引用格式：崔俊峰，杨金路，王民，等.基于随机森林算法的泥页岩孔隙度预测[J].油气地质与采收率，2023，30(6)：13-21.
CUI Junfeng, YANG Jinlu, WANG Min, et al.Shale porosity prediction based on random forest algorithm[J]J.Petroleum Geology and Recovery Efficiency, 2023, 30(6): 13-21.

基于随机森林算法的泥页岩孔隙度预测

崔俊峰[1]，杨金路[2]，王 民[2]，王 鑫[2]，吴 艳[2]，余昌琦[2]
（1.中国石油勘探开发研究院，北京 100083；2.中国石油大学（华东）深层油气重点实验室，山东 青岛 266580）

摘要：准确、快速地获取泥页岩孔隙度对页岩油空间分布及勘探目标预测具有重要意义。针对利用测井响应方程预测孔隙度精度较低的问题，建立一种基于随机森林算法的孔隙度预测模型，与BP神经网络、支持向量机和XGBoost算法进行预测精度对比，并利用SHAP方法分析测井参数的重要性和影响范围。研究结果表明：随机森林算法可以很好地预测泥页岩孔隙度，且预测效果好于BP神经网络、支持向量机和XGBoost算法；基于随机森林算法的泥页岩孔隙度预测在渤海湾盆地某凹陷应用发现，对模型预测孔隙度最重要的前3项测井参数为补偿中子、自然伽马和普通视电阻率；基于随机森林算法的泥页岩孔隙度预测模型可以快速识别单井孔隙度，不仅可以弥补因无法连续取心而难以获取完整孔隙度分布特征的问题，还能大幅提高孔隙度预测效率与精度。

关键词：随机森林；机器学习；测井；孔隙度预测；泥页岩
文章编号：1009-9603(2023)06-0013-09 DOI：10.13673/j.pgre.202212025
中图分类号：TE19 **文献标识码**：A

Shale porosity prediction based on random forest algorithm

CUI Junfeng[1], YANG Jinlu[2], WANG Min[2], WANG Xin[2], WU Yan[2], YU Changqi[2]
（1.*Research Institute of Petroleum Exploration & Development, PetroChina, Beijing City*, 100083, *China*; 2.*Laboratory of Deep Oil and Gas, China University of Petroleum (East China), Qingdao City, Shandong Province*, 266580, *China*）

Abstract: Precise and fast acquisition of shale porosity is important for the prediction of the spatial distribution of shale oil and the exploration target. To address the problem of low accuracy of porosity prediction using logging response equation, a porosity prediction model based on random forest algorithm is established, and the prediction accuracy is compared with those of BP neural network, support vector machine, and XGBoost algorithm, and the importance and influence range of logging parameters are analyzed by SHAP method. The results show that the random forest algorithm can better predict shale porosity, and the prediction effect is better than BP neural network, support vector machine, and XGBoost algorithm; the application of shale porosity prediction based on random forest algorithm in a depression in Bohai Bay Basin finds that the top three most important logging parameters for model prediction of porosity are compensation neutron, natural gamma, and ordinary apparent resistivity; the shale porosity prediction model based on random forest algorithm can quickly identify the porosity of a single well, which can not only compensate for the difficulty of obtaining the complete porosity distribution characteristics due to the inability of continuous coring but also significantly improve the efficiency and accuracy of porosity prediction.

Key words: random forest; machine learning; logging; porosity prediction; shale

北美页岩油勘探的成功掀起了全球范围内页岩油勘探的热潮[1]。中国页岩油可采资源量为74×

收稿日期：2022-12-24。
作者简介：崔俊峰(1970—)，男，陕西西安人，高级工程师，硕士，从事油气地质综合研究。E-mail：yjy_cjf@petrochina.com.cn。
通信作者：王民(1981—)，男，河北石家庄人，教授，博士。E-mail：wangm@upc.edu.cn。
基金项目：国家自然科学基金项目“非常规油气地质评价”(41922015)和“电磁波辐射页岩油原位转化中的非热效应机理及其意义”(42072147)。

10^8～372×10^8 t[2]，展现出广阔的前景。孔隙度是影响页岩储层储油能力和产油能力的重要因素，为获得准确的岩心孔隙度，可通过氮气吸附、高压压汞、扫描电镜、核磁共振等实验联合表征[2-3]，但受限于取心数量和实验费用，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，可间接实现孔隙度预测，但由于储层非均质性强，测井曲线之间存在大量的信息冗余，采用线性方程和经验统计公式无法较好地描述孔隙度，因此许多学者尝试通过机器学习方法预测孔隙度[4-11]。使用BP神经网络算法[5]基于测井数据预测孔隙度，其结果易陷入局部极值，一般配合相关的优化算法；将模糊逻辑算法与BP神经网络算法相结合，其结果优于BP神经网络算法的预测结果[6-8]；使用遗传算法[9-10]和帝国竞争算法[11]也可以对BP神经网络算法的参数选取进行优化，跳出局部极值，其优化后的预测结果相比BP神经网络算法更精确。YASIN等将支持向量机和粒子群优化算法相结合，成功预测出巴基斯坦萨万气田Lower Goru储层的孔隙度分布[12]。在同类研究中，机器学习算法在渗透率预测的表现也较为优异[13-17]。但以上算法多为单一机器学习算法，准确率普遍较低，有待进一步提升。随机森林算法通过基于集成决策树的学习算法，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[18-20]。为此，笔者以渤海湾盆地某凹陷A段孔隙度预测为目标，通过随机森林算法构建孔隙度预测模型，建立孔隙度的快速获取方法，进而探索将该方法应用于全井段孔隙度分布预测，以期为后续储层评价提供支撑。

1　方法原理

1.1　随机森林算法

随机森林（Random Forest，RF）算法作为一种基于CART决策树的集成学习算法，被广泛应用于分类或回归问题[21-22]。CART决策树适应于离散型数据，能够提取潜藏在列数据间的规则，但面对缺失数据时十分困难，并且在构建时极易出现过拟合的情况，性能具有一定的局限性[23-24]。

随机森林算法在CART决策树的基础上，通过随机有放回的抽取样本和随机无放回的抽取特征形成新样本集进行训练，并将生成的多棵CART决策树组成随机森林模型。随机森林算法作为一种组合分类器，其算法简单、易于实现、泛化能力强，在分类、回归问题上表现优异。

随机森林算法的工作流程（图1）为：①确定随机森林训练参数，如输入特征的可能的子集S_f，特征属性集F，决策树个数n，随机特征个数m等。②从训练样本集S中随机有放回抽取n个子样本集，再对每个子样本集的特征进行随机抽取后，训练决策树。③汇总决策树结果并输出。

1.2　SHAP算法

在完成孔隙度预测模型训练后，使用SHAP算法量化不同测井参数对模型预测孔隙度的重要性和影响范围[25]，目的是为了优选出对模型预测孔隙度最重要的测井参数，进而达到更准确的预测效果。SHAP算法的计算公式为：

$$\varphi_j=\sum_{S_j\subseteq\{x_1,\cdots,x_p\}\atop\{x_j\}}\frac{|S|!\left(p-|S|-1\right)!}{p!}\left[f_x\left(S\cup\left\{x_j\right\}\right)-f_x\left(S_f\right)\right] \tag{1}$$

2　数据处理与模型建立

2.1　数据处理

2.1.1　数据选取

研究采用的孔隙度数据和测井资料来自渤海湾盆地某凹陷3口井，目的层为A段，共获取373个样本数据。选取8种测井参数作为样本属性值，分别为自然伽马（GR）、普通视电阻率测井（R_4）、声波时差（AC）、补偿中子（CNL）、密度（DEN）和自然伽马能谱测井（K，U和TH）。每一类样本数据均由9

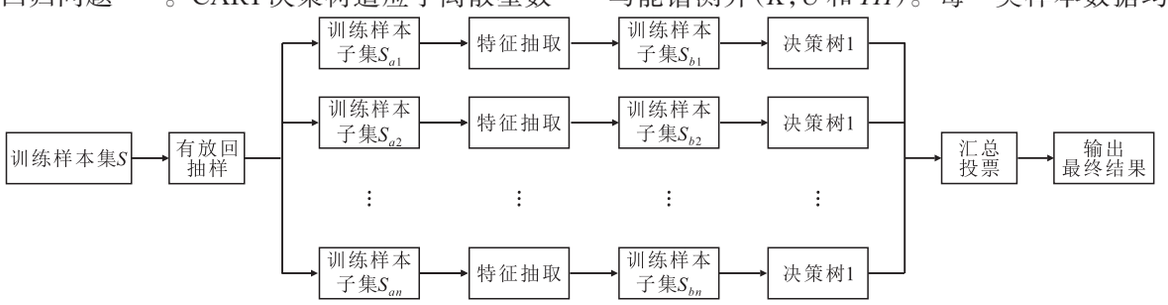


图1　随机森林算法工作流程

Fig.1　Flowchart of random forest algorithm

维向量组成，包括8维不同参数值及1维孔隙度标签。

2.1.2　数据归一化

由于各类测井数据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，会影响孔隙度的预测结果，为了消除量纲对模型预测效果的影响，需对数据进行归一化。通过最大和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，且样本数据经归一化后将按照7:3的比例随机划分为训练样本集和测试样本集。其定义如下：

$$x^* = \frac{x' - x'_{\min}}{x'_{\max} - x'_{\min}} \tag{2}$$

2.1.3　数据分析

由数据集中孔隙度与测井参数相关矩阵（图2）可以看出，孔隙度与部分测井参数具有相关性，其中*GR*，R_4，*AC*，*CNL*及*DEN*与孔隙度相关性较为明显，*K*，*U*和*TH*与孔隙度的相关性不明显。*GR*与R_4，*DEN*呈负相关，与*AC*，*CNL*呈正相关。R_4与*AC*，*CNL*呈负相关。*AC*，*CNL*和*DEN*三者之间存在强相关性，且*AC*与*CNL*呈正相关，*AC*与*DEN*呈负相关，*CNL*与*DEN*呈负相关。对于*K*，*U*和*TH*测井参数，除*K*，*TH*与*AC*，*CNL*的相关性较高外，其他的相关性均较低。

2.2　模型建立、调优及评价标准

为了获得最优的模型预测性能，需选取适合的

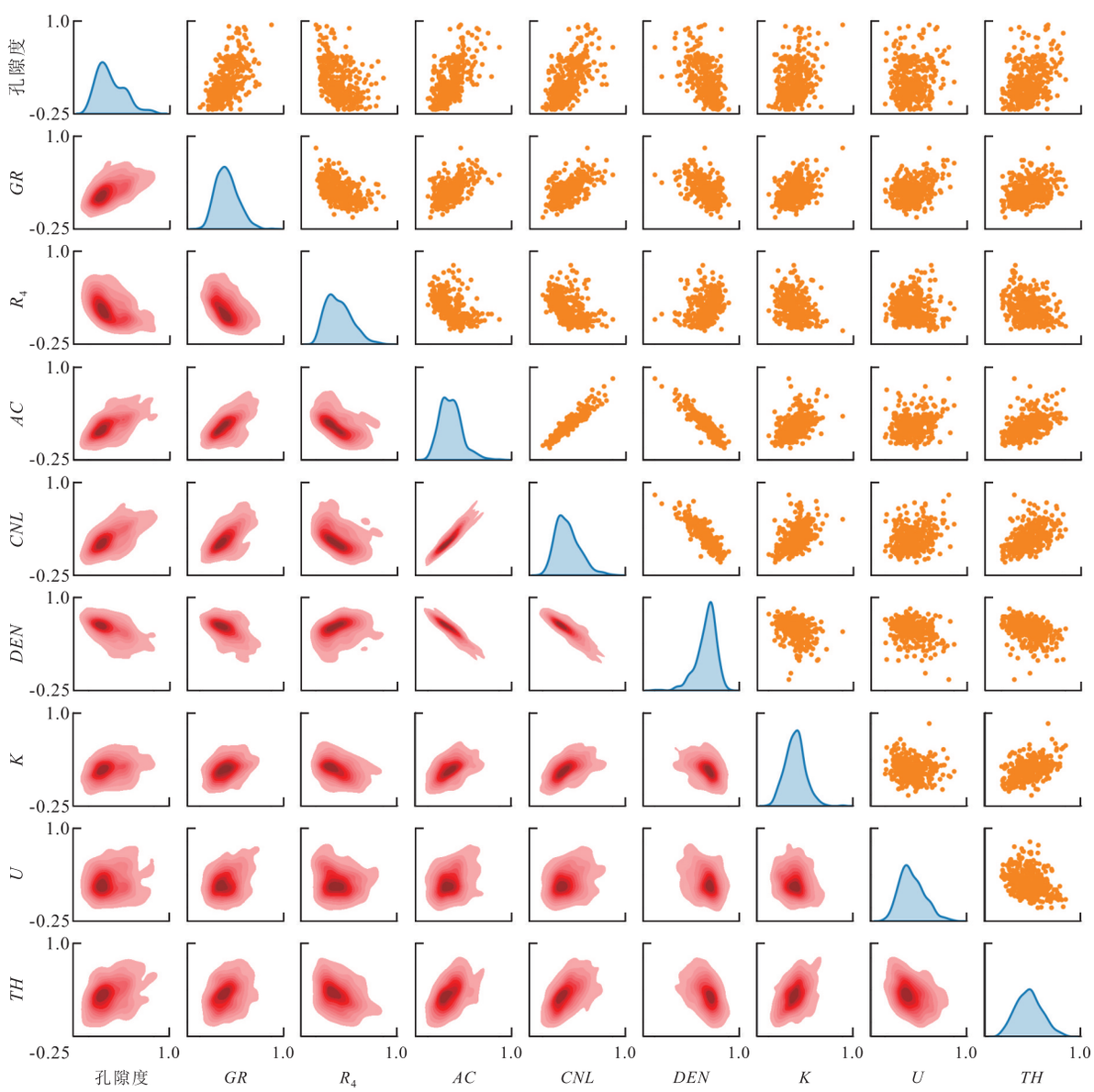


图2　数据集中孔隙度与测井参数相关矩阵

Fig.2　Correlation matrix of porosity and logging parameters in dataset

模型参数。针对BP神经网络、支持向量机、随机森林和XGBoost算法的4个模型，列出对其模型性能影响较大的参数及搜索范围（表1）。由于样本数量较少，故使用网格搜索和五折交叉验证对模型参数进行调优。通过五折交叉验证，将数据集随机分成5份数量相等的子样本集，轮流取1个子样本集作为测试数据集，其他子样本集用作训练。将每次试验得出的测试评分取平均值，作为模型效果的评估值。最终确定各模型最优参数组合（表1）。评价标准选择决定系数R^2：

$$R^2 = 1 - \frac{SS_{\text{residual}}}{SS_{\text{total}}} = 1 - \frac{\sum_{i=1}(\hat{y}^{(i)} - \bar{y})^2}{\sum_{i=1}(y^{(i)} - \bar{y})^2} \quad (3)$$

表1　不同孔隙度预测模型参数调优

Table1　Parameter tuning for different porosity prediction models

算法	模型参数	搜索范围	最优值
BP神经网络	学习率	0.000 1 ~ 0.2	0.1
	隐藏层神经元数量	10 ~ 300	180
支持向量机	*C*	0.001 ~ 500	5
	Gamma（核函数自带参数）	0.001 ~ 500	2
随机森林	迭代次数	10 ~ 400	120
	最大树深度	3 ~ 15	6
	内部节点再划分所需最小样本数	1 ~ 50	10
	叶子节点最小样本数	1 ~ 50	1
XGBoost	迭代次数	10 ~ 400	60
	最大树深度	3 ~ 15	7
	Gamma（核函数自带参数）	0 ~ 0.6	0.3
	子节点最小权重和	1 ~ 10	2
	样本采样率	0.5 ~ 1	0.9
	特征采样率	0.5 ~ 1	0.6

3　实验结果与分析

3.1　模型预测结果分析

完成模型参数调优后，对建立的孔隙度预测模型的预测效果进行验证。使用表1中确定的最优参数组合，BP神经网络、支持向量机、随机森林及XGBoost算法的4种模型对训练样本集及测试样本集的预测结果如图3所示。4种模型预测的孔隙度与实测孔隙度均呈近正比例关系，但R^2存在差距。XGBoost算法对训练样本集的拟合能力最强，其R^2值为0.998 9（图3d）；随机森林算法次之，R^2值为0.974 4（图3c）；支持向量机和BP神经网络算法的R^2值分别为0.942（图3b）和0.932（图3a）。对于测试样本集，随机森林算法取得的R^2值最高，达到0.900 2（图3c），预测能力较好；XGBoost算法的R^2值为0.888 1（图3d），支持向量机和BP神经网络算法的R^2值分别为0.892 4（图3b）和0.894 3（图3a）。

综上所述，随机森林算法虽然对训练样本集的拟合能力稍弱于XGBoost算法，但对测试样本集的表现好于XGBoost算法，表明其泛化能力更强，更有利于推广利用。

3.2　测井参数重要性分析

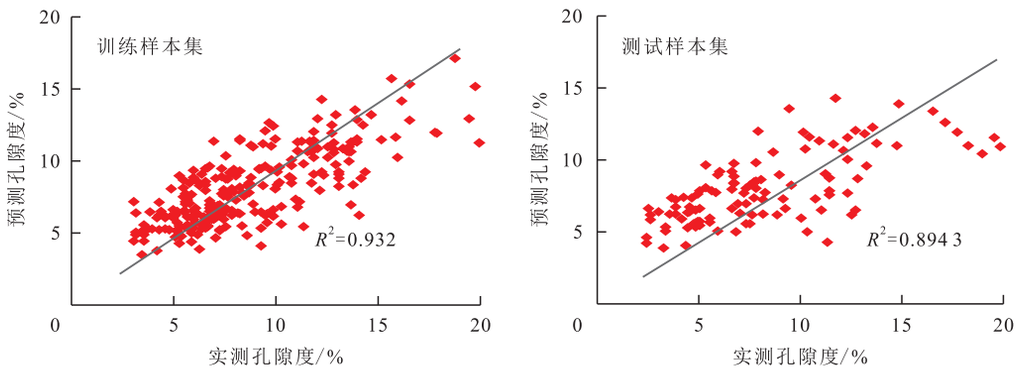
在完成回归模型训练后，使用SHAP算法量化不同测井参数对模型预测孔隙度的重要性。由测井参数的重要程度以及对孔隙度的影响分析（图4a）可以看出，其横坐标为SHAP值（基准孔隙度差值），纵坐标为各参数按重要性排序，由下至上测井参数重要性增大。当散点基准孔隙度差值大于0时，说明相对于基准孔隙度，该样本具有正贡献，而小于0则具有负贡献。

图4b定量表示出每项测井参数对孔隙度预测的贡献值。可以看出，*CNL*是影响孔隙度的最关键因素，相较于基准孔隙度差值（图4a），*CNL*最高可增加孔隙度约为3.8%；其次，*GR*也同样对孔隙度影响较大，最高可增加孔隙度为1.9%；R_4的重要性位列第3，最高可增加孔隙度约为2.6%；*U*和*AC*的影响较小，最高可增加孔隙度分别约为1.4%和0.8%；而*DEN*，*TH*和*K*对孔隙度的影响则最小。

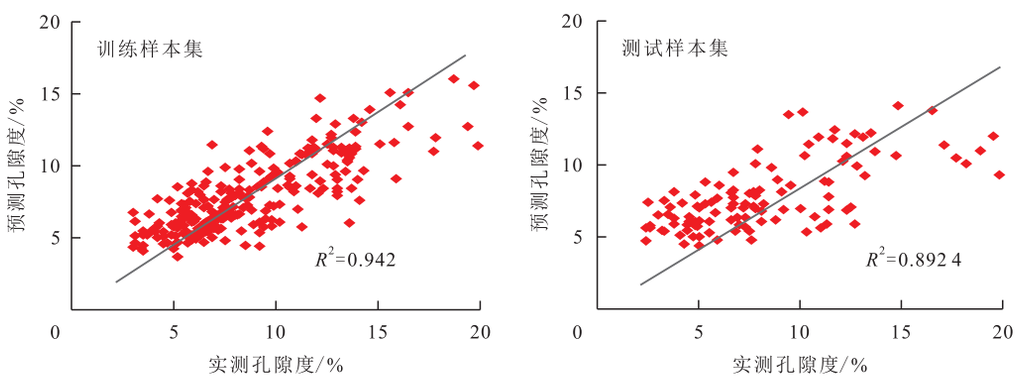
3.3　测井参数影响范围分析

将不同测井参数与SHAP值分别建立交汇图，可用于判断测井参数对孔隙度的影响范围（图5）。图5的横坐标为归一化后的测井参数，可换算为原始测井数据分布，纵坐标为该参数的SHAP值（基准孔隙度差值）。基准孔隙度差值为0表示数据集中孔隙度的平均水平。

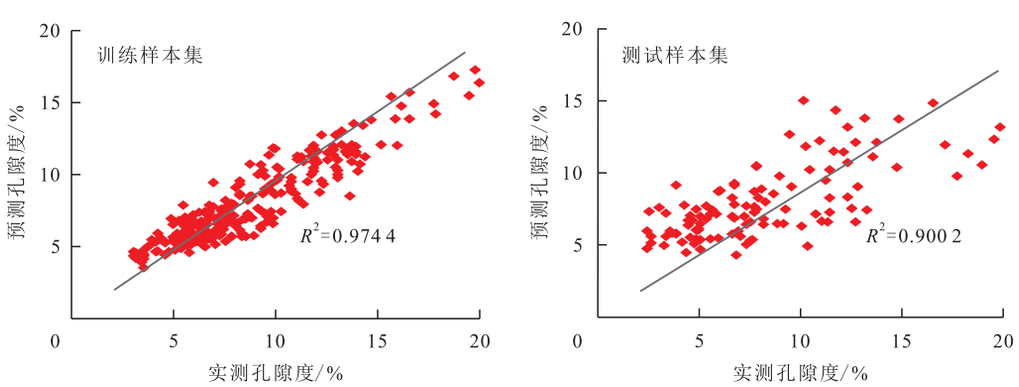
图5a—5d为图4b中SHAP值最大的4项孔隙度预测关键测井参数。*CNL*是影响孔隙度预测的最重要参数，随着*CNL*值的增加，基准孔隙度差值由负变正且逐渐增大，说明孔隙度逐渐增大；当*CNL*值大于25%，孔隙度明显增大。*GR*也具有相同的变化趋势，随着*GR*增加，基准孔隙度差值也增加；当*GR*大于62 API，孔隙度明显增大。R_4测井数据具有相反的变化趋势，R_4增加，基准孔隙度差值降低；当R_4小于11 Ω•m时，孔隙度明显增大。*U*和*DEN*与基准孔隙度差值呈负相关，当*U*值小于



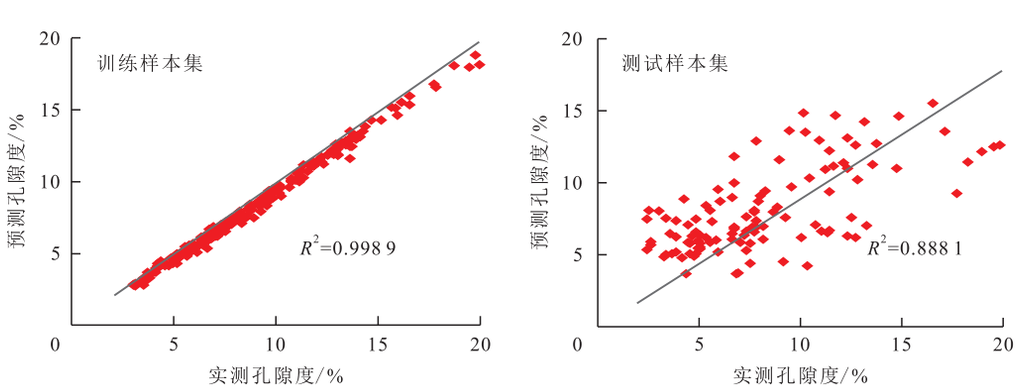
a—BP神经网络算法



b—支持向量机算法



c—随机森林算法



d—XGBoost算法

图3　不同算法孔隙度预测模型预测效果评价

Fig.3　Evaluation of prediction effect of porosity prediction models with different algorithms

3.6%、密度小于2.5 g/cm³时，孔隙度明显增加。

声波时差与基准孔隙度差值大致呈正相关，在声波时差大于88 μs/ft时，孔隙度明显增加；*TH*与基准孔隙度差值的关系不明显；*K*与基准孔隙度差值

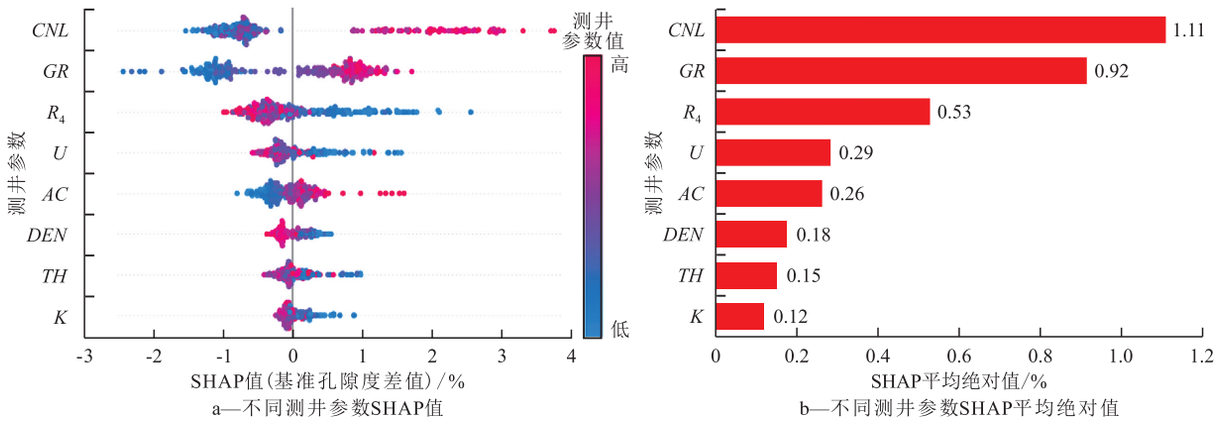


图4　孔隙度预测中的测井参数重要性分析

Fig.4　Analysis of importance of logging parameters in porosity prediction

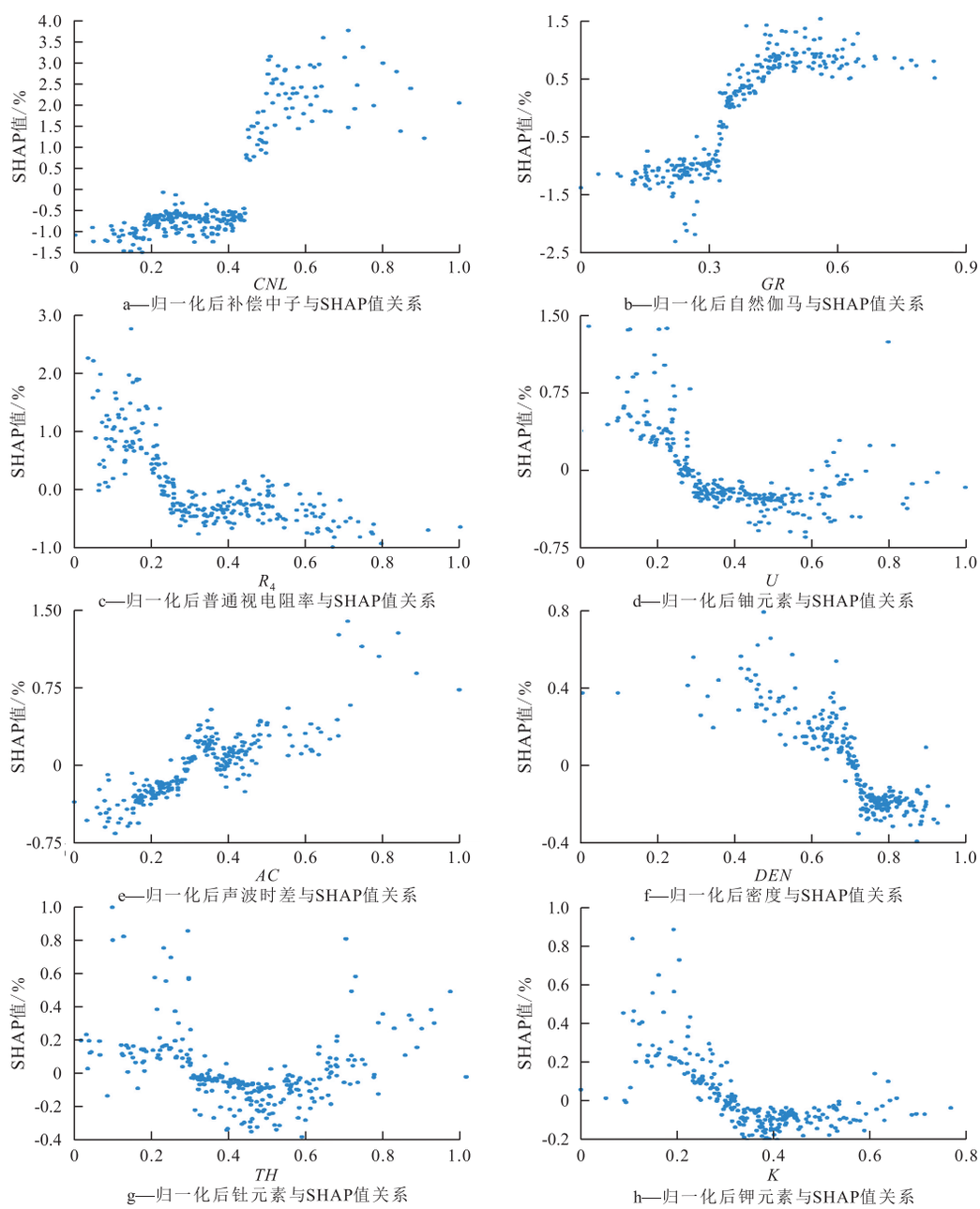


图5　孔隙度预测中测井参数影响范围分析

Fig.5　Analysis of influence ranges of logging parameters in porosity prediction

大致呈负相关。

3.4　单井孔隙度连续预测评价

BP神经网络、支持向量机、随机森林、XGBoost算法对于渤海湾盆地某凹陷单井X的孔隙度预测结果(图6)显示,埋深为3 100~3 180 m的4种模型的预测孔隙度多高于实测孔隙度,埋深为3 180~3 400 m的4种模型均可以较好地拟合出页岩孔隙度,但随机森林算法预测孔隙度的误差更小,预测能力更强。将基于随机森林算法的孔隙度预测模型推广应用于准噶尔盆地某凹陷单井Y,结果(图7)显示,在埋深为3 110~3 135和3 165~3 190 m存在少量孔隙度预测值偏离实测值,总体上孔隙度预测模型的预测值可以较好地拟合实测数据点。

综上所述,随机森林算法在页岩孔隙度预测中具有很好的准确性和较好的应用效果,不仅可以弥补因无法连续取心而难以获取完整孔隙度分布特征的问题,还能大幅提高孔隙度预测效率。

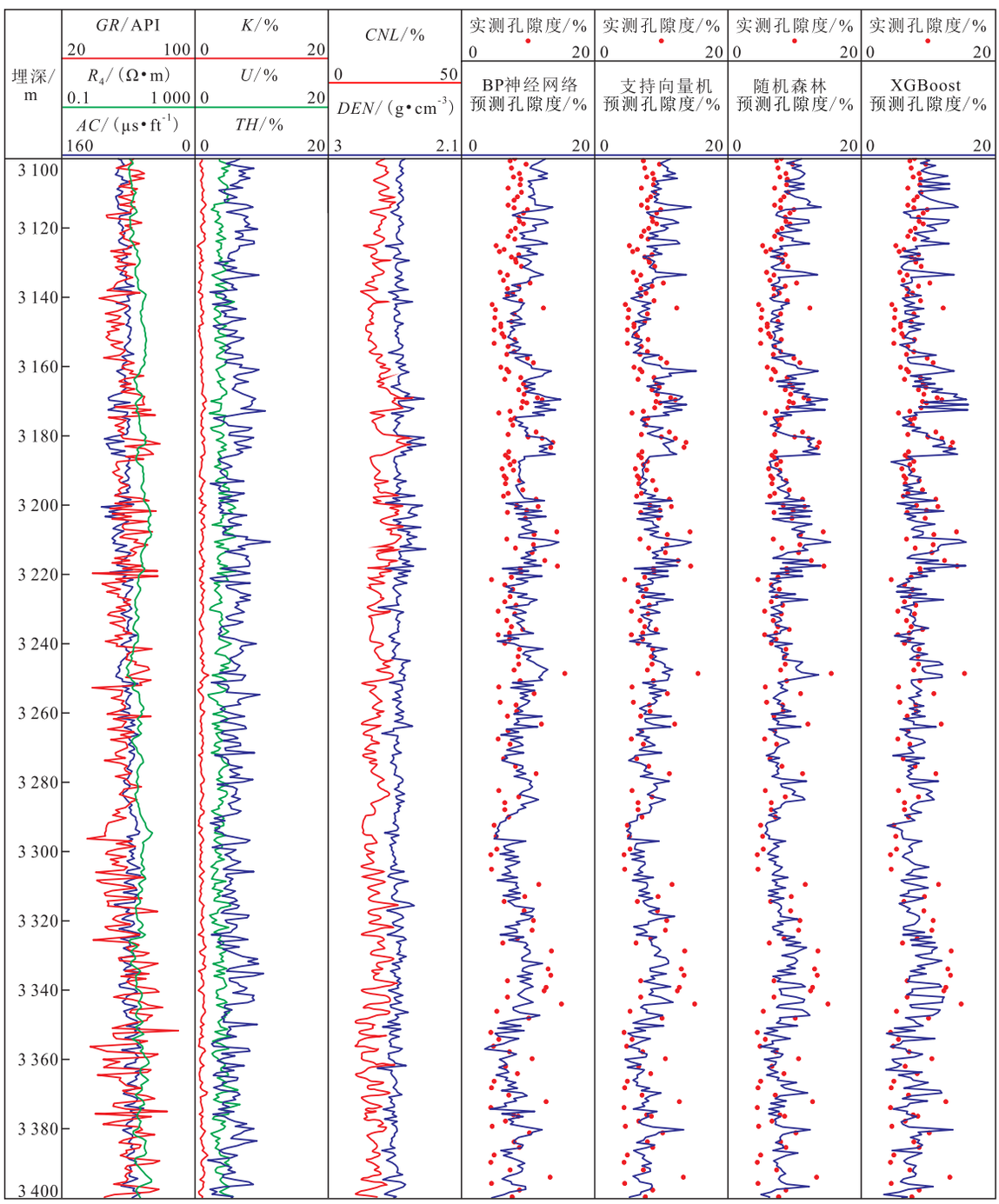


图6　渤海湾盆地某凹陷单井X不同孔隙度预测模型预测结果对比

Fig.6　Comparison of prediction results of different porosity prediction models for single well X in a depression in Bohai Bay Basin

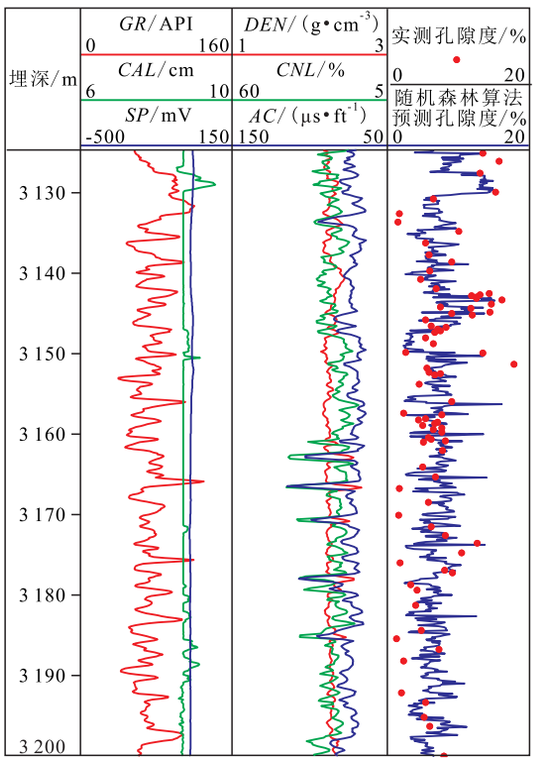


图7　基于随机森林算法的准噶尔盆地某凹陷单井Y孔隙度预测结果

Fig.7　Porosity prediction of Well Y in a depression in Junggar Basin based on random forest algorithm

4　结论

随机森林算法具有泛化能力强，有利于推广利用的优势。建立的基于随机森林算法的泥页岩孔隙度预测效果好于BP神经网络、支持向量机和XGBoost算法。对于渤海湾盆地某凹陷，该模型预测孔隙度的最重要的3项测井参数为补偿中子、自然伽马和普通视电阻率。针对利用常规测井响应方程预测孔隙度精度较低的问题，基于随机森林算法的孔隙度预测模型在泥页岩孔隙度预测中具有很好的应用前景，该模型不仅可以弥补因无法连续取心而难以获取完整孔隙度分布特征的问题，还能大幅提高孔隙度预测效率和精度，从而达到快速、准确预测单井孔隙度的目的，指导页岩油勘探开发。

符号解释

a，b——抽取的训练样本个数，个；

C——惩罚系数；

$f_x(S_f)$——特征子集S_f的预测；

F——特征属性集；

i——第i个特征；

j——第j个特征；

m——随机特征个数，个；

n——决策树个数，个；

p——输入特征的个数，个；

R^2——决定系数；

S——训练样本集；

S_{an}，S_{bn}——训练样本子集；

S_f——输入特征的可能的子集；

SS_{residual}——残差平方和；

SS_{total}——总离差平方和；

$\dfrac{|S|!(p-|S|-1)!}{p!}$——子集$S_f$的特征组合情况占比；

x'——样本；

x——原始测井数据；

x^*——归一化后的测井数据；

x_j——样本的第j个特征；

x_p——样本的第p个特征；

y——模型预测值；

$\hat{y}$——样本真实值；

$\bar{y}$——样本真实平均值；

φ——特征边际贡献；

φ_j——特征j的贡献度。

参考文献

[1] HACKLEY P C，FISHMAN N，WU T，et al. Organic petrology and geochemistry of mudrocks from the Lacustrine Lucaogou Formation，Santanghu Basin，Northwest China：Application to lake basin evolution［J］. International Journal of Coal Geology，2016，168：20-34.

[2] 吴伟，梁志凯，郑马嘉，等.页岩储层孔隙结构与分形特征演化规律[J].油气地质与采收率，2022，29(4)：35-45.
WU Wei，LIANG Zhikai，ZHENG Majia，et al. Pore structures in shale reservoirs and evolution laws of fractal characteristics［J］. Petroleum Geology and Recovery Efficiency，2022，29(4)：35-45.

[3] 徐云龙，张洪安，李继东，等.渤海湾盆地东濮凹陷陆相页岩层系储集特征及其主控因素[J].断块油气田，2022，29(6)：729-735.
XU Yunlong，ZHANG Hongan，LI Jidong，et al. Reservoir characteristics and its main controlling factors of continental shale strata in Dongpu Sag，Bohai Bay Basin［J］. Fault-Block Oil and Gas Field，2022，29(6)：729-735.

[4] 胡素云，赵文智，侯连华，等.中国陆相页岩油发展潜力与技术对策[J].石油勘探与开发，2020，47(4)：819-828.
HU Suyun，ZHAO Wenzhi，HOU Lianhua，et al. Development potential and technical strategy of continental shale oil in China［J］. Petroleum Exploration and Development，2020，47(4)：819-828.

[5] HELLE H B，BHATT A，URSIN B. Porosity and permeability prediction from wireline logs using artificial neural networks：a north sea case study［J］. Geophysical Prospecting，2001，49

(4)：431-444.

[6] MALKI H A，BALDWIN J. A neuro-fuzzy based oil/gas producibility estimation method［C］. Proceedings of the 2002 International Joint Conference on Neural Networks. Honolulu：IEEE，2002：896-901.

[7] REZAEE M R，KADKHODAIE-ILKHCHI A，ALIZADEH P M. Intelligent approaches for the synthesis of petrophysical logs［J］. Journal of Geophysics and Engineering，2008，5(1)：12-26.

[8] OLATUNJI S O，SELAMAT A，ABDULRAHEEM A. A hybrid model through the fusion of type-2 fuzzy logic systems and extreme learning machines for modelling permeability prediction［J］. Information Fusion，2014，16：29-45.

[9] AHMADI A　M，ZENDEHBOUDI S，LOHI A，et al. Reservoir permeability prediction by neural networks combined with hybrid genetic algorithm and particle swarm optimization［J］. Geophysical Prospecting，2013，61(3)：582-598.

[10] KHALIFAH H A，GLOVER P W J，LORINCZI P. Permeability prediction and diagenesis in tight carbonates using machine learning techniques［J］. Marine and Petroleum Geology，2020，112：104096.

[11] JAMSHIDIAN M，HADIAN M，ZADEH M M，et al. Prediction of free flowing porosity and permeability based on conventional well logging data using artificial neural networks optimized by imperialist competitive algorithm-a case study in the South Pars gas field［J］. Journal of Natural Gas Science and Engineering，2015，24：89-98.

[12] YASIN Q，SOHAIL G M，KHALID P，et al. Application of machine learning tool to predict the porosity of clastic depositional system，Indus Basin，Pakistan［J］. Journal of Petroleum Science and Engineering，2021，197：107975.

[13] ZHANG Guoyin，WANG Zhizhang，MOHAGHEGH S，et al. Pattern visualization and understanding of machine learning models for permeability prediction in tight sandstone reservoirs［J］. Journal of Petroleum Science and Engineering，2021，200：108142.

[14] GHOLAMI R，SHAHRAKI A R，PAGHALEH M J. Prediction of hydrocarbon reservoirs permeability using support vector machine［J］. Mathematical Problems in Engineering，2012，2012：670723.

[15] AL-ANAZI A F，GATES I D. Support vector regression to predict porosity and permeability：effect of sample size［J］. Computers and Geosciences，2012，39：64-76.

[16] ZHONG Zhi，CARR T R，WU Xinming，et al. Application of a convolutional neural network in permeability prediction：a case study in the Jacksonburg-Stringtown oil field，West Virginia，USA［J］. Geophysics，2019，84(6)：B363-B373.

[17] ZHOU Kaibo，HU Yangxiang，PAN Hao，et al. Fast prediction of reservoir permeability based on embedded feature selection and Light GBM using direct logging data［J］. Measurement Science and Technology，2020，31(4)：045101.

[18] BIAU G，SCORNET E. A random forest guided tour［J］. Test，2016，25(2)：197-227.

[19] WANG Pu，CHEN Xiaohong，WANG Benfeng，et al. An improved method for lithology identification based on a hidden Markov model and random forests［J］. Geophysics，2020，85(6)：1-56.

[20] FENG Runhai. Improving uncertainty analysis in well log classification by machine learning with a scaling algorithm［J］. Journal of Petroleum Science and Engineering，2021，196：107995.

[21] BREIMAN L. Random forests［J］. Machine Learning，2001，45(1)：5-32.

[22] ROKACH L. Ensemble-based classifiers［J］. Artificial Intelligence Review，2010，33：1-39.

[23] MYLES A J，FEUDALE R N，LIU Yang，et al. An introduction to decision tree modeling［J］. Journal of Chemometrics，2004，18(6)：275-285.

[24] SONG Yanyan，LU Ying. Decision tree methods：applications for classification and prediction［J］. Shanghai Archives of Psychiatry，2015，27(2)：130-135.

[25] LUNDBERG S M，LEE S I. A unified approach to interpreting model predictions［C］. Proceedings of the 31st International Conference on Neural Information Processing System. California：Curran Associates，2017：4 768-4 777.

编辑　邹潋滟