基于极端梯度提升树算法的测井曲线重构方法

齐春生，丁 磊，王 一，郑志锋，张 茜

（中海石油（中国）有限公司 海南分公司，海口 570311）

摘要：南海西部海域油田测井资料采集过程中经常会出现测井曲线缺失的问题，为油气层测井解释精确评价带来了较大困难。为解决该问题，基于 XGBoost 算法综合利用测井、气测录井资料实现测井曲线重构。该方法支持处理稀疏数据，并且通过网格搜索实现模型参数自动优选。通过在靶区对比不同机器学习算法以及不同测井曲线的重构效果，该方法的调参难度低、计算精度高，可以有效提高曲线重构的工作效率。将该方法应用到南海西部油田低阻水淹层识别工作中，取得了良好的应用效果。

关键词：XGBoost；曲线重构；稀疏数据；网格搜索；水淹层

中图分类号：TE151 **文献标志码**：A **文章编号**：1671－1807(2022)10－0405－06

南海西部海域油气田在进行测井资料采集的过程中，由于仪器故障等原因，往往会出现部分井段测井资料丢失的情况，为测井解释精确评价带来困难。虽然可以通过现场随钻上提复测或电缆补测的方式重新采集测井资料，但这种方式一方面增加了作业成本，另一方面复测得到的测井资料可能会由于长时间泥浆浸泡出现失真的现象。

针对该问题，研究人员通常使用各类数学算法，根据已有测井曲线进行缺失曲线重构，达到缺失测井曲线补全的目的。目前，国内外学者使用的测井曲线重构方法主要有：①使用经验公式拟合法对曲线进行校正，主要有 Gardner 公式[1]和 Faust 公式[2]；②使用岩石物理建模的方式针对不同性质储层进行建模正演[3]；③使用多元线性回归方法进行拟合，建立现有曲线与目标曲线的线性相关关系[3-6]。以上几种方法虽然计算精度满足要求，但都存在应用局限性。近年来，随着机器学习技术的不断发展，研究人员开始尝试使用机器学习算法如 BP 神经网络[7]、MRGC 图形聚类[8]、SVM 支持向量机[9]等进行测井曲线重构，取得了良好的应用效果。

研究人员经过长期的应用实践后发现，机器学习算法与传统方法相比，在具备高度参数化、适用范围广的特点的同时，也存在着参数众多、容易发生过拟合现象的缺点；不仅如此，由于海上油田各井测井系列不同，导致难以建立大规模训练样本，也影响了机器学习模型的精度。

为解决传统机器学习算法在测井曲线重构中面临的问题，本文使用极端梯度提升树 XGBoost 算法[10]，开发了一种支持参数自动优选的测井曲线重构方法。将其应用到靶区 Z 油田的测井曲线重构工作中，以期取得更便捷、更准确的应用效果。

1 XGBoost 算法概述

1.1 算法原理

XGBoost 算法是梯度提升树的一种，其主要思想是使用贪心策略和二次最优化的方法，逐步在集成中添加弱分类器，每一个弱分类器都通过拟合前序分类器的残差对其进行改正，由此得到强分类器[11]。

XGBoost 算法使用的树模型为 CART 回归树模型，其预测模型以及目标损失函数为

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{1}$$

$$\mathrm{Obj}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

式中：K 为树的总数；f_k 为第 k 棵树；$\hat{y}_i$ 为样本 x_i 的预测结果；$l(y_i, \hat{y}_i)$ 为样本 x_i 的训练误差，$\Omega(f_k)$ 为第 k 棵树的正则项。

采用加法模型计算第 t 个模型，并对其目标损失函数进行泰勒二阶展开，可以得到

$$\hat{y}_i^{\,t} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{\,t-1} + f_t(x_i) \tag{3}$$

$$\mathrm{Obj}(\theta)^t \approx \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{t-1}) + \partial_{\hat{y}_i^{t-1}} l(y_i, \hat{y}_i^{t-1}) f_t(x_i) +$$

收稿日期：2022-05-17

作者简介：齐春生（1977—），男，河南周口人，中海石油（中国）有限公司海南分公司，工程师，研究方向为油气勘探开发数字化。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$\frac{1}{2}\partial^2_{\hat{y}_i^{t-1}} l(y_i,\hat{y}_i^{t-1})\ f_t(x_i)^2]+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2+C \tag{4}$$

式中：$\mathrm{Obj}(\theta)^t$ 为第 t 个模型的目标损失函数；$\partial_{\hat{y}_i^{t-1}} l(y_i,\hat{y}_i^{t-1})$ 为 $l(y_i,\hat{y}_i^{t-1})$ 关于 $\hat{y}_i^{t-1}$ 的一阶导数；$\partial^2_{\hat{y}_i^{t-1}} l(y_i,\hat{y}_i^{t-1})$ 为 $l(y_i,\hat{y}_i^{t-1})$ 关于 $\hat{y}_i^{t-1}$ 的二阶导数；$\hat{y}_i{}^t$ 为第 t 个模型的预测结果；f_t 为第 t 棵树；γ、λ 为常数；T 为回归树叶节点个数；w_j 为叶节点权重；C 为前 $t-1$ 棵树的复杂度。

式(4)中，$i=1\sim n$ 求和代表在样本中遍历，$j=1\sim T$ 求和代表在树叶子节点遍历，可以将该式全部转换为在树叶子节点遍历的方式，得到

$$\mathrm{Obj}(\theta)^t\approx\sum_{j=1}^{T}\Big[(\sum_{i\in I_j}k_i)\,w_j+\frac{1}{2}(\sum_{i\in I_j}h_i+\lambda)\,w_j^2\Big]+\gamma T \tag{5}$$

式中：$k_i=\partial_{\hat{y}_i^{t-1}} l(y_i,\hat{y}_i^{t-1})$；$h_i=\partial^2_{\hat{y}_i^{t-1}} l(y_i,\hat{y}_i^{t-1})$；$I_j=\{i\mid q(x_i)=j\}$ 表示在第 j 个叶子节点上的样本；w_j 为第 j 个叶子的得分值。

令 $G_j=\sum_{i\in I_j}k_i$，$H_j=\sum_{i\in I_j}h_i$，并对 w_j 求偏导，令偏导数等于0，求解得

$$w_j^*=-\frac{G_j}{H_j+\lambda} \tag{6}$$

$$\mathrm{Obj}^*=-\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda}+\gamma T \tag{7}$$

确定了回归树的结构和每个叶子节点得分值计算方法之后，XGBoost 算法利用贪婪算法，遍历所有特征的划分点，根据式(8)计算分裂后的得分增益，确定划分方向。

$$\mathrm{Score}=\frac{1}{2}\Big[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\Big]-\gamma \tag{8}$$

式中：$\frac{G_L^2}{H_L+\lambda}$ 为分裂后左子树得分；$\frac{G_R^2}{H_R+\lambda}$ 为分裂后右子树得分；$\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}$ 为不分裂得分；γ 为新增子树复杂度代价。

1.2 算法特点

XGBoost 算法与传统机器学习算法相比做出了多项改进，主要表现在防止过拟合、特征重要性计算以及稀疏矩阵计算支持上。

1)防止过拟合。XGBoost 算法除了在损失函数 $\mathrm{Obj}(\theta)$ 中加入正则项 $\Omega(f_k)$ 外，还使用了 Shrinkage 缩减法和 Column Subsampling 列抽样法，有效防止过拟合的发生。

2)特征重要性计算。XGBoost 算法通过统计提升树中每个特征分裂点的得分得到特征重要性，由此直观展示每个特征对机器学习任务的贡献价值。

3)稀疏矩阵计算支持。当训练样本中某特征包含缺失值时，XGBoost 算法可以使用稀疏感知分割算法分别将该特征划分到左子树和右子树，当划分到左子树时有

$$\begin{cases}G_R=G_R+k_j,\quad G_L=G-G_R\\ H_R=H_R+h_j,\quad H_L=H-H_R\end{cases} \tag{9}$$

当划分到右子树时有

$$\begin{cases}G_L=G_L+k_j,\quad G_R=G-G_L\\ H_L=H_L+h_j,\quad H_R=H-H_L\end{cases} \tag{10}$$

然后根据式(9)和式(10)分别计算得分收益，选择得分最高的方向划分子树。

综上所述，XGBoost 算法可以有效避免过拟合现象的发生，还提供了特征重要性计算以及稀疏矩阵计算支持，能够解决传统机器学习算法在测井曲线重构中面临的问题。

2 XGBoost 算法测井曲线重构效果研究

为测试 XGBoost 算法测井曲线重构效果，以靶区 Z 油田 22 井作为测试井，进行两方面研究：①分别使用测试井的本井测井曲线以及临井测井曲线组成训练样本重构电阻率曲线，对比分析 XGBoost 算法与传统机器学习算法 BP 神经网络、SVM 支持向量机的计算效果；②分别基于特征数据完整以及存在特征数据缺失的临井测井曲线重构电阻率曲线，分析 XGBoost算法针对稀疏测井曲线训练样本的计算效果。

2.1 测井曲线重构效果对比实验

靶区 Z 油田测试井以及 6 口临井均测量了自然伽马、电阻率、补偿中子、密度测井以及气测录井资料，本实验首先选取测试井 2 512～2 650 m 深度段测录井曲线数据组成本井训练样本，选取测试井 2 651～2 879 m 深度段测录井曲线数据作为测试样本，并使用 XGBoost 算法进行电阻率曲线重构；然后取临井同层段测录井曲线数据经标准化后组成临井训练样本，并使用 BP 神经网络、SVM 支持向量机以及 XGBoost 算法进行电阻率曲线重构，得到图 1 所示的电阻率曲线重构效果对比图。

图 1 中共有 4 个电阻率重构道，道中红色实线为真实电阻率曲线，蓝色实线为重构电阻率曲线。前 3 道为使用 BP 神经网络、SVM 支持向量机以及 XGBoost 算法学习临井训练样本得到的电阻率重

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

构结果,第 4 道为使用 XGBoost 算法学习本井训练样本得到的电阻率重构结果。可以看出:①与 BP 神经网络、SVM 支持向量机算法相比,使用 XGBoost算法重构出的电阻率曲线与真实值更为吻合;②使用本井训练样本和临井训练样本学习得到的 XGBoost 算法模型效果接近。上述现象说明 XGBoost 算法适用于测井曲线重构,当一口井全井段需要重构测井曲线时,可以使用临井数据作为训练样本对目标井进行测井曲线重构,但是由于各井测井资料测量仪器和环境不同,需要首先对测井曲线进行标准化处理;当一口井仅有部分井段需要重构测井曲线时,可以直接使用该井其他深度段测井曲线作为训练样本对目标井段进行测井曲线重构,这种方式无需对曲线进行标准化,更为方便快捷。

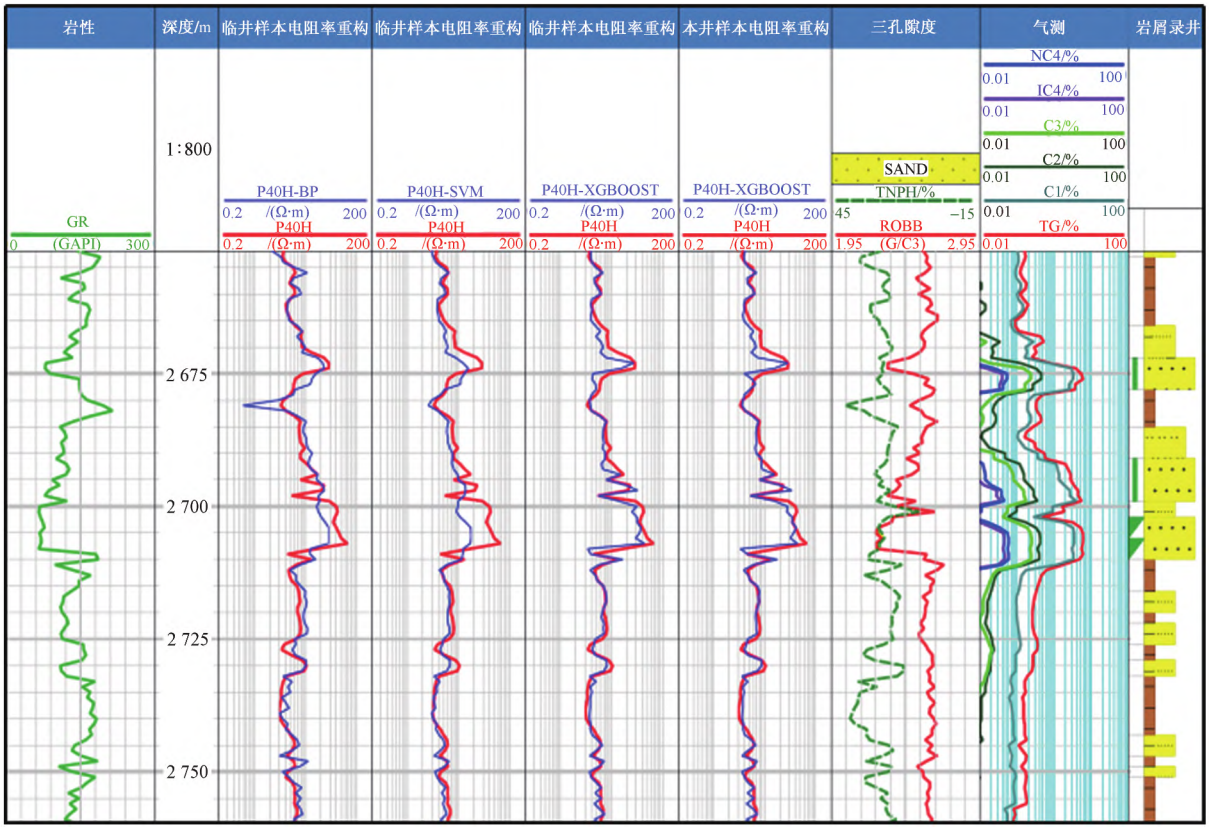

图 1 测试井电阻率重构效果对比

2.2 稀疏训练样本曲线重构实验

本实验基于临井训练样本,首先取训练样本中 5 口井,分别删除部分气测总烃曲线、补偿中子测井曲线值组成稀疏训练样本,然后取剩余一口特征数据完整井的测录井数据组成完整训练样本。基于上述 3 个训练样本使用 XGBoost 算法进行电阻率曲线重构,得到图 2 所示的电阻率曲线重构效果对比图。

图 2 中共有 3 个电阻率重构道,道中红色实线为真实电阻率曲线,蓝色实线为重构电阻率曲线。前两道为基于稀疏训练样本学习得到的电阻率重构结果,第 3 道为基于去除稀疏数据后的完整训练样本学习得到的电阻率重构结果。可以看出:①当测井训练样本存在特征数据缺失时,XGBoost 算法依然可以有效地完成测井曲线重构任务,而且由于大量有用信息得以保留,模型精度相较于数据较少的完整训练样本有显著提高;②对比基于两种稀疏训练样本学习得到的电阻率重构效果,基于缺失补偿中子曲线的训练样本训练得到的重构效果更好。针对该现象,分析实验 1 中原始训练样本模型特征重要性(表 1),总烃曲线的特征重要性为 0.495,补偿中子曲线的特征重要性为 0.018,说明总烃曲线特征重要性更高,该曲线缺失对电阻率重构任务精度影响更大。上述现象说明 XGBoost 算法可以有效应对测井训练样本数据缺失的情况,既能有效提高模型精度,又可以节约数据准备时间,可以为测井曲线重构带来极大帮助。

2.3 小结

综上,使用 XGBoost 算法进行测井曲线重构相比于传统算法有以下优势:

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

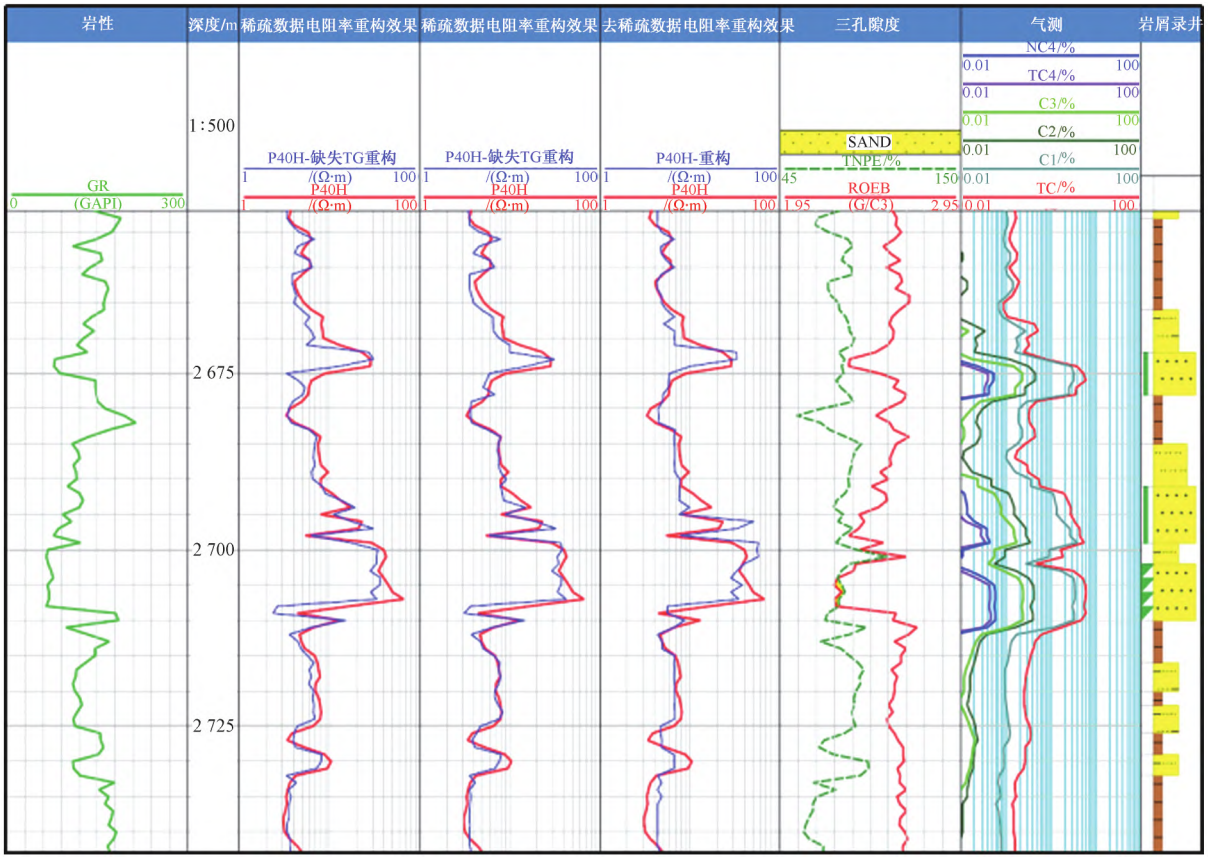

图 2　稀疏训练样本电阻率重构效果对比

表 1　XGBoost 算法模型特征重要性

曲线名	特征重要性
总烃	0.495
甲烷	0.210
伽马	0.127
井径	0.082
密度	0.068
补偿中子	0.018

1)对单一井来说，测井曲线重构不需要进行归一化处理；对于多井来说，算法可以直接对包含稀疏数据的多井训练样本进行处理，所以可以大幅节约数据准备时间。

2)得益于更多的训练样本数据，使用 XGBoost 算法可以有效提高测井曲线重构精度。

3　应用实例

3.1　XGBoost 算法测井曲线重构流程

基于 XGBoost 算法重构测井曲线的流程主要包括以下 4 个步骤：

1)选择本井或周围临井测录井曲线数据(临井测井曲线数据需要事先进行曲线标准化处理)组成训练样本。

2)对训练样本进行数据清理，删除异常值和重复值，并按比例将训练样本分为训练集和验证集。

3)使用 XGBoost 算法学习训练集，设置算法参数变化范围，使用网格搜索[12]得到最佳的参数组合，并在验证集上验证模型效果。

4)将得到的 XGBoost 算法模型应用于实际井进行测井曲线重构。

3.2　电阻率重构在低阻油田水淹层识别中的应用

靶区 Z 油田 A 油组经过多年的开发生产，构造高部位井点局部已经出现水淹。该油组为低阻油层，水淹类型为盐水水淹，故油层与水淹层电阻率响应特征类似，无法使用常规定性识别图版识别水淹层。随着油田挖潜工作的实施，如何准确识别水淹层并进行合理避射成为油田下一步深度开发的关键。

油田水淹过程中，随着储层内烃类物质的含量和组分发生变化，气测录井曲线也会随之发生变化。因此，综合分析测井、气测录井资料，可以帮助研究人员识别水淹层[13-14]。

基于上述原理，本文以靶区未水淹井为训练井，以训练井自然伽马曲线、物性曲线和气测录井曲线作为学习曲线，以电阻率曲线作为预测曲线，使用 XGBoost 算法建立未水淹电阻率重构模型，将该模型应用于水淹井得到原始未水淹电阻率曲线，

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

对比该曲线与真实电阻率曲线形态差异,由于水淹类型为盐水水淹,所以水淹后电阻率曲线会低于原始未水淹电阻率曲线,由此识别水淹层。

3.2.1 模型建立

取训练井中自然伽马、密度、补偿中子、钻时、总烃、甲烷、乙烷、丙烷、异丁烷、正丁烷曲线数据为特征学习曲线,取电阻率曲线为预测曲线。XGBoost算法需要调整的参数及其变化范围见表2。

表2 XGBoost算法参数调整变化范围

名称	最小值	最大值	步长
max_depth	3	10	1
learning_rate	0.01	0.2	0.01
min_child_weight	1	11	2
subsample	0.5	1	0.1
colsample_bytree	0.5	1	1

使用网格搜索遍历34 560种参数组合,得到最优参数组合:max_depth为3,learning_rate为0.19,min_child_weight为1,subsample为1,colsample_bytree为0.8。该模型在验证集的应用效果如图3所示。可以看出重构电阻率值与真实电阻率值接近,分布在45°线附近,说明XGBoost算法重构电阻率模型精度较高,可以应用于水淹井的未水淹原始电阻率曲线计算工作中。

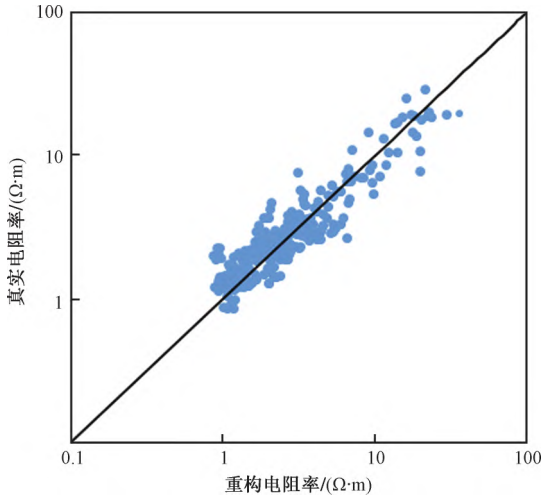

图3 XGBoost算法电阻率重构模型验证效果

3.2.2 水淹识别

将XGBoost算法计算得到的未水淹原始电阻率重构模型应用于Z油田水淹井B1H井,得到图4

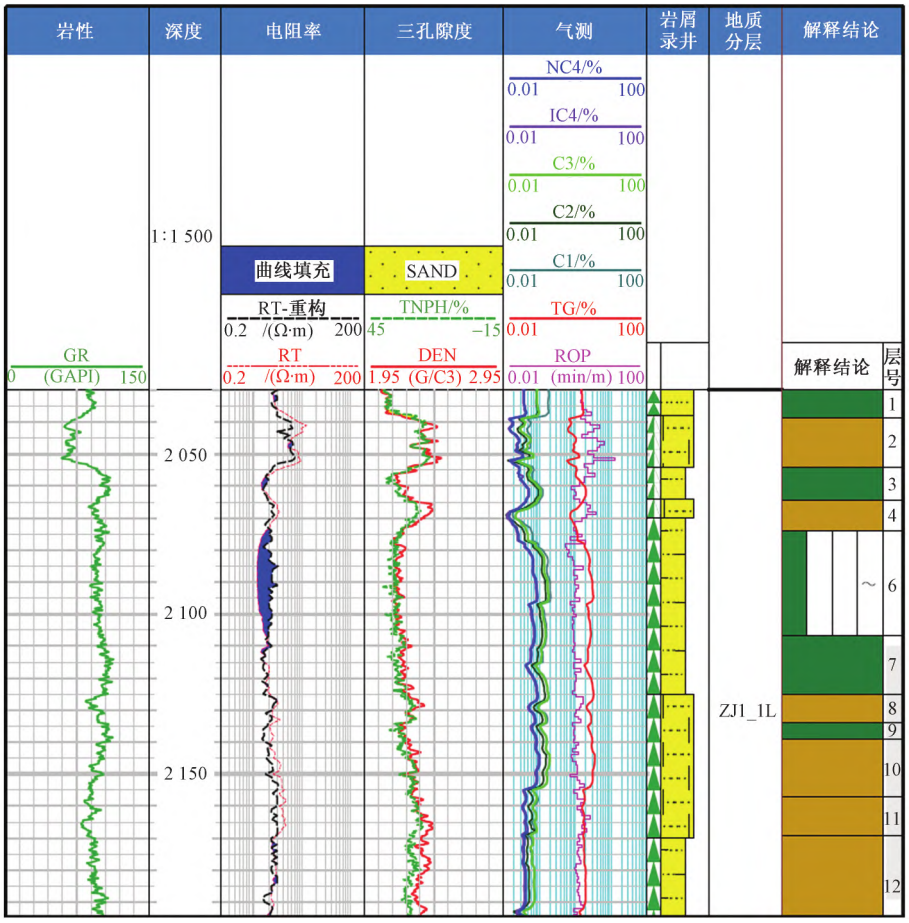

图4 B1H井原始未水淹电阻率重构剖面图

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

所示的原始未水淹原始电阻率重构剖面图。从图中可以看出，电阻率道中原始未水淹电阻率与实测电阻率在未水淹层段基本重合，在水淹层段出现明显包络现象，与实际情况相符，说明使用未水淹原始电阻率和实测电阻率交会可以有效完成靶区低阻油田水淹层定性识别任务。

4 结论

1)基于 XGBoost 算法的测井曲线重构方法计算精度高、调参难度低，而且支持对稀疏数据的处理，可以节约大量数据准备时间，为测井曲线重构带来极大帮助。

2)利用包络法将 XGBoost 测井曲线重构方法应用到靶区 Z 油田 A 油组低阻水淹层识别工作中，取得了较好的应用效果。

参考文献

[1] 袁全社，周家雄，李勇，等. 声波测井曲线重构技术在储层预测中的应用[J]. 中国海上油气，2009，21(1)：23-26.

[2] 陈钢花，王永刚. Faust 公式在声波曲线重构中的应用[J]. 勘探地球物理进展，2005(2)：125-128，10.

[3] 李宁，于鹏，田军，等. 基于龙凤山营城组储层预测的声波和密度测井曲线重构方法探讨：以 LB1 井为例[J]. 油气藏评价与开发，2016，6(4)：18-22.

[4] 李婷婷，陈井瑞，郭岳. 复杂地质条件下密度曲线重构在储层预测中的应用[J]. 中国锰业，2019，37 (5)：19-23.

[5] 王俊瑞，梁力文，邓强，等. 基于多元回归模型重构测井曲线的方法研究及应用[J]. 岩性油气藏，2016，28(3)：113-120.

[6] 余为维，冯磊，杜艳艳. 基于特征曲线重构的波阻抗反演在复杂储层预测中的应用[J]. 科学技术与工程，2019，19(4)：58-65.

[7] 宋梅远. 测井曲线重构在哈山地区油气水层判识中的综合应用[J]. 科学技术与工程，2014，14(19)：211-216.

[8] 何苗，于海峰，田中元，等. 乍得 B 盆地密度和声波时差曲线异常自动识别与重构[J]. 地球物理学进展，2018，33(5)：1911-1918.

[9] 唐何兵，刘传奇，韦红，等. 基于支持向量机的声波曲线预测在水平井随钻深度预测中的应用[J]. 物探与化探，2017，41(2)：256-261.

[10] 闫星宇，顾汉明，肖逸飞，等. XGBoost 算法在致密砂岩气储层测井解释中的应用[J]. 石油地球物理勘探，2019，54(2)：447-455，241.

[11] Aurelien Geron. 机器学习实践：基于 Scikit-Learn 和 TensorFlow[M]. 北京：机械工业出版社，2018：177-180.

[12] Gavin Hackeling. Scikit-learn 机器学习[M]. 北京：人民邮电出版社，2019：86-88.

[13] 段仁春. 运用气测资料识别水淹层方法探讨[J]. 录井工程，2008，19(1)：48-51.

[14] 黄保纲，宋洪亮，申春生，等. 利用气测资料判断调整井油层水淹程度的尝试[J]. 中国海上油气，2011，23(3)：170-174.

Reconstruction Method of Logging Curves Based on Extreme Gradient Boosting Method

QI Chunsheng, DING Lei, WANG Yi, ZHENG Zhifeng, ZHANG Xi

(Hainan Branch of CNOOC Ltd., Haikou 570311, China)

Abstract: In the process of well logging data acquisition in the western area of South China Sea, the lack of well logging curve often occurs, which makes it difficult to accurately evaluate the well logging interpretation of oil and gas reservoirs. In order to solve this problem, Based on XGBoost algorithm, the logging curve reconstruction is realized by comprehensively using logging and gas logging data. The method could support the processing of sparse data, and realize the automatic optimization of model parameters through grid search. By comparing the different machine learning algorithms and the reconstruction effect of different logging curves in the target area, the new method has the advantages of low parameter adjustment difficulty, high calculation accuracy, and it can effectively improve the work efficiency. Furthermore, the method is applied to the identification of low resistivity water flooded zones in the western South China Sea Oilfield, and good application result is obtained.

Keywords: XGBoost; reconstruct log curve; sparse data; grid search; water flooded zone

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net