C., Fishman, N., Wu, T., et al., 2016. Organic Petrology and Geochemistry of Mudrocks from the Lacustrine Lucaogou Formation, Santanghu Basin, Northwest China: Application to Lake Basin Evolution. *International Journal of Coal Geology*, 168: 20−34. https://doi.org/10.1016/j.coal.2016.05.011

Hu, S.Y., Zhao, W.Z., Hou, L.H., et al., 2020. Development Potential and Technical Strategy of Continental Shale Oil in China. *Petroleum Exploration and Development*, 47(4): 819−828 (in Chinese with English abstract).

Li, B. Y., Pang, X. Q., Dong, Y. X., et al., 2019a. Lithofacies and Pore Characterization in an Argillaceous-Siliceous-Calcareous Shale System: A Case Study of

the Shahejie Formation in Nanpu Sag, Bohai Bay Basin, China. *Journal of Petroleum Science and Engineering*, 173: 804－819. https://doi.org/10.1016/j.petrol.2018.10.086

Li, J. B., Wang, M., Chen, Z. H., et al., 2019b. Evaluating the Total Oil Yield Using a Single Routine Rock-Eval Experiment on As-Received Shales. *Journal of Analytical and Applied Pyrolysis*, 144: 104707. https://doi.org/10.1016/j.jaap.2019.104707

Li, J. B., Jiang, C. Q., Wang, M., et al., 2020. Adsorbed and Free Hydrocarbons in Unconventional Shale Reservoir: A New Insight from NMR T1-T2 Maps. *Marine and Petroleum Geology*, 116: 104311. https://doi.org/10.1016/j.marpetgeo.2020.104311

Li, Q.Q., Lan, B.F., Li, G.Q., et al., 2021. Element Geochemical Characteristics and Their Geological Significance of Wufeng-Longmaxi Formation Shales in North Margin of the Central Guizhou Uplift. *Earth Science*, 46(9): 3172－3188 (in Chinese with English abstract).

Li, S.X., Zhou, X.P., Guo, Q.H., et al., 2021. Research on Evaluation Method of Movable Hydrocarbon Resources of Shale Oil in the Chang 7_3 Sub-Member in the Ordos Basin. *Natural Gas Geoscience*, 32(12): 1771－1784 (in Chinese with English abstract).

Lin, M. R., Xi, K. L., Cao, Y. C., et al., 2021. Petrographic Features and Diagenetic Alteration in the Shale Strata of the Permian Lucaogou Formation, Jimusar Sag, Junggar Basin. *Journal of Petroleum Science and Engineering*, 203: 108684. https://doi. org/10.1016/j. petrol.2021.108684

Liu, B., Shi, J.X., Fu, X.F., et al., 2018. Petrological Characteristics and Shale Oil Enrichment of Lacustrine Fine-Grained Sedimentary System: A Case Study of Organic-Rich Shale in First Member of Cretaceous Qingshankou Formation in Gulong Sag, Songliao Basin, NE China. *Petroleum Exploration and Development*, 45(5): 828－838 (in Chinese with English abstract).

Liu, Z.B., Liu, G.X., Hu, Z.Q., et al., 2019. Lithofacies Types and Assemblage Features of Continental Shale Strata and Their Significance for Shale Gas Exploration: A Case Study of the Middle and Lower Jurassic Strata in the Sichuan Basin. *Natural Gas Industry*, 39(12): 10－21 (in Chinese with English abstract).

Lu, S.F., Li, J.Q., Zhang, P.F., et al., 2018. Classification of Microscopic Pore-Throats and the Grading Evaluation on Shale Oil Reservoirs. *Petroleum Exploration and Development*, 45(3): 436－444 (in Chinese with English abstract).

Lundberg, S. M., Lee, S. I., 2017. A Unified Approach to Interpreting Model Predictions. 31st Conference on Neural Information Processing Systems, Long Beach.

Nie, H.K., Zhang, P.X., Bian, R.K., et al., 2016. Oil Accumulation Characteristics of China Continental Shale. *Earth Science Frontiers*, 23(2): 55－62 (in Chinese with English abstract).

Su, S. Y., Jiang, Z. X., Shan, X. L., et al., 2019. Effect of Lithofacies on Shale Reservoir and Hydrocarbon Bearing Capacity in the Shahejie Formation, Zhanhua Sag, Eastern China. *Journal of Petroleum Science and Engineering*, 174: 1303－1308. https://doi. org/10.1016/j. petrol.2018.11.048

Wang, G. C., Carr, T. R., Ju, Y. W., et al., 2014. Identifying Organic-Rich Marcellus Shale Lithofacies by Support Vector Machine Classifier in the Appalachian Basin. *Computers & Geosciences*, 64: 52－60. https://doi.org/10.1016/j.cageo.2013.12.002

Wang, M., Ma, R., Li, J. B., et al., 2019. Occurrence Mechanism of Lacustrine Shale Oil in the Paleogene Shahejie Formation of Jiyang Depression, Bohai Bay Basin, China. *Petroleum Exploration and Development*, 46(4): 789－802 (in Chinese with English abstract).

Wang, P., Chen, X. H., Wang, B. F., et al., 2020. An Improved Method for Lithology Identification Based on a Hidden Markov Model and Random Forests. *Geophysics*, 85(6): IM27－IM36. https://doi. org/10.1190/geo2020-0108.1

Wang, Z.M., Jiang, Y.Q., Fu, Y.H., et al., 2022. Characterization of Pore Structure and Heterogeneity of Shale Reservoir from Wufeng Formation-Sublayers Long-1_1 in Western Chongqing Based on Nuclear Magnetic Resonance. *Earth Science*, 47(2): 490－504 (in Chinese with English abstract).

Wu, S.T., Zhu, R.K., Cui, J.G., et al., 2015. Characteristics of Lacustrine Shale Porosity Evolution, Triassic Chang 7 Member, Ordos Basin, NW China. *Petroleum Exploration and Development*, 42(2): 167－176 (in Chinese with English abstract).

Zeng, H.B., Wang, F.R., Luo, J., et al., 2021. Characteristics of Pore Structure of Intersalt Shale Oil Reservoir by Low Temperature Nitrogen Adsorption and High Pressure Mercury Pressure Methods in Qianjiang Sag. *Bulletin of Geological Science and Technology*, 40(5): 242－252 (in Chinese with English abstract).

Zhang, B., Mao, Z.G., Zhang, Z.Y., et al., 2021. Black Shale Formation Environment and Its Control on Shale Oil Enrichment in Triassic Chang 7 Member,

Ordos Basin, NW China. *Petroleum Exploration and Development*, 48(6): 1127－1136 (in Chinese with English abstract).

附中文参考文献

胡素云, 赵文智, 侯连华, 等, 2020. 中国陆相页岩油发展潜力与技术对策. 石油勘探与开发, 47(4): 819－828.

李琪琪, 蓝宝锋, 李刚权, 等, 2021. 黔中隆起北缘五峰-龙马溪组页岩元素地球化学特征及其地质意义. 地球科学, 46(9): 3172－3188.

李士祥, 周新平, 郭芪恒, 等, 2021. 鄂尔多斯盆地长 7_3 亚段页岩油可动烃资源量评价方法. 天然气地球科学, 32(12): 1771－1784.

柳波, 石佳欣, 付晓飞, 等, 2018. 陆相泥页岩层系岩相特征与页岩油富集条件——以松辽盆地古龙凹陷白垩系青山口组一段富有机质泥页岩为例. 石油勘探与开发, 45(5): 828－838.

刘忠宝, 刘光祥, 胡宗全, 等, 2019. 陆相页岩层系岩相类型、组合特征及其油气勘探意义——以四川盆地中下侏罗统为例. 天然气工业, 39(12): 10－21.

卢双舫, 李俊乾, 张鹏飞, 等, 2018. 页岩油储集层微观孔喉分类与分级评价. 石油勘探与开发, 45(3): 436－444.

聂海宽, 张培先, 边瑞康, 等, 2016. 中国陆相页岩油富集特征. 地学前缘, 23(2): 55－62.

王民, 马睿, 李进步, 等, 2019. 济阳坳陷古近系沙河街组湖相页岩油赋存机理. 石油勘探与开发, 46(4): 789－802.

王子萌, 蒋裕强, 付永红, 等, 2022. 基于核磁共振表征渝西地区五峰组-龙一$_1$亚段页岩储层孔隙结构及非均质性. 地球科学, 47(2): 490－504.

吴松涛, 朱如凯, 崔京钢, 等, 2015. 鄂尔多斯盆地长 7 湖相泥页岩孔隙演化特征. 石油勘探与开发, 42(2): 167－176.

曾宏斌, 王芙蓉, 罗京, 等, 2021. 基于低温氮气吸附和高压压汞表征潜江凹陷盐间页岩油储层孔隙结构特征. 地质科技通报, 40(5): 242－252.

张斌, 毛治国, 张忠义, 等, 2021. 鄂尔多斯盆地三叠系长 7 段黑色页岩形成环境及其对页岩油富集段的控制作用. 石油勘探与开发, 48(6): 1127－1136.