文章编号：1001-1749(2024)02-0146-08

# 基于 BO-XGBoost 优化方法的砂岩厚度预测方法研究与应用

刘烜良[1]，张军华[1]，白青林[2]，王福金[2]，刘中伟[2]，焦红岩[2]
(1. 中国石油大学(华东)地球科学与技术学院，青岛 266580；
2. 中国石化胜利油田分公司现河采油厂，东营 257068)

**摘 要**：河道砂体预测在油气勘探中具有十分重要的作用，但在实际勘探开发中，复杂河道砂体地层埋深及河道本身结构的复杂性会使其井震关系较差，进而导致砂岩预测精度较低。针对该问题，笔者利用地震多属性信息，发挥密井网优势，利用 BO-XGBoost 砂岩厚度预测方法，在验证集占比为 25%时取得最佳的预测效果，且预测效果要好于常规 SVM 和 XGBoost 方法。研究方法对同类复杂储层的砂岩厚度预测，有借鉴意义。

**关键词**：贝叶斯优化；XGBoost；砂体；厚度预测
**中图分类号**：P631.4 **文献标识码**：A **DOI**：10.3969/j.issn.1001-1749.2024.02.03

## 0 引言

河道砂体由于地质特征明显，很受油田勘探开发人员的关注。但对复杂河道砂体，由于河道宽度小，叠置、交叉严重，垂向厚度薄、互层多等原因，导致其井震关系较差，地震属性数据中往往没有明显的河道特征，即使是在密井网工区，其储层预测依然有很大的难度。

XGBoost(extreme gradient boosting，极限梯度提升)，自 2016 年由 Chen 等[1]提出后便受到广泛关注。2021 年，谷宇峰等[2]利用粒子群算法(Particle Swarm Optimization，PSO)对 XGBoost 进行优化，并将其应用于渗透率预测，通过与其他算法对比发现，PSO-XGBoost 适用性更强，预测精度更高，稳定性最好。2022 年，张家臣等[3]将 XGBoost 应用于测井曲线重构，通过 K 折交叉验证将 XGBoost 模型性能与梯度提升决策树、随机森林和全连接神经网络三种方法进行对比，验证结果表明，基于 XGBoost 的测井曲线重构方法在准确性和稳定性方面都取得了更好的效果，并且表现出较强的泛化能力。2022 年，邢强等[4]将 XGBoost 应用到储层类型识别中，提高了识别精度与识别效率。2021 年，丁阳阳等[5]将 XGBoost 应用于煤层识别领域，利用 XGBoost 机器学习方法，实现了多种测井信息融合，有效地提高了煤体结构的识别精度和识别效率。2019 年，闫星宇[6]将 XGBoost 应用于致密砂岩气储层参数预测中，与随机森林方法和支持向量机算法进行比较发现，XGBoost 算法能准确地预测孔隙度、渗透率并对致密砂岩气层进行有效识别且效果由于另外两种算法。此外，XGBoost 在交通，医学，环境科学，计算机科学等领域也有着广泛应用[7-9]。

由于该算法参数较多，通常需要优化算法对其进行优化，常用的优化算法主要有网格搜索(Grid Search)，贝叶斯优化算法(Bayesian Optimization，BO)和群体优化算法(如粒子群算法，遗传算法)。针对砂岩厚度预测问题，网格搜索效率相对较低，对于群体优化算法而言样本过少，所以选取贝叶斯优化算法对 XGBoost 进行超参数优化。已有学者将 BO-XGBoost 应用在电子测量，隧道工程等领域[10-14]，但在储层预测领域应用较少。这里将 BO-XGBoost 应用于密井网工区的复杂砂体预测，并讨论了验证集占比对预测模型的影响，具有一定的参考价值。

收稿日期：2022-09-14
基金项目：国家自然科学基金项目(42072169)；胜利油田项目(30200003-21-ZC0631)
第一作者：刘烜良(1997—)，男，硕士，主要从事储层预测工作，E-mail：306750796@qq.com。

# 1　方法原理

## 1.1　XGBoost 方法原理

XGBoost 属于提升类算法，通过将多个弱分类器的结果组合起来，变为强分类器输出最终结果。砂岩厚度预测属于回归问题，其回归预测值表达式为：

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

其中 $\hat{y}_i$ 为最终预测值，$x_i$ 为输入样本，$f_k(x_i)$为第 $k$ 棵树所得预测值。通过运用集成树原理，将所有决策树预测值相加得到最终预测值。XGBoost 在增加决策树的同时保存上一轮迭代的预测值以此来提高模型的精度。迭代函数表达式为：

$$\hat{y}_i^{0}=0 \tag{2}$$

$$\hat{y}_i^{(1)}=f_1(x_i)=\hat{y}_i^{(0)}+f_1(x_i) \tag{3}$$

$$\hat{y}_i^{(2)}=f_1(x_i)+f_2(x_i)=\hat{y}_i^{(1)}+f_2(x_i) \tag{4}$$

……

$$\hat{y}_i^{(t)}=\sum_{k=1}^{t}f_k(x_i)=\hat{y}_i^{(t-1)}+f_t(x_i) \tag{5}$$

其中 $\hat{y}_i^{(t)}$ 为前轮迭代模型输出预测值之和，$\hat{y}_i^{(t-1)}$ 为当前 $t-1$ 轮迭代模型预测值之和，$f_t(x_i)$为第 $t$ 棵树取值。引入正则项：

$$\Omega(f_t)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 \tag{6}$$

其中 $\gamma$ 为惩罚力度，$T$ 为叶子节点数量，$\omega_j$ 为当前叶子节点权重，$\lambda$ 为叶子权重正则化参数。正则项决定了树的复杂程度，它可以限制树模型的叶子数量，从而避免模型过拟合，提高泛化能力。损失函数定义为：

$$l(y_i,\hat{y}_i)=(y_i-\hat{y}_i)^2 \tag{7}$$

通过最小化损失函数来构建预测模型，将正则项与树模型损失函数结合构成 XGBoost 目标函数：

$$Obj^{(t)}=\sum_{i=1}^{n}l(y_i,\hat{y}_i^{(t-1)}+f_t(x_i))+\Omega(f_t)+C \tag{8}$$

其中 $C$ 为常数。对目标函数做二阶泰勒展开：

$$Obj^{(t)}=\sum_{i=1}^{n}\{l(y_i,\hat{y}_i^{(t-1)})+g_if_t(x_i)+\frac{1}{2}h_if_t^2(x_i)\}+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2+C \tag{9}$$

其中 $g_i=\partial_{\hat{y}^{(t-1)}i}l(y_i,\hat{y}_i^{(t-1)})$，$h_i=\partial^2_{\hat{y}^{(t-1)}}l(y_i,\hat{y}_i^{(t-1)})$，分别表示预测误差对当前模型的一阶导和二阶导。对 $\omega_j$ 进行一阶推导，得到 XGBoost 最优目标函数：

$$Obj=-\frac{1}{2}\sum_{j=1}^{T}\frac{G2_j}{H_j+\lambda}+\lambda T \tag{10}$$

上式也称为打分函数(scoring function)，它是衡量树结构好坏的标准，值越小代表树的结构越好。对(9)式求极值得到最优解为：

$$\omega_j^{*}=-\frac{G_j}{H_j+\lambda} \tag{11}$$

式中 $G_j=\sum_{i\in I_j}g_i$，表示映射为叶子节点 $j$ 的所有输入样本的一阶导之和，$H_j=\sum_{i\in I_j}h_i$，表示映射为叶子节点 $j$ 的所有输入样本的二阶导之和，$I_j$ 表示叶子节点样本集。

## 1.2　贝叶斯优化方法原理

贝叶斯优化是一种基于概率分布的超参数优化算法，其主要核心部分为先验函数和采集函数。先验函数采用高斯回归(GP，Gaussian Process)，并利用 GP 将先验概率模型转化为后验概率分布。后验函数采用最大期望提升(probability of improvement，PI)。

GP 可以看作一个函数，将所要优化的 XGBoost 超参数看作输入 $x$，输出为高斯分布的均值和方差。输入与输出之间的关系由均值函数和协方差函数(核函数)决定，即：

$$f(x)\sim GP(m(x),k(x,x')) \tag{12}$$

式中，$m(x)=E[f(x)]$，$k(x,x')=E[(f(x)-m(x))(f(x')-m(x'))]$。$m(x)$为均值函数，$k(x,x')$为协方差函数。$f(x)$为目标函数。高斯过程的每个 $x$ 都有一个对应的高斯分布，对于 XGBoost 超参数 $X=(x_1,x_2\cdots x_t)$即 $D=(X,y)$，$y=\{f(x_1),f(x_2)\cdots f(x_t)\}$，则存在一个联合高斯分布，可表示为：

$$\begin{bmatrix}f(x_1)\\ \vdots\\ f(x_t)\end{bmatrix}:N\left(0,\begin{bmatrix}k(x_1,x_1) & \cdots & k(x_1,x_t)\\ \vdots & \ddots & \vdots\\ k(x_t,x_1) & \cdots & k(x_t,x_t)\end{bmatrix}\right) \tag{13}$$

若加入新的样本 $x_{t+1}$ 并更新协方差矩阵，记为 $\boldsymbol{K}$，则联合高斯分布可表示为：

$$\begin{pmatrix}f_{1:t}\\ f_{t+1}\end{pmatrix}\sim N(0,\begin{bmatrix}K & k\\ kT & k(x_{t+1},x_{t+1})\end{bmatrix}) \tag{14}$$

其中 $\boldsymbol{k}=[k(x_{t+1},x_1)k(x_{t+1},x_2)\cdots k(x_{t+1},x_t)]$，

进一步可以得到 $f_{t+1}$ 的后验概率分布，其表达式为：

$$P(f_{t+1}|D_{1:t},x_{t+1})\sim N(u,\sigma^2) \tag{15}$$

$$u=\boldsymbol{k}^T\boldsymbol{K}^{-1}f_{1:t} \tag{16}$$

$$\sigma^2=k(x_{t+1},x_{t+1})-\boldsymbol{k}^T\boldsymbol{K}^{-1}\boldsymbol{k} \tag{17}$$

采集函数(PI)表达式为：

$$f_{PI}(x)=\Phi\left(\frac{u(x)-f(x^+)-\xi}{\sigma(x)}\right) \tag{18}$$

式中 $\Phi$ 为正态分布积累密度函数，$u(x)$，$\sigma(x)$分别是后验模型得到的目标函数值的均值和方差，$f(x^+)$为当前最佳目标函数值，$\xi$ 为大于等于 0 的可调参数。

### 1.3 BO-XGBoost 储层预测流程的设计

首先对提取出的地震属性进行优选，将优选后的地震属性进行剔除异常值，归一化等预处理；然后将预处理后的井点处的属性与井点处的砂岩厚度组成数据集来训练砂岩厚度预测模型，利用贝叶斯优化并引入交叉验证来优化 XGBoost 超参数，对整个研究区的砂岩厚度进行预测。对比参数优化前后 XGBoost 的预测效果。基于 BO-XGBoost 模型砂岩厚度预测流程如图 1 所示。

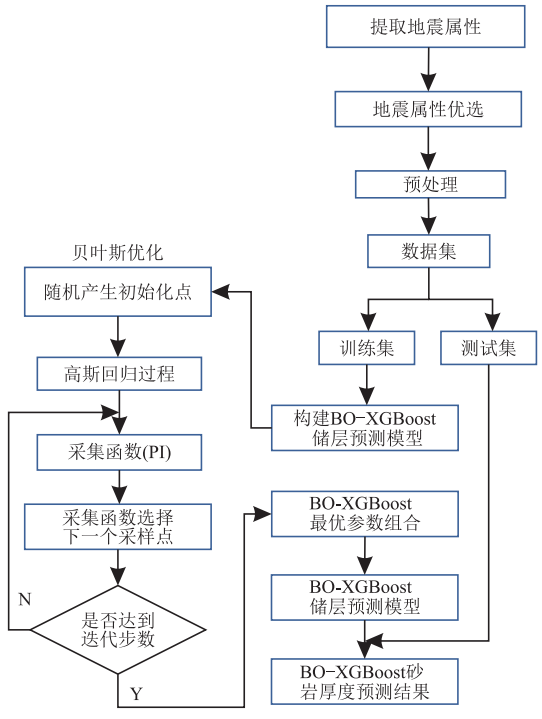


**图 1** BO- XGBoost 模型砂岩厚度预测流程

Fig. 1 Sandstonethickness prediction process of BO- XGBoost model

## 2 BO-XGBoost 储层预测模型

### 2.1 研究区概况及地震属性提取

这里以胜利油田通 61 老油区为研究实例。研究区位于东营南坡王家岗鼻状构造带，主力储层为沙二上断控水下分流河道。该块区自 1974 年投入开发，目前已进入高勘探开发成熟期。研究区在东部老油区有一定的代表性，具有典型的地质与地球物理特点。工区采用密井网，有各类井 89 口，部分井井距小于 200 m。地震属性是描述复杂河道砂体的重要手段，根据地震属性的地质意义和以往的经验，本文提取弧长、平均振幅等 11 种地震属性，利用基于核相似性度量的特征选择方法，优选出弧长、平均振幅、平均能量、带宽、主频、能量半时、正振幅和、过零点个数 8 种地震属性，这 8 种地震属性的地质含义是明确的：弧长属性对于阻抗差较大的储层效果较好；平均振幅属性可以有效识别振幅异常，层序特征以及岩性信息；平均能量可以用于评价地震波能量；带宽与地震资料的品质相关性较大；主频属性可以用于识别岩性和流体变化信息；能量半时属性反映了分析时窗内能量相对变化关系，能够指示沉积环境与岩性岩相变化；正振幅和属性可以用于识别特殊岩性类型；过零点个数属性能够较好地描述地震波振幅过零点的个数，包含了岩性、流体等信息。图 2 和图 3 分别为 8 种地震属性的井震关系图和井点厚度与地震属性的 Pearson 相关系数图，可以看出整个研究区的井震关系较差，与井点厚度相关性最好的平均振幅属性相关系数也仅有 0.11。且单一属性预测精度较低，希望采用地震多属性信息，发挥密井网优势，利用 BO-XGBoost 方法提高预测精度。

### 2.2 验证集占比研究

训练集和验证集如何设置，它们对机器学习储层预测有什么样的影响，是一项很有理论研究价值的工作。一般来说，训练集的占比越高，相应模型的预测精度也会越高，但验证集过少，会使利用的井数据减少，失去了预测的价值。为了搞清楚这一问题，这里用 XGBoost 来进行试验，验证集占比由 10%逐渐增加到 90%，取值间隔为 5%，平均绝对误差大小作为评价标准。经实验，当验证集占比为 25%时，验证集预测绝对误差最小，为 1.53 m。图 4 给出了

部分不同验证集占比的预测结果，可以看到：①当验证集占比较小(25%)时，大多数井预测结果都比较好；②随着验证集占比的提高，验证井增加、训练井减少，实际厚度与预测厚度井点分布逐渐分散，到75%已比较分散，且有两个井的误差较大。所以我们将验证集占比固定为 25%。

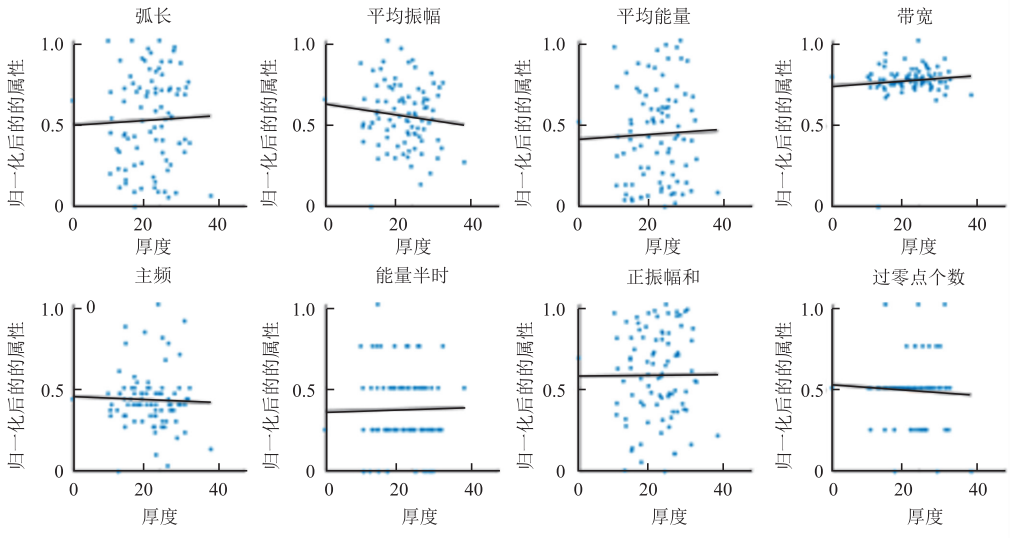


**图 2**　井震关系图

Fig. 2　Well earthquake relationship


**图 3**　Pearson 相关系数图

Fig. 3　Pearson correlation coefficient

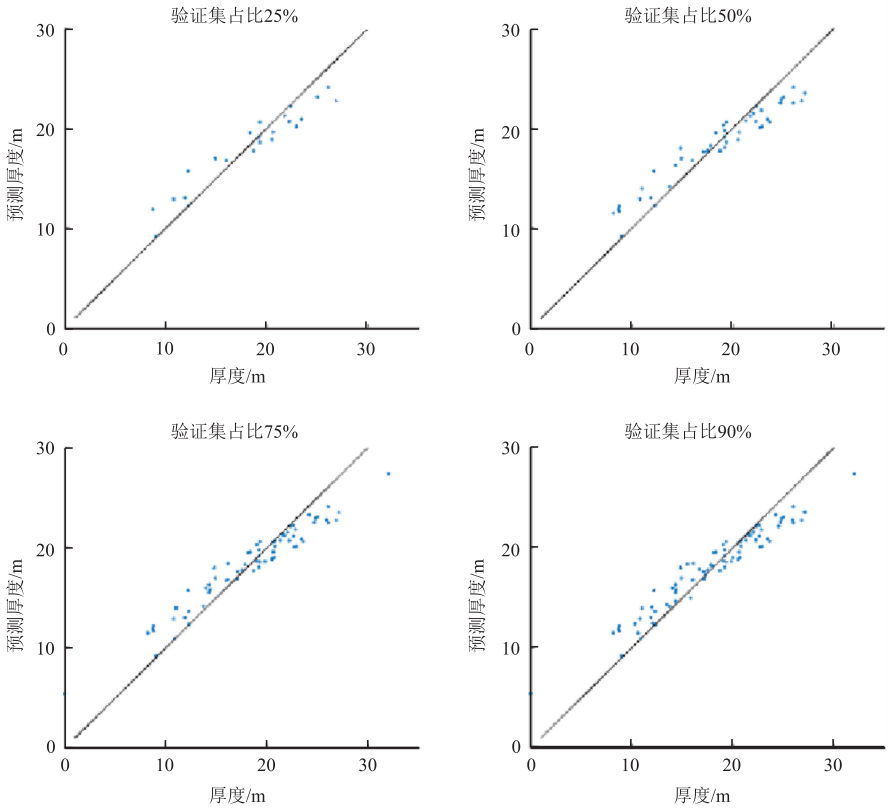


**图 4** 验证集占比与预测精度散点分析图

Fig. 4 Scatter analysis of validation set proportion and prediction accuracy

### 2.3 BO-XGBoost 储层预测

首先随机选取三口井作为检验井，将剩余 86 口井用于模型训练，验证集占比固定为 25%，通过多次测试，确定 XGBoost 超参数组合大体寻优范围，然后利用贝叶斯优化辅以交叉验证对 XGBoost 算法进行超参数寻优，主要调节参数的含义，参数寻优范围及最终寻优结果如表 1 所示。利用优化后的超参数组合对测试集进行预测，并与优化前 XGBoost 预测结果和常用 SVM 预测结果进行性能对比，对比结果如表 2 所示。

**表 1** XGBoost 主要参数

Tab. 1 Main parameters of XGBoost

| XGBoost 参数 | 参数含义 | 参数取值范围 | 优化前参数 | 最终参数 |
|---|---|---|---|---|
| n_estimators | 决策树数量 | 100－200 | 150 | 189 |
| max_depth | 树的最大深度 | 4－10 | 7 | 9 |
| min_child_weight | 最小叶子结点权重 | 1－10 | 3 | 1 |
| learning_rate | 学习率 | 0.05－0.5 | 0.04 | 0.053 |
| subsample | 样本采样比例 | 0.5－1 | 0.75 | 0.856 |
| Colsample_bytree | 特征随机采样比例 | 0.5－1 | 0.5 | 0.55 |
| gamma | 节点分裂折减系数 | 0－10 | 6 | 4.3 |
| alpha | 正则系数 | 0－10 | 3 | 3.556 |

从预测结果可以看出，BO-XGBoost 的 RMSE 最小，相应的检验井绝对误差也最小，预测精度最高，SVM 的 RMSE 最大，检验井误差也相应最大，预测精度相对较低，预测效果 BO-XGBoost＞XGBoost＞SVM。图 5 为三种方法对整个研究区砂岩厚度预测结果，图 6 为研究区平均振幅属性图。通过对比可以发

现，平均振幅属性中显示的砂岩位置与预测结果中砂岩厚度分布差异较大，这一定程度上表明研究区井震关系很差；对比三种方法的预测结果，从黑色圆圈处可以看出，SVM 预测结果相对较差，XGBoost 与 BO-XGBoost 预测结果更加精细且砂岩展布规律相差不大，但从检验井预测误差来看，BO-XGBoost 效果更佳。

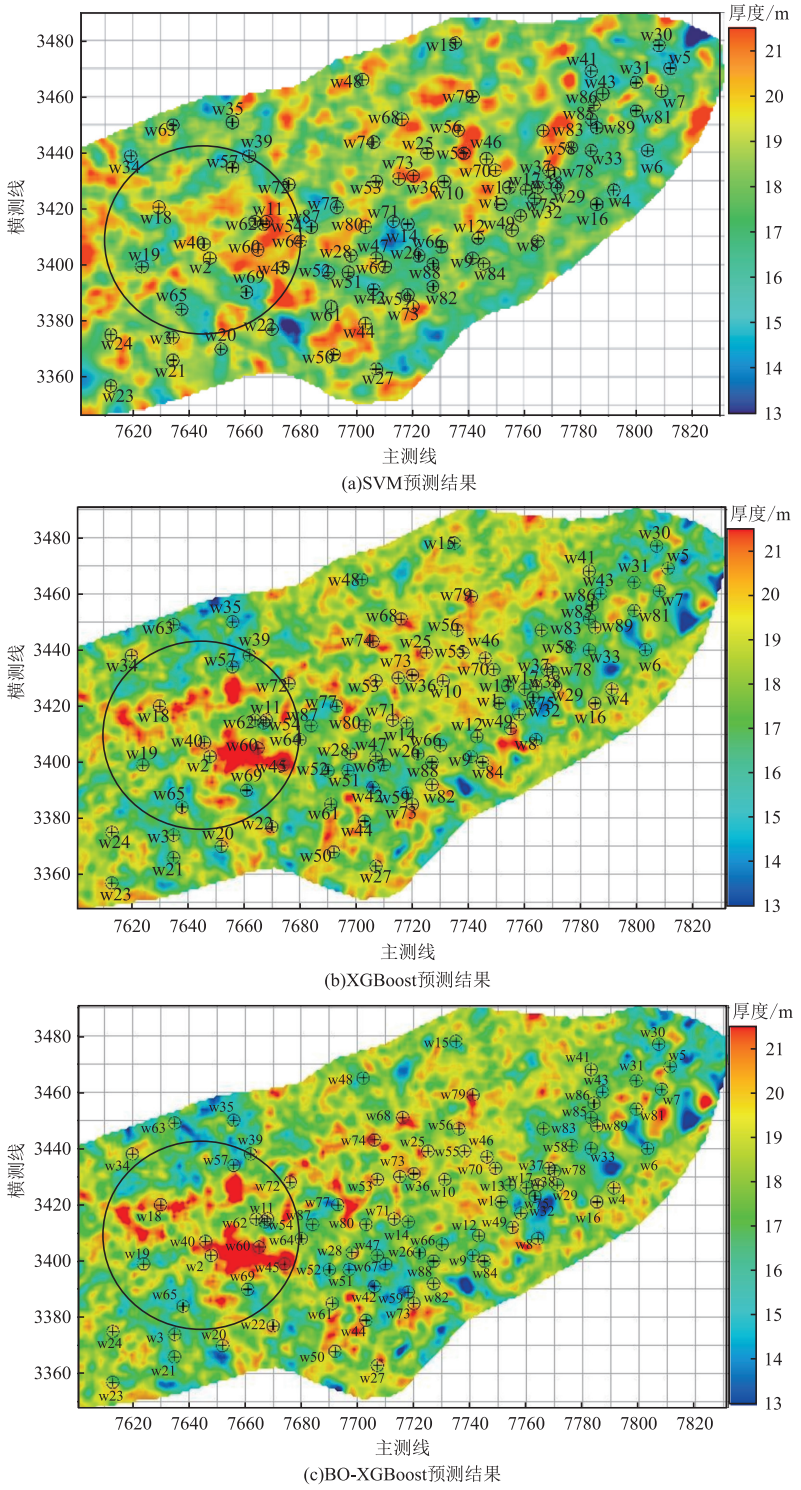


**图 5**　研究区砂岩厚度预测图

Fig. 5　Prediction of sandstone thickness in the study area

**表 2** 预测结果性能对比

Tab. 2 Performance comparison of prediction results

| 模型 | RMSE | 井名 | 预测厚度/m | 实际厚度/m | 预测绝对误差/m |
|---|---|---|---|---|---|
| SVM | 3.376 | W25 | 18.46 | 20.5 | 2.04 |
| | | W32 | 17.54 | 14.3 | 3.24 |
| | | W84 | 18.70 | 22.2 | 3.50 |
| XGBoost | 1.843 | W25 | 18.76 | 20.5 | 1.74 |
| | | W32 | 16.85 | 14.3 | 2.55 |
| | | W84 | 19.84 | 22.2 | 2.36 |
| BO-XGBoost | 1.572 | W25 | 19.08 | 20.5 | 1.42 |
| | | W32 | 16.08 | 14.3 | 1.78 |
| | | W84 | 20.88 | 22.2 | 1.32 |

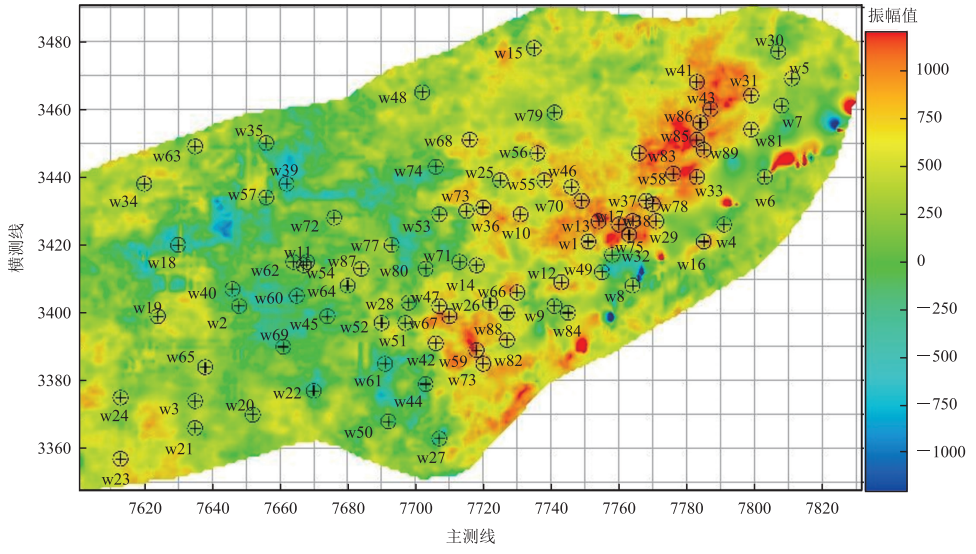


**图 6** 平均振幅属性

Fig. 6 Average amplitude attribute

## 3 结论

基于 BO-XGBoost 的砂岩厚度预测方法，利用贝叶斯优化 XGBoost 参数，再辅以交叉验证，结合地震多属性信息和密井网优势建立预测模型，在井震关系较差的情况下，实现了对研究区砂岩厚度的预测，且预测精度比 SVM 和未优化的 XGBoost 方法更高，预测的砂体分布形态也与研究区砂岩厚度图分布规律相一致，值得在储层预测中推广。

**参考文献：**

[1] CHEN T Q, GUESTRIN C. XGBoost: A scalable tree boosting system [C] //Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco, USA, 2016: 785- 794.

[2] 谷宇峰，张道勇，鲍志东. 测井资料 PSO-XGBoost 渗透率预测[J]. 石油地球物理勘探，2021，56(1)：26-37.
GU Y F, ZHANG D Y, BAO Z D. Premeability prediction using PSO-XGBoost based on logging data[J]. Oil Geophysical Prospecting, 2021, 56(1): 26-37. (In Chinese)

[3] 张家臣，邓金根，谭强，等. 基于 XGBoost 的测井曲线重构方法[J]. 石油地球物理勘探，2022，57(3)：697-705.
ZHANG J C, DENG J G, TAN Q, et al. Reconstruction of well logs based on XGBoost [J]. Oil Geophysical Prospecting, 2022, 57(3): 697-705. (In Chinese)

[4] 邢强，张晋言，王镇方，等. 基于 XGBoost 的测井解释规则库自动获取方法[J]. 石油物探，2022，61(2)：356-363.
XING Q, ZHANG J Y, WANG Z F, et al. Automatic acquisition of a rule base for logging interpretation u-

sing the XGBoost algorithm[J]. Geophysical Prospecting for Petroleum, 2022,61(2): 356-363. (In Chinese)

[5] 丁阳阳，赵军龙，李兆明，等. 基于 XGBoost 算法的煤体结构测井识别技术研究[J]. 地球物理学进展，2022，37(3)：998-1006.
DING Y Y, ZHAO J L, LI Z M, et al. Research on logging recognition technology of coal structure based on XGBoost algorithm. [J] Progress in Geophysics , 2022,37(3):998-1006. (In Chinese)

[6] 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.
YAN X Y, GU H M, XIAO Y F, et al. XGBoost algorithm applied in the interpretation of tight-sand gas reservoir on well logging data [J]. Oil Geophysical Prospecting, 2019,56(2): 447-455. (In Chinese)

[7] 陆家兴，陈明，秦玉芳，等. 基于 LINCS-L1000 扰动信号通过 SAE-XGBoost 算法预测药物诱导下的细胞活性[J]. 生物工程学报，2021，37(4)：1346-1359.
LU J X, CHEN M, QIN Y F, et al. Prediction of drug-induced cell viability by SAE-XGBoost algorithm based on LINCS-L1000 perturbation signal[J]. Chinese Journal of Biotechnology. 2021, 37(4): 1346-1359. (In Chinese)

[8] BIN YU, WENYING QIU, CHENG CHEN, et al. SubMito-XGBoost: predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting[J]. Bioinformatics, 2020, 36(4), 1074-1081.

[9] D. ZHANG, L. QIAN, B. MAO, et al. A data-driven design for fault detection of wind turbines using random forests and XGBoost[J]. IEEE Access, 2018 (6), 21020-21031.

[10] 何军，林广东，申小军，等. 基于贝叶斯优化 XGBoost 的隧道沉降量预测[J]. 计算机系统应用，2022，31(7)：379-385.
HE J, LIN G D, SHEN X J, et al. Prediction of tunnel subsidence based on bayes optimized XGBoost[J]. Computer Systems & Applications, 2022, 31(7): 379-385. (In Chinese)

[11] 孙斌，储芳芳，陈小惠. 基于贝叶斯优化 XGBoost 的无创血压预测方法[J]. 电子测量技术，2022，45(7)：68-74.
SUN B, CHU F F, CHEN X H. Non-invasive blood pressure detection method based on Bayesian optimization XGBoost[J]. Electronic Measurement Technology. 2022, 45(7): 68-74. (In Chinese)

[12] 黄琦，沈建国，蒋敏兰，等. 基于高光谱成像技术与 BO-XGBoost 的花生含水率无损检测研究[J/OL]. 中国粮油学报：1-10[2022-09-13].
HUANG Q, SHEN J G, JIANG M L, et al. Nondestructive detection of peanut moisture content based on near infrared hyperspectral technique and BO-XGBoost [J/OL]. Journal of the Chinese Cereals and Oils Association: 1-10[2022-09-13]. (In Chinese)

[13] 顾天下，刘勤明，叶春明. 基于 BO-XGBoost 与集成学习方法的供应链金融信用评价研究[J/OL]. 上海理工大学学报：1-8[2022-09-13].
GU T X, LIU Q M, YE C M. Credit evaluation of supply chain finance based on BO-XGBoost and ensemble learning method[J/OL]. Journal of University of Shanghai for Science and Technology: 1-8[2022-09-13]. (In Chinese)

[14] 王媛彬，李媛媛，韩骞，等. 基于 PCA-BO-XGBoost 的矿井回采工作面瓦斯涌出量预测[J]. 西安科技大学学报，2022，42(2)：371-379.
WANG Y B, LI Y Y, HAN Q, et al. Gas emission prediction of the stope in coal mine based on PCA-BO-XGBoost[J]. Journal of Xi'an University of Science and Technology, 2022, 42(2) : 371－379. (In Chinese)

## Research and application of sandstone thickness prediction method based on BO-XGBoost optimization method

LIU Xuanliang[1], ZHANG Junhua[1], BAI Qinglin[2], WANG Fujin[2], LIU Zhongwei[2], JIAO Hongyan[2]

(1. School of Geosciences, China University of Petroleum (East China), Qingdao 266580, China;
2. Xianhe oil production plant, Shengli Oilfield Company, SINOPEC, Dongying, 257068, China)

**Abstract**: Channel sand body prediction plays a vital role in oil and gas exploration, but in actual exploration and development, the buried depth of complex channel sand body and the complexity of channel structure will make its well seismic relationship poor, leading to low sandstone prediction accuracy. Given this problem, this paper uses the seismic multi-attribute information, gives full play to the advantages of the dense well pattern, and uses the BO-XGBoost sandstone thickness prediction method to achieve the best prediction effect when the validation set proportion is 25%. The prediction effect is better than the conventional SVM and XGBoost methods. The research method can be used for reference in sandstone thickness prediction of similar complex reservoirs.

**Keywords**: bayesian optimization; XGBoost; sand body; thickness prediction