

利用混合模型 CRBM-PSO-XGBoost 识别致密砂岩储层岩性

谷宇峰¹, 张道勇¹, 鲍志东²

[1. 自然资源部 油气资源战略研究中心, 北京 100034; 2. 中国石油大学(北京) 地球科学学院, 北京 102249]

摘要: 识别储层岩性是开展地层对比和沉积展布等地质基础研究工作的重要前提。致密砂岩储层多为砂、泥岩薄互层, 形成的岩性不仅种类较多, 而且大部分岩性的测井响应也较为相似, 为相关的岩性识别工作带来了困难。基于机器学习技术在模式识别上具有强大的分析性能, 选用了稳定性好且计算效率高的 XGBoost 模型来解决致密砂岩储层岩性识别问题。该模型在建模过程中需要较多的经验参数参与, 且计算速度随着自变量的增加而逐渐降低, 为此提出了采用 CRBM 模型和 PSO 模型对其进行改进。CRBM 模型具备数据提取功能, 可从源数据中挖掘出更少但更利于建模的新自变量, 而 PSO 模型可通过迭代计算确定 XGBoost 所有经验参数的最优值。以姬塬油田西部长 4+5 段部分取心井资料为基础, 通过设计两个实验来验证提出的混合模型的预测能力。为加强验证效果, 在实验中加入 PNN 和 SVM 的优化模型进行对比。实验后发现所提出模型的预测准确率最高, 均在 90% 以上。实验结果表明所提出的模型不仅能有效解决致密砂岩储层岩性识别问题, 且较以往经典预测模型更具推广应用性。

关键词: XGBoost 模型; CRBM 模型; PSO 模型; 岩性识别; 致密砂岩储层; 姬塬油田; 鄂尔多斯盆地

中图分类号: TE19 **文献标识码:** A

Lithology identification in tight sandstone reservoirs using CRBM-PSO-XGBoost

Gu Yufeng¹, Zhang Daoyong¹, Bao Zhidong²

[1. Strategic Research Center of Oil and Gas Resources, Ministry of Natural Resources, Beijing 100034, China;
2. College of Geosciences, China University of Petroleum (Beijing), Beijing 102249, China]

Abstract: Lithology identification is universally regarded as a prerequisite for some fundamental geological work such as stratigraphic correlation and analysis of sedimentary system distribution. The tight sandstone reservoirs are generally composed of multiple thin sandstone-mud alternating beds, featuring a variety of lithological types, most of which are similar in log responses, posing great difficulty for lithology identification. Given the powerful analytical performance of machine learning techniques shown on pattern recognition, XGBoost, a stable and highly efficient model is adopted to predict lithology of the tight sandstone reservoirs. However, this model needs many empirical parameters to complete its prediction, and its computing efficiency will be reduced when it deals with more independent variables. The CRBM and PSO models are, therefore, introduced to improve the prediction performance. CRBM is functional to extract features from original data so that fewer but more significant independent variables can be created by the CRBM in modeling. While PSO is effective to determine the optimal values via iterative computation for those empirical parameters used by XGBoost. Based on the data derived from some cored wells of Chang 4 + 5 Members in western Jiyuan oilfield, two experiments are designed to validate the prediction performance of the hybrid model proposed, CRBM-PSO-XGBoost. Furthermore, two optimized models formed by PNN and SVM are selected to make a comparison, as a way to consolidate the validation effect. The experiment results manifest that the prediction accuracy comes to peak by the proposed model, up to over 90%. In all, the hybrid model proposed in the study, capable of identifying lithology of tight sandstone reservoirs, is of wider application to lithology prediction compared with previous models.

Key words: XGBoost model, CRBM model, PSO model, lithology identification, tight sandstone reservoir, Jiyuan oilfield, Ordos Basin

收稿日期: 2020-08-12; 修订日期: 2021-08-12。

第一作者简介: 谷宇峰(1988—), 男, 助理研究员, 储层表征和测井解释技术。E-mail: aaaaa3377@126.com。

准确获取储层岩性数据对于开展地层对比、沉积展布分析和地质建模等地质基础工作至关重要,因此岩性识别一直被视为是一项关键研究内容^[1-6]。经典的岩性识别工具为交会图。交会图一般是由2种或3种测井曲线构成的,而选取的曲线需要对各种所需识别的岩性都能有独特的响应范围^[4-6]。故而,当测井数据点散落在交会图中,各个岩性对应的测井响应划分条件便可通过观察来明确。对于砂质较纯或者岩性成分简单的储层,该工具能够有效地解决其岩性识别问题,但随着油气勘探范围的扩大,非常规油气逐渐成为了油气研究方向上的核心对象,使得更多关于非常规储层的岩性识别问题被提出。由于非常规储层岩性成分复杂,且多数岩性在测井曲线上具有相似的响应特征,导致以交会图为代表的经典识别工具难以适用,为此众多地球物理学家发展了以机器学习技术为主的岩性识别方法^[7-15]。

机器学习是人工智能的一个重要分支,其主要计算原理是在对学习数据分析的基础之上,建立自变量与因变量的线性或非线性网络映射关系,之后根据该关系完成对预测数据的处理。在模式识别领域,目前得到广泛应用的模型有 KNN (K-nearest neighbors), PNN (probabilistic neural network) 和 SVM (support vector machine) 等^[7-15]。KNN 模型是聚类分析中的代表,主要是依据预测数据点与学习数据点之间距离的远近进行模式判断。由于不需要对学习数据进行训练,所以 KNN 模型计算效率较高,而且因为预测数据点只会被聚类到距离最近的学习数据点的模式中,所以即使学习样本中含有少量错误样点,该模型仍能够进行有效的判断,体现出其容错能力。赵彤彤等(2018)使用了一种基于模糊熵的 KNN 模型进行了岩性识别研究,并取得了较好的识别效果^[7]。张梓童等(2019)分析了 KNN 模型的计算原理,并通过验证明确了该模型在岩性识别上是一种有效手段^[8]。PNN 模型是一种基于概率密度分析模式的识别模型,其主要思想是先利用学习样本建立各个模式的概率密度分布,之后根据预测样本在分布中的概率值判断归属。由于该模型采用了概率分析,较 KNN 模型在学习样本处理上有更高的容错性,而且概率密度分布的建立无需训练,所以理论上其预测效率也较高。赵杰和李春华(2009)在岩性识别中尝试了 PNN 模型,并取得了不错的识别效果^[10]。陈刚(2018)以随钻测井资料为基础,利用 PNN 模型对煤层岩性进行了识别,验证了模型的有效性^[11]。SVM 模型是监督学习中的代表,其主要计算思想是先通过训练找到影响模式判断的最重要

的学习样本(称为支持向量),之后依据这些样本完成数据预测。由此可见,该模型的预测效率和预测效果完全取决于支持向量的数量和质量,而不是通过全部样本学习的,这就大大提升了模型的容错能力和计算效率。李政宏等(2020)分析了机器学习技术在岩性识别中的重要性,并证明了 SVM 模型是一种有效识别模型^[13]。根据 SVM 模型的特性,林香亮等(2020)使用 PCA (principal component analysis) 模型对其进行改进,并验证了 PCA-SVM 混合模型在砂砾岩岩性识别上具有良好的应用效果^[14]。虽然上述模型在许多应用案例中都得到了一定程度的认可,但仍难以推广。由于每个预测样本都要和所有学习样本进行计算,所以随着学习样本容量的扩大, KNN 和 PNN 模型的计算效率将会严重降低;随着学习样本维度的增加,样点间的距离值也会随之增大,从而加大了 KNN 模型在模式判断上的不确定性;建模需要多种经验参数参与,如 PNN 模型的概率密度分布窗口长度和 SVM 模型的惩罚系数等,导致预测模型和预测结果难以确保为最优。

梯度提升是一类优秀的模式识别技术,其主要计算原理是先将目标值与计算值之间的差值作为训练对象,再通过一系列 CART (classification and regression tree) 回归树训练将差值减小,最后凭借由这些训练后的回归树组成的强学习器完成预测^[16-19]。XGBoost 模型是这类技术的代表。由于该模型引入了正则化项,并精细化了学习公式,所以在训练过程中大概率地避免了过拟合现象的发生,确保了训练的可靠性;由于融入了并行计算技术,模型的计算效率不会随着训练样本容量的扩大而出现严重衰减的现象。因此,该模型因其训练稳定且计算效率高在岩性识别中也得到了关注^[16-19]。Dev 和 Eden (2019) 利用了该模型进行了岩性识别,并验证了模型的有效性^[18]。闫星宇等(2019)采用 XGBoost 模型研究了渗透率预测和储层评价问题,发现该模型的应用对于测井解释发展具有重要意义^[19]。但需注意, XGBoost 模型在建模过程中也有两点不足:建模需要较多经验参数参与,使得模型状态难以确保为最优;当样本维度变高时,为提高计算效率,模型可通过自身的随机自变量采样技术来实现,但这种随机处理方式也会使得建模后模型状态难以确保为最优。为使 XGBoost 模型在建模后达到最佳状态,本文采用 PSO (particle swarm optimization) 模型和 CRBM (continuous restricted Boltzmann machine) 模型对其进行改进。PSO 模型能有效解决参数优化问题,而 CRBM 模型因具有数据提取功能可从源数据中挖掘出更少且对因变量预测更为重要的新自变量,由此

解决了样本随机降维的问题^[20-25]。至此,本文提出利用 CRBM - PSO - XGBoost 混合模型来解决非常规储层岩性识别问题。下文将对模型的计算原理和预测效果逐一分析。

1 计算原理

在本文中,由于岩性识别问题是利用由测井资料和岩性观察数据建立的模型来解决,因此学习数据集中自变量应由测井曲线构成,而应变量由岩性观察数据构成,可表示为 $A = \{X_{mn}, Y_m\}$, 其中 X_{mn} 为测井数据矩阵,表示有 m 个样本,而每个样本由 n 条曲线构成, Y_m 为岩性观察数据向量,有 m 个样本。岩性观察数据在程序中为字符信息,难以应用,为此采用 one-hot coding(独热编码)技术进行编码^[16-19]。例如,岩性观察数据为细砂岩,与之对应设定的原始编码为 2,而识别的岩性共有 5 种,则最终采用的编码是一个长度为 5 且第二个元素为 1 的零向量,可表示为 $[0, 1, 0, 0, 0]$ 。所以, Y_m 在编码后可进一步变为 Y_{mK} , 其中 K 为识别岩性种类,也是每个样本的长度。此时,因变量的样本可表示为 $y_i = (y_{i1}, y_{i2}, \dots, y_{iK})$ 。XGBoost 模型采用 CART 回归树进行迭代训练,并在训练后形成一个预测模型。预测模型称为强分类器,其表达式如下^[16]:

$$F_k(\mathbf{x}_i) = \sum_{d=1}^D \sum_{j=1}^J w_{jk,d} \eta \quad (1)$$

式中: $F_k(\mathbf{x}_i)$ 为作用在样本 $\mathbf{x}_i$ 上的第 k 类强分类器, $i = 1, 2, \dots, m$; $w_{jk,d}$ 为在第 d 次迭代中第 k 棵回归树(即第 k 类回归树)的第 j 个叶节点中所有样本的替代值,无量纲, $k = 1, 2, \dots, K$, $d = 1, 2, \dots, D$, $j = 1, 2, \dots, J$; η 为学习速率,无量纲。

以公式(1)为基础,确定样本 $\mathbf{x}_i$ 被分到第 k 类岩性的概率由下面的 softmax 函数计算^[16-19]:

$$Prob_k(\mathbf{x}_i) = e^{F_k(\mathbf{x}_i)} / \sum_{l=1}^K e^{F_l(\mathbf{x}_i)} \quad (2)$$

式中: $Prob_k(\mathbf{x}_i)$ 表示概率值,无量纲。

在所得的概率值中选择最大值对应的岩性标记为样本 $\mathbf{x}_i$ 的预测岩性。公式(1)中的 $w_{jk,d}$ 由下式确定^[16-19]:

$$w_{jk,d} = -G_{jk,d} / (H_{jk,d} + \lambda) \quad (3)$$

$$G_{jk,d} = \sum_{l=1}^K \partial_{f_{k,d-1}(\mathbf{x}_i)} L[y_{il}, f_{l,d-1}(\mathbf{x}_i)] \quad (4)$$

$$H_{jk,d} = \sum_{l=1}^K \partial_{f_{k,d-1}(\mathbf{x}_i)}^2 L[y_{il}, f_{l,d-1}(\mathbf{x}_i)] \quad (5)$$

式中: $\mathbf{x}_i \in R_{jk,d}$, $R_{jk,d}$ 为在第 d 迭代中第 k 棵回归树的第

j 个叶节点对应的区域; $f_{k,d-1}(\mathbf{x}_i)$ 为在第 $d-1$ 次迭代中作用在样本 $\mathbf{x}_i$ 的第 k 类学习器,其形式以公式(1)为准; L 为损失函数,一般以交叉熵形式为主,在本文中为 $-y_{il} \ln[P_{l,d-1}(\mathbf{x}_i)]$, P 为公式(2)所示的 softmax 函数; λ 为正则化系数,无量纲。

由于 XGBoost 模型在应用前要设置好框架,其中有较多经验参数需要确定,如迭代次数、回归树分裂次数、正则化系数和学习速率等,因此建模后模型难以确保在最优状态。本文采用在多目标最优化问题上计算效率高的 PSO 模型对其进行优化。在执行 PSO 模型之前,先要设定种群^[20-22]。种群包含许多种子,而每个种子由需要优化的参数构成,所以种群可表示为:

$$\Gamma = \{\sigma_i \mid \sigma_i = (\sigma_{i1}, \sigma_{i2}, \dots, \sigma_{iz}), i = 1, 2, \dots, q\} \quad (6)$$

其中, q 为种子数量, σ_i 为第 i 个种子,包含 z 个参数。之后,PSO 模型通过下面的迭代公式计算各参数的最优值^[20-22]:

$$\begin{cases} \Delta_i^j = \{\sigma_i^j \mid \sigma_i^j = \operatorname{argmin}(\|\mathbf{O}_i^j - \mathbf{O}_{\text{std}}\|_2), j = 1, 2, \dots, t\} \\ \Delta_g^j = \{\sigma_g^j \mid \sigma_g^j = \operatorname{argmin}(\|\mathbf{O}_g^j - \mathbf{O}_{\text{std}}\|_2)\} \\ \mathbf{W}_i^{j+1} = \omega \mathbf{W}_i^j + c_1 r_1 (\Delta_i^j - \sigma_i^j) + c_2 r_2 (\Delta_g^j - \sigma_i^j) \\ \sigma_i^{j+1} = \sigma_i^j + \mathbf{W}_i^{j+1} \end{cases} \quad \text{s. t. } \begin{cases} \sigma_i^j \in [\sigma_{\max,1}, \sigma_{\max,2}] \\ \mathbf{W}_i^j \in [-\mathbf{W}_{\max}, \mathbf{W}_{\max}] \end{cases} \quad (7)$$

式中: Δ_i^j 为第 i 个种子在前 j 次迭代中拥有最小计算误差的表达式; Δ_g^j 代表在前 j 次迭代中拥有最小计算误差的种子; σ_i^j 代表在第 j 次迭代中的第 i 个种子; σ_g^j 代表在第 j 次迭代中任意一种子; $\mathbf{O}_i^j$ 为第 i 个种子在第 j 次迭代中的计算结果,无量纲; $\mathbf{O}_g^j$ 代表任意一个种子在第 j 次迭代中的计算结果,无量纲; $\mathbf{O}_{\text{std}}$ 为目标结果,无量纲; t 为迭代次数,无量纲; $\mathbf{W}_i^j$ 为第 i 个种子在第 j 次迭代中的步长,无量纲; ω 为重力系数,无量纲; c_1 与 c_2 为加速系数,无量纲; r_1 和 r_2 为弹性系数,无量纲; $\sigma_{\max,1}$ 与 $\sigma_{\max,2}$ 为种子参数值界限,无量纲; $\mathbf{W}_{\max}$ 为迭代步长最大界限,无量纲; $\|\cdot\|_2$ 为 2-范数; argmin 为求最小值函数。

为计算方便,公式(7)中计算结果和目标结果应以原始编码组成。前人研究结果显示,在迭代前期采用较大的 ω 有利于全局搜索,而到后期采用较小的 ω 有利于局部搜索,为了能够让 PSO 模型高效地进行优化计算,本文采用 LDIW (Linear decreasing inertia weight) 算法使 ω 能够在迭代中自适应地改变^[20-22]。

XGBoost 模型计算效率一般随着样本中自变量个数的增加而降低,因此 Chen 和 Guestrin (2016) 在创造

该模型时为提高其运行速率提出了自变量随机采样算法^[16]。该算法是在考虑自变量较多时,随机选择几个自变量重组学习样本,以让 XGBoost 模型能够通过处理容量更小的学习样本来快速建模。由于自变量是随机选的,难以保证这些自变量都能影响因变量的变化,为此本文提出采用 CRBM 模型方法对源数据进行处理,以实现在源数据降维的同时确保得到的新自变量都为关键变量。CRBM 模型是通过连接可见层和隐含层之间的权重将源数据进行转换,以此实现自变量由多变少的目的^[23-25]。为确保转换质量,CRBM 模型一般要将提取的特征或者称新自变量反转化到可见层中,并将重构的数据与源数据进行对比,此时如果两者之间的误差在允许的范围内,则表明转化是有效的。CRBM 模型的框架一般可表示为^[23-25]:

$$\begin{cases} P_{\theta}(\mathbf{h}_j | \mathbf{V}) = S(\mathbf{V}^T \mathbf{W} + \sigma N) \\ P_{\theta}(\mathbf{v}_i | \mathbf{H}) = S(\mathbf{W} \mathbf{H} + \sigma N) \\ S(\mathbf{x}_i) = \phi_l + (\phi_h - \phi_l) / (1 + e^{-\mu x_i}) \end{cases} \quad (8)$$

式中: P 为概率激活函数,由 S 函数(即 sigmoid 函数)确定; $\mathbf{V}$ 为可见层数据矩阵; $\mathbf{W}$ 为权值矩阵; $\mathbf{H}$ 为隐含层数据矩阵; $\mathbf{v}_i$ 为第 i 个可见层数据向量; $\mathbf{h}_j$ 为第 j 个隐含层数据向量; σ 为设定的噪音方差,无量纲; N 为以标准正态分布为基准的噪音,无量纲; ϕ_h 和 ϕ_l 为 S 函数的上、下渐近线; μ 为噪音控制参数,无量纲,当减小时能够让 S 函数从确定状态变为二值随机分布状态; θ 为迭代参数集,即在迭代中需要确定的参数集。

公式(8)表明源数据可通过第一分式转到隐含层中,而隐含层数据可通过第二分式转回到可见层,用于检验重构数据质量。公式(8)中需要确定的参数有 $\mathbf{W}$ 和 μ ,因此 θ 只包含该两个参数,对应的迭代公式为^[23-25]:

$$\begin{cases} \Delta w_{ij,k} = P_{\theta}(\mathbf{h}_j | \mathbf{V}^0) \mathbf{v}_i^0 - P_{\theta}(\mathbf{h}_j | \mathbf{V}^1) \mathbf{v}_i^1 \\ \Delta \mu_k = [P_{\theta}(\mathbf{h}_j | \mathbf{V}^0)^2 - P_{\theta}(\mathbf{h}_j | \mathbf{V}^1)^2] / \mu_{k-1}^2 \end{cases} \quad (9)$$

式中: $\Delta w_{ij,k}$ 为 $\mathbf{W}$ 第 i 行第 j 列元素在第 k 次迭代中的迭代步长,无量纲; $\Delta \mu_k$ 为在第 k 次迭代中 μ 的迭代步长,无量纲, $k=1,2,\dots,K$;右上角标 0 和 1 分别表示迭代开始前的可见层数据和由隐含层重构得到的可见层数据。

由于 CRBM 模型在每次迭代中都是将所有样本一次性进行处理,这会使其计算效率在样本量较大的情况降低,为此本文提出利用 mini-batch 技术在每次迭代中分批处理样本^[25]。假设 mini-batch 数据集共有 n 个,每个包含 χ 个样本,则公式(9)可改写为^[25]:

$$\begin{cases} \Delta w_{ij,k} = P_{\theta}(\mathbf{h}_j | \mathbf{V}^0) \mathbf{v}_i^0 - P_{\theta}(\mathbf{h}_j | \mathbf{V}^m) \mathbf{v}_i^m \\ \Delta \mu_k = [P_{\theta}(\mathbf{h}_j | \mathbf{V}^0)^2 - P_{\theta}(\mathbf{h}_j | \mathbf{V}^m)^2] / \mu_{k-1}^2 \end{cases} \quad (10)$$

公式(10)表明由隐含层重构的可见层数据将在处理最后一个 mini-batch 后得到。在确定步长之后,并在融合 mini-batch 技术情况下,CRBM 模型的迭代公式可表示为^[25]:

$$\theta^k = \xi \theta^{k-1} + (\eta / \chi) \Delta \theta^k \quad (11)$$

式中: ξ 为动量系数,无量纲。

根据上述,本文提出一种用于处理岩性识别问题的新混合模型 CRBM - PSO - XGBoost,设计的模型如图 1a 所示。该图显示模型由 CRBM, PSO 和 XGBoost 3 个部分构成。CRBM 模块首先处理训练数据,将原始的 n 个变量缩减为 n_1 个关键变量。之后,由新变量组成的数据将被 XGBoost 模块处理。XGBoost 处理过程可看作是一系列 CART 回归树的处理过程。因每棵树的结构及树的数量都是由经验参数确定的,表示 XGBoost 模块并非在最优状态,为此 XGBoost 模块在处理后需要经 PSO 模型进行参数调整以达最优状态。经 PSO 模型优化后, XGBoost 模块可用于处理预测数据,解决识别问题。提出模型的计算流程显示在图 1b 中:第一步,原始数据将被分为训练数据和预测数据,其中训练数据包含岩性数据,而预测数据不包含岩性数据;第二步,训练数据由公式(11)代表的 CRBM 模块处理,而公式(11)中的参数则由公式(8)和(10)确定;第三步, PSO - XGBoost 模块将处理由关键变量组成的新训练数据,其中 XGBoost 模块结构经验参数先以 PSO 种子形式确定,之后被带入公式(7)中进行迭代以达最优状态;最后,经优化的 XGBoost 模块将处理预测数据,完成预测任务。需要注意的是,公式(7)中的 $\mathbf{O}_i^t$ 和 $\mathbf{O}_i^t$ 由公式(2)确定。

2 实验验证

2.1 数据来源及实验设计

本次实验以姬塬油田西部长 4 + 5 段(延长组 4 + 5 段)致密砂岩储层为验证对象。姬塬油田处于鄂尔多斯盆地中西部天环坳陷和陕北斜坡两个一级构造单元之间,其整体构造形态呈一宽缓的北东 - 南西倾向的单斜(图 2)^[26-29]。油田发育多个含油层系,其中长 4 + 5 段是主力开发层系之一。长 4 + 5 段为浅水三角洲沉积,主要发育三角洲前缘亚相,其储层的形成受控于沉积展布,多为水下分流河道、水下天然堤和河口坝^[26-29]。油田西部目前共有 2 000 多口探井和评价井,但只有少部分井具有岩心资料,因此为完成长 4 + 5 段的精细地层格架建立和沉积展布规律分析等

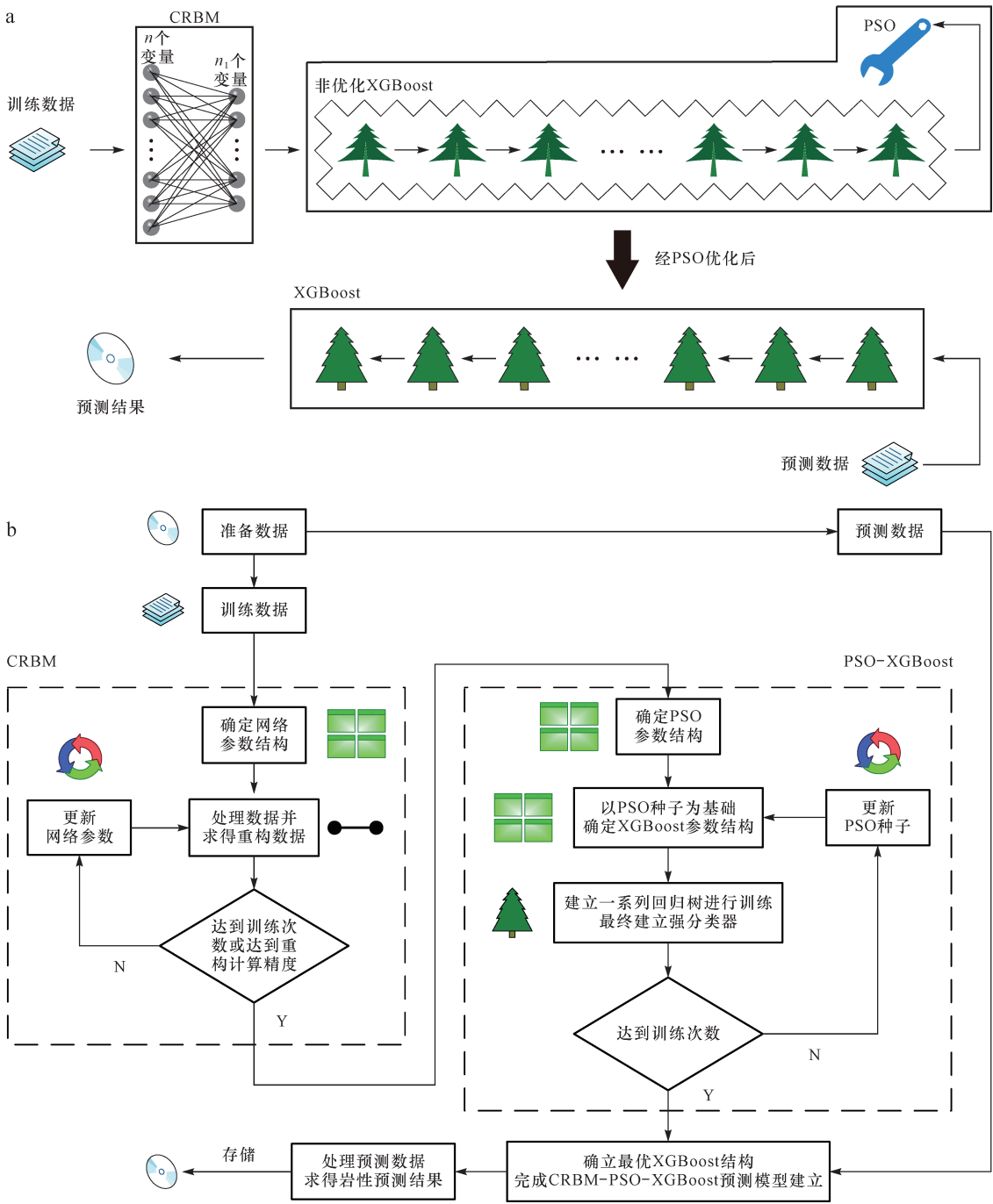


图 1 CRBM - PSO - XGBoost 模型结构(a) 及其计算流程(b)

Fig. 1 Structure(a)and computing flow (b) of the CRBM-PSO-XGBoost,a hybrid model proposed

工作,岩性识别成为一项关键研究内容。根据多口探井的岩心资料观察,识别出目的层储层主要岩性共 8 种,分别为中砂岩、细砂岩、粉-细砂岩、粉砂岩、泥质粉-细砂岩、泥质粉砂岩、粉砂质泥岩和泥岩。依据经典交会图的设计原理,本次选用测井数据中显示孔隙性的 AC(声波时差)、含泥性的 GR(自然伽马)和含油性的 AT90(阵列感应电阻率)来划分岩性。图 3a 为 AC - GR - AT90 三维交会图,可见 8 种岩性的测井数据点在图中融杂在一起,难以进行区分。图 3b—d 为 3 种曲线

两两组合形成的二维交会图,同样,各个岩性的数据点在图中仍有很程度上的重合,导致划分标准难以建立。图 3 表明目的层的主要岩性不能由二维或三维交会图进行识别,其原因是多种岩性具有相似的测井响应特征,使有效的岩性 - 测井响应识别关系无法形成。为此,本文采用机器学习方法来解决岩性识别问题,并根据在引言中的分析提出一种新的混合模型 CRBM - PSO - XGBoost。模型结构及其计算流程已经在计算原理中进行了说明,并用图 1 进行了展示,这里不再赘述。

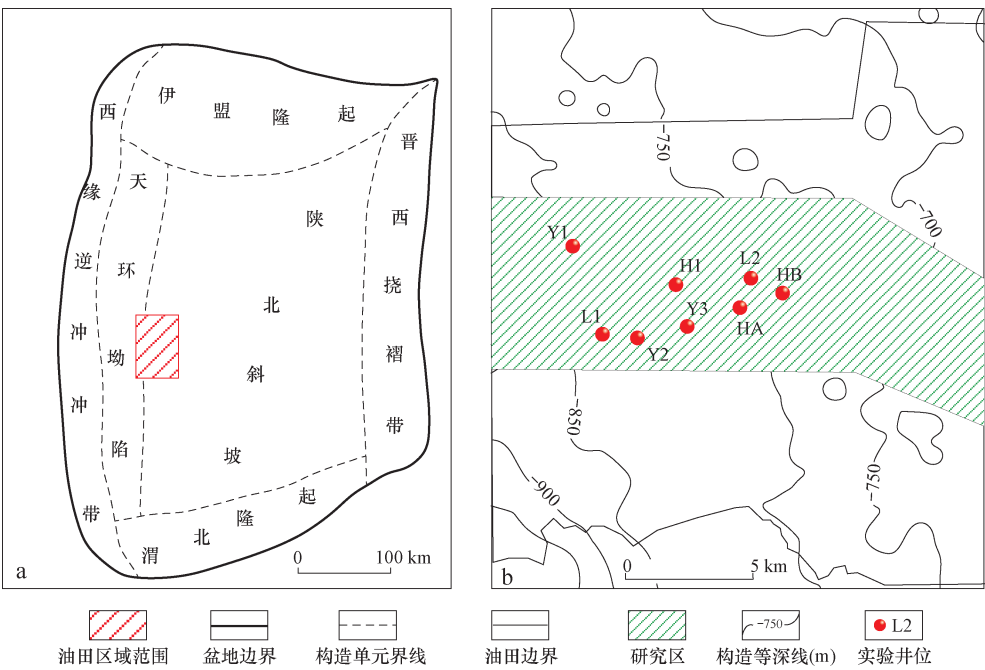


图 2 鄂尔多斯盆地构造区划(a)及姬塬油田西部工区概况(b)
Fig. 2 Structural division (a) and outline of western Jiyuan oilfield (b), Ordos Basin

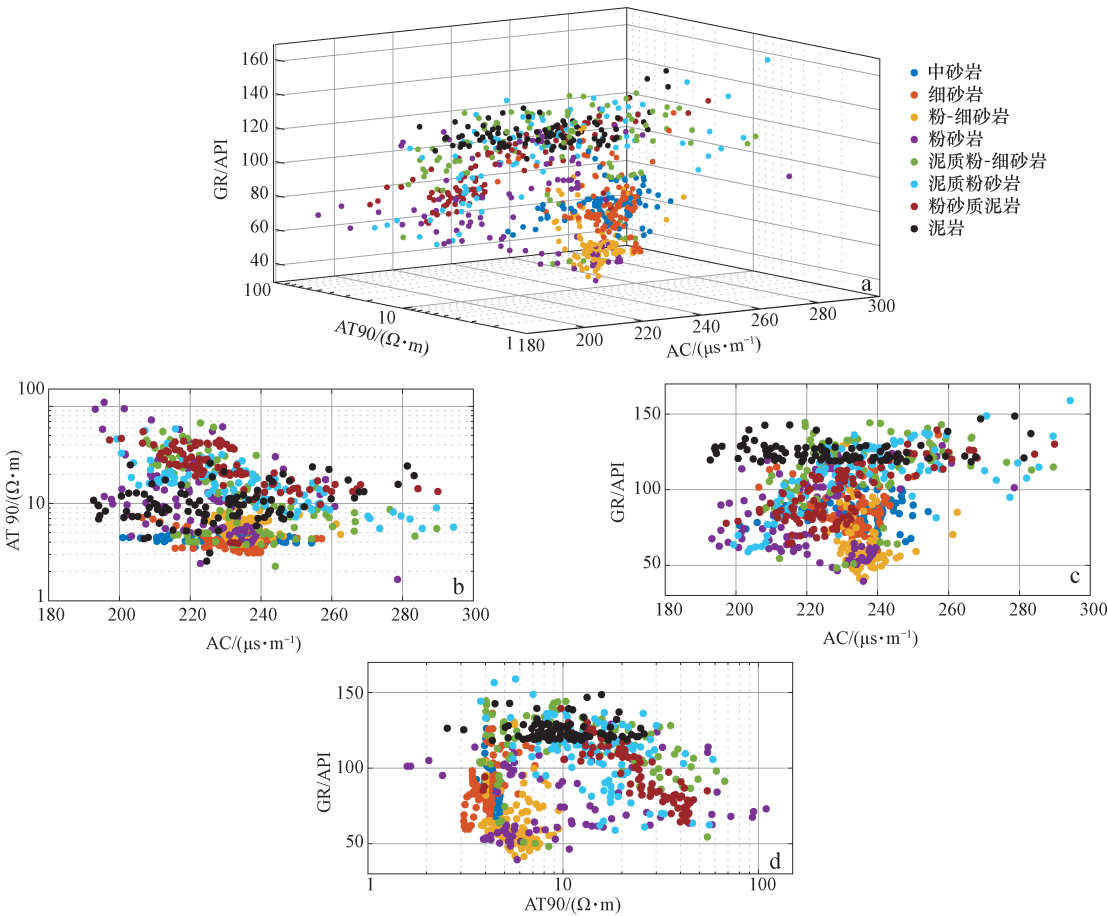


Fig. 3 3D or 2D cross correlations used to identify 8 major types of lithology of target formations
a. AC - GR - AT90 三维交会图; b. AC - AT90 二维交会图; c. AC - GR 二维交会图; d. GR - AT90 二维交会图

为验证所提出模型的预测能力,本文采用研究区 8 口取心井的测井及岩心观察数据进行实验,井位如图 2c 所示。岩心观察数据是根据每一个测井深度点从岩心柱上观察所得到的岩性数据。Y1, Y2, Y3, L1, L2 和 H1 井设为训练井,即利用这些井的资料组成学习样本,而 HA 和 HB 井设为验证井,即利用两口井的资料组成预测样本。学习样本共 4 606 个,而预测样本每口井有 300 个。样本由测井曲线和岩心信息组成,其中测井曲线有 11 种,分别是 AC(声波时差)、SP(自然电位)、GR(自然伽马)、CNL(补偿中子)、DEN(补偿密度)、PE(光电吸收截面指数)和阵列感应电阻率(AT10, AT20, AT30, AT60, AT90)。岩心信息为岩心观察数据的原始编码经 one-hot coding 技术转换后得到的信息。本次规定中砂岩、细砂岩、粉-细砂岩、粉砂岩、泥质粉-细砂岩、泥质粉砂岩、粉砂质泥岩和泥岩对应的原始编码依次为数字 1 到 8。所以,针对某一测井深度点,如果从岩心上观察到的岩性为细砂岩,则与之对应的原始编码为 2,岩心信息为[0, 1, 0, 0, 0, 0, 0, 0]。为简易说明,岩心信息在表中用原始编码进行了展示。实验共有两个,其中第一个实验是利用由 Y2, Y3, L2, H1 资料组成的含有 3 060 个样本的学习数据进行预测,第二个是利用全部学习样本进行预测,其目的是检验预测模型的识别准确率是否会在增加训练样本量的情况下有所提升。为增强验证效果,实验中加入了 PNN 和 SVM 模型进行对比。

2.2 实验过程及结果

在实验 1 中,先验证 XGBoost 模型是否在嵌入 PSO 模型和 CRBM 模型后其预测能力有所改变,之后再对所有预测模型进行对比。对于 XGBoost 模型,一组根据前人研究成果设定的经验参数如表 1 的第 3 列上部分所示。进行优化前,PSO 模型的计算参数也要设定。一组经验参数展示在了表 1 中部。需要指出的是,PSO 模型较 XGBoost 模型更容易找到一组理想的设置参数,这是因为 PSO 模型目的就是模型参数进行调优以确保预测结果最为可靠,而这点很容易在 PSO 模型迭代计算中实现,即在不动 XGBoost 模型预先设置参数的情况下,PSO 模型可通过简单的参数调试甚至不用参数调试即可令 XGBoost 模型参数达到最优化,所以 PSO 模型的嵌入虽然增加了预先设置参数的工作量,但在实际操作上减少了调参工作量^[20-22]。对于 CRBM 模型,一组设定的经验参数如表 2 所示。

为使 PSO - XGBoost 混合模型更快速地完成建模,CRBM 模型应从源数据中提取更少的特征,因此依据前人研究经验,隐含层神经元个数可设置为测井曲线个数的一半^[23-25]。由于 CRBM 模型目的只是提取数据特征,所以只要提取的特征满足迭代条件即认为 CRBM 模型完成了任务,这使得该模型的调参工作也变得非常简单。图 4 展示了测井源数据和由 CRBM 模型得到的重构数据的对比情况。可见,各曲线的两种数据的吻合度非常高,表明 CRBM 模型对源数据的提取是有效的。在完成 CRBM 模型处理后,分别利用原始学习样本和提取特征数据对 PSO - XGBoost 混合模型进行训练。图 5 显示了 PSO 模型作用在 XGBoost 模型上的优化过程。可以明显地看出,PSO 模型能够有效地优化 XGBoost 模型,尤其是在处理提取特征数据的情况下。在完成 XGBoost, PSO - XGBoost 和 CRBM - PSO - XGBoost 3 种模型的训练之后,预测目标便可进行处理,这里先以 HA 井为例。3 种模型的预测准确率分别为 51.00%, 80.33% 和 92.67%,显示出提出模型的预测能力最强。图 6 以柱状图的形式展示了部分岩性预测结果。图中取心道的信息为实际取心柱的观察结果,已经通过深度校正归位。为便于分析,在取心道上选择了 20 个样点进行对比。通过观察发现,XGBoost 模型结果中有 10 个错误样点(No. 1, 3, 5, 10, 12, 14, 15, 17, 18, 19),PSO - XGBoost 混合模型有 6 个(No. 3, 6, 12, 14, 16, 18),而提出模型仅有 2 个(No. 12 和 18)。由对比可知,XGBoost 模型在使用经验参数的情况下得到的预测结果不能准确地反映岩性实际分布规律,而经 CRBM 模型和 PSO 模型优化后,其预测能力得到明显提升,得到的预测结果非常可靠,可有效地反映储层岩性分布情况。

在明确 PSO 模型和 CRBM 模型的嵌入对 XGBoost 模型的预测能力有提升作用之后,提出模型将与 PNN 和 SVM 模型进行对比。由于对比模型在建模时也需要用到经验参数,因此为使所有验证模型在预测时都能达到最佳状态,PSO 模型和 CRBM 模型也将对 PNN 和 SVM 进行优化。表 2 记录了所有验证模型参数的设置及其优化结果。对于验证井 HA,表 3 记录了用 3 种优化模型得到的预测结果。可见,在相同优化的条件下,XGBoost 模型以 92.67% 的高识别准确率成为预测能力最强的模型。图 7 以柱状图的形式展示了 HA 井部分岩性预测结果。通过观察发现,CRBM - PSO - PNN 模型预测结果有 6 个错误样点(No. 2, 6, 9, 14, 16, 18),CRBM - PSO - SVM 模型有 4 个错误样点(No. 2, 9, 13, 19),与提出模型的 2 个错误样点相比,

表 1 验证模型参数设置及优化结果

Table 1 Parameters selected for the validation model and corresponding optimized data

PNN		SVM	XGBoost
初始参数 (是否需要优化)	$\alpha = X(0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1)$ (是)	$c = 1$ (是) 核函数 = RBF (否) $\gamma = 1$ (是)	$D = 100$ (是) $\eta = 0.01$ (是) $\max_depth = 3$ (是) $\lambda = 0.1$ (是) $\min_chile_weight = 0.1$ (是) $\delta = 0$ (是) 损失函数 = 交叉熵损失函数 (否)
PSO 参数设定	$\alpha (10^{-2}, 2 \times 10^0)$	$q = 10$ $t = 100$ $\omega_{\max}, \omega_{\min} = 0.9, 0.4$ $c_1, c_2 = 1.5$ r_1 和 r_2 , 范围 $[0, 1]$	
		各优化参数的 $\sigma_{\max_1}$ 和 $\sigma_{\max_2}$	
			$D (10^2 0, 10^4)$ $\eta (10^{-3}, 10^0)$ $\max_depth (3 \times 10^0, 10^1)$ $\lambda (10^0, 10^2)$ $\min_chile_weight (10^0, 10^2)$ $\delta (10^0, 10^1)$
		各优化参数的 $W_{\max}$	
优化后参数	$\alpha (0.1)$	$c (10)$ $\gamma (10)$	$D(50)$ $\eta(0.1)$ $\max_depth(5)$ $\lambda (10)$ $\min_chile_weight(10)$ $\delta(5)$
			$D = 730$ $\eta = 0.2$ $\max_depth = 5$ $\lambda = 1.3$ $\min_chile_weight = 1.2$ $\delta = 0$

注: α 为概率密度分布窗口长度, 因识别岩性共有 8 种, 因此向量长度为 8; c 为惩罚系数; RBF 为径向基函数; γ 为平滑因子; D 为迭代次数; η 为学习速率; $\max_depth$ 为最大回归树深度, 即回归树最大分裂次数; λ 为正则项系数; $\min_chile_weight$ 为叶节点最小样本权重之和, 如果样本二阶导之和小于该值, 则停止选用最佳分裂点进行划分; δ 为最小分裂梯度下降值, 如果叶节点中样本对应的梯度下降值小于该值, 则停止选用最佳分裂点进行划分; q 为种子数量; t 为迭代次数; $\omega_{\max}, \omega_{\min}$ 为 LDIW 算法中最大和最小重力系数; c_1, c_2 为加速系数; r_1, r_2 为弹性系数; 表中所有参数均为无量纲。

表中数据格式为 (最小值, 最大值)。

表 2 CRBM 参数设置

Table 2 Parameters selected for the CRBM model

参数	值
最大迭代次数 (K)	100
迭代精度 (ϵ)	0.001
可见层神经元个数 (n_v)	11
隐含层神经元个数 (n_h)	6
动量系数 (ξ)	0.9
学习速率 (η)	0.1
噪音控制参数 (μ)	0.05
噪音方差 (σ)	0.2
S 函数上渐近线 (ϕ_h)	1
S 函数下渐近线 (ϕ_l)	0
mini-batch 学习样本容量 (X)	100

反映出 PNN 和 SVM 优化模型所给的预测结果不能准确地描述储层岩性分布状况。

分析 HA 井预测结果之后, 再利用 3 种模型对 HB 井进行预测。建模时, PSO 模型和 CRBM 模型所用的参数设置不变。表 3 给出了 3 种优化模型的预测结果。结果显示, 混合模型 CRBM – PSO – XGBoost 的预测准确率仍为最高, 达 90.33%, 再次表明提出模型的预测能力最强。

在实验 2 中, 所有 4 606 个学习样本将用于建模, 其目的是检验在训练更多学习样本的情况下, 各验证模型的预测能力是否有所加强。建模时, PSO 模型和 CRBM

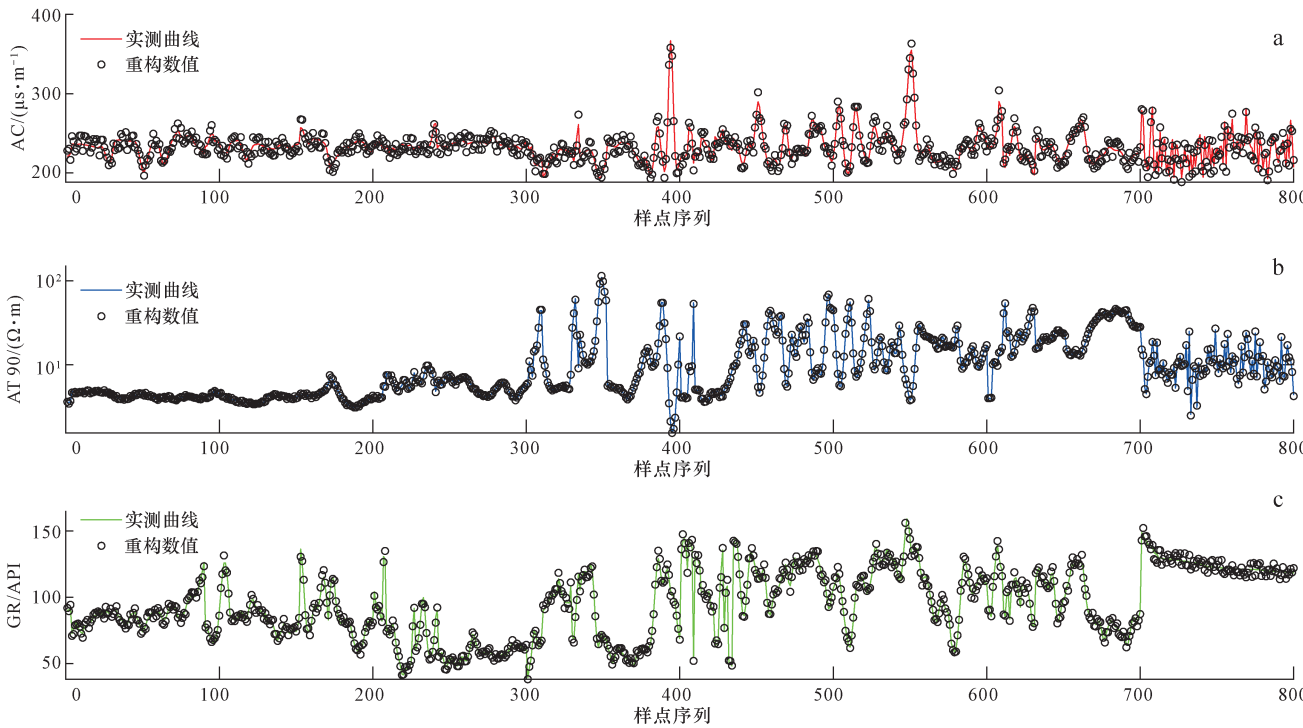


图 4 CRBM 重构数据检测

Fig. 4 Testing of reconstructing data for the CRBM model

a. AC 实测曲线与其重构数据比较;b. AT 90 实测曲线与其重构数据比较;c. GR 实测曲线与其重构数据比较

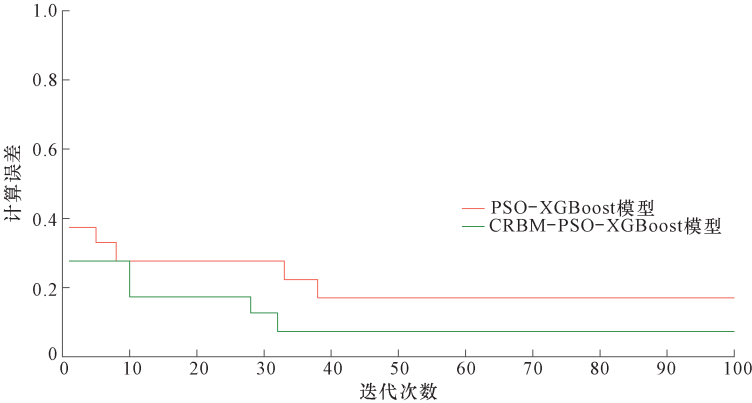


图 5 PSO - XGBoost 和基于 CRBM 的 PSO - XGBoost 训练优化过程

Fig. 5 Training optimization of PSO-XGBoost and CRBM-based PSO-XGBoost

模型所用的参数设置与实验 1 的一致。表 3 记录了用 3 种优化模型得到的两口验证井的预测结果。通过对比发现:①在训练更多学习样本后,各验证模型的预测准确率都有所提升,表明增大训练样本容量是提高模型预测能力的一种有效途径;②提出模型的识别准确率最高,都超过了 90%,不仅显示出提出模型所得的预测结果可靠性高,还再次证明了该模型的预测性能最佳。

表 3 中的计算时间数据显示 CRBM - PSO - PNN 耗时最长,而 CRBM - PSO - SVM 的最短。对于 PNN 模型,由于每个预测样本在预测时都要与全部训练

样本进行计算,因此耗时最长,而且这种耗时会随着训练样本容量的扩大而急剧增加(对比实验 1 和实验 2 的数据),表明在处理大数据时该模型效率低。SVM 模型预测时,采用的是支持向量而不是全部学习样本,所以耗时短。但与 XGBoost 模型相比,SVM 模型在两个实验中的计算时间也都仅快 10 s 左右,这在岩性识别问题中优先考虑预测准确率的情况下,并没有显示出该模型的预测效率高,反而体现了 XGBoost 模型的预测效率高。因此,综合来看,提出模型的预测效率最高,即使是在增大训练样本容量的条件下。

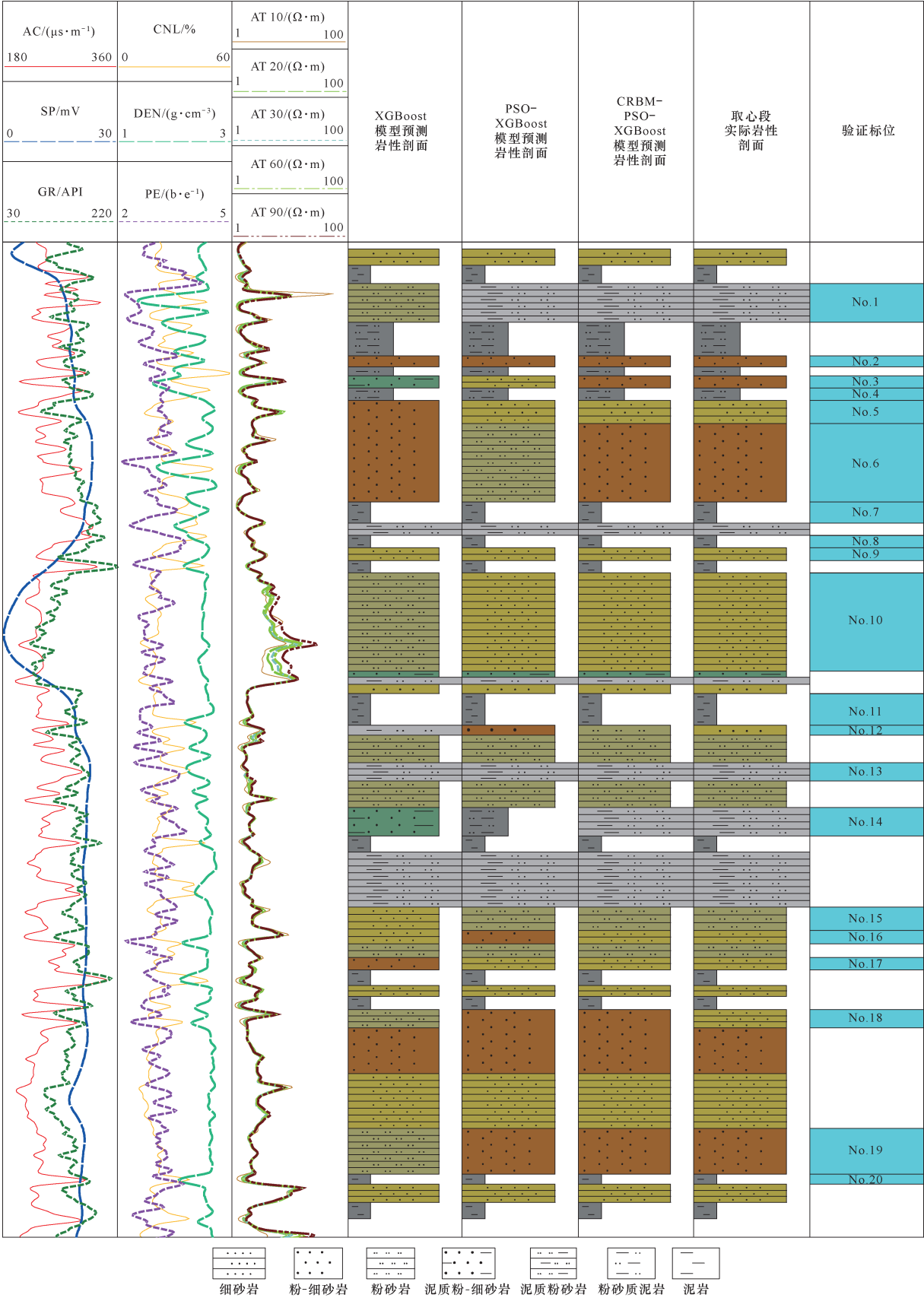


图 6 实验 1 中 HA 井岩性预测信息柱状图

Fig. 6 Columns showing the predicted lithology in Well HA derived from the Experiment 1

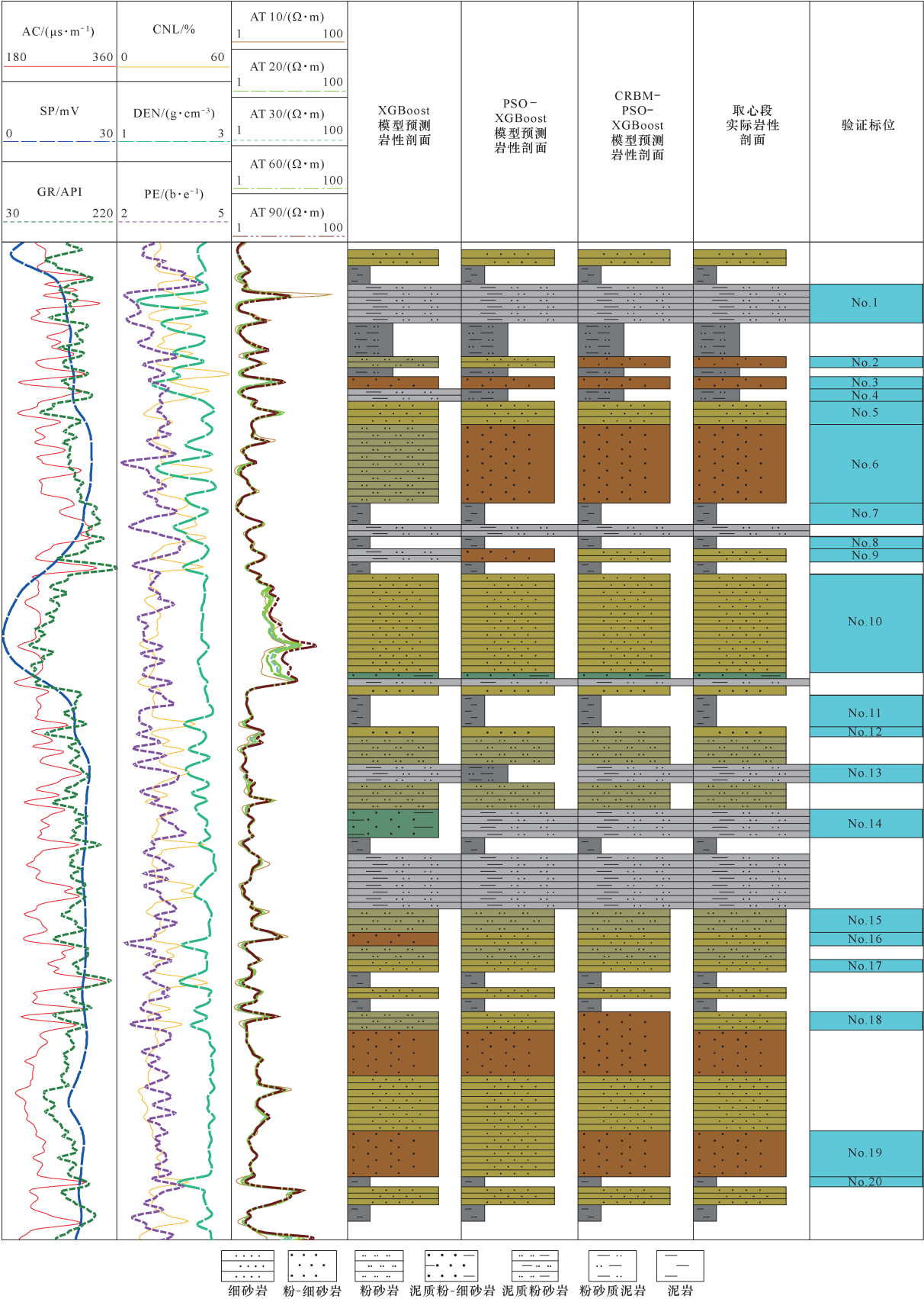


图 7 实验 2 中 HA 井岩性预测信息柱状图

Fig. 7 Columns showing the predicted lithology in Well HA derived from Experiment 2

表 3 验证井预测准确率和计算时间信息统计

Table 3 Data summary of prediction accuracy and computing time of validation wells

预测模型	实验 1		实验 2	
	HA 井	HB 井	HA 井	HB 井
CRBM – PSO – PNN	79.00/50.269 2	75.33/50.269 1	83.00/120.348 5	80.67/120.348 6
CRBM – PSO – SVM	83.67/25.487 1	77.00/25.487 2	88.33/43.247 3	84.33/43.247 2
CRBM – PSO – XGBoost	92.67/35.983 7	90.33/35.983 6	96.33/54.453 9	93.67/54.453 7

注:表中数据为识别准确率(%) / 计算时间(s);计算时间包含建模和预测两部分耗时。

3 结论

- 1) PSO 在建模过程中能够优化 XGBoost 多种经验参数,为强化模型预测能力奠定了基础。
- 2) CRBM 模型可从源数据中提取更少但更利于分析因变量的新特征或称新自变量,为提高 PSO – XGBoost 混合模型计算效率提供了途径。
- 3) CRBM – PSO – PNN, CRBM – PSO – SVM 和 CRBM – PSO – XGBoost 的预测能力可在训练更多学习样本的情况下得到提升,表明扩大训练样本容量是提升各验证模型预测能力的一种有效手段。
- 4) CRBM – PSO – XGBoost 相比于 CRBM – PSO – PNN 和 CRBM – PSO – SVM 能给出更为可靠的预测结果,且耗时也较短,表明该模型预测效率高,在解决致密砂岩储层岩性识别问题上更具推广性。

参 考 文 献

[1] 孔强夫,杨才,李浩,等. 基于图论聚类和最小临近算法的岩性识别方法[J]. 石油与天然气地质,2020,41(4):884–890.
Kong Qiangfu, Yang Cai, Li Hao, et al. A lithology recognition method based on multi-resolution graph-based clustering and K-nearest neighbor: A case study from the Leikoupo formation carbonate reservoirs in western Sichuan Basin [J]. Oil & Gas Geology, 2020, 41(4):884–890.

[2] 张天付,黄理力,倪新锋,等. 塔里木盆地柯坪地区下寒武统吾松格组岩性组合及其成因与勘探意义[J]. 石油与天然气地质,2020,41(5):928–940.
Zhang Tianfu, Huang Lili, Ni Xinfeng, et al. Lithological combination, genesis and exploration significance of the lower Cambrian Wusonggeer formation of Kalpin area in Tarim Basin: Insight through the deepest Asian onshore well-Well Luntan 1 [J]. Oil & Gas Geology, 2020, 41(5):928–940.

[3] 黄彦庆,刘忠群,林恬,等. 川东北元坝地区徐家河组三段基于相控的相对优质储层预测[J]. 石油与天然气地质,2021,42(4):863–872.
Huang Yanqing, Liu Zhongqun, Lin Tian, et al. Lithofacies-based prediction of relatively high-quality reservoirs of the Xu 3 member in Yuanba area, northeastern Sichuan Basin [J]. Oil & Gas Geology,

2021, 42(4):863–872.

[4] 郑见中. 辽河拗陷兴隆台潜山带中生界储层测井评价方法[J]. 石油地质与工程,2019,33(6):29–33.
Zheng Jianzhong. Logging evaluation method of Mesozoic reservoir in Xinlongtai buried hill belt of Liaohe depression [J]. Petroleum Geology and Engineering, 2019, 33(6):29–33.

[5] 闫林,冉启全,高阳,等. 新疆芦苇沟组致密油赋存形式及可动性评价[J]. 油气藏评价与开发,2017,7(6):20–26.
Ran Lin, Ran Qiquan, Gao Yang, et al. Tight oil occurrence form and recoverability evaluation of tight oil reservoir in Lucaogou Formation of Xinjiang [J]. Reservoir Evaluation and Development, 2017, 7(6):20–26.

[6] 袁述武,李想,史乐,等. 克拉玛依油田六、七、九区石炭系内幕有利火山岩储层岩性分布预测[J]. 大庆石油地质与开发,2019,38(6):40–45.
Yuan Shuwu, Li Xiang, Shi Le, et al. Prediction of the lithology distribution of favorable volcanic reservoirs in the inside Carboniferous system in Block 6, 7 and 9 of Karamay Oilfield [J]. Petroleum Geology & Oilfield Development in Daqing, 2019, 38(6):40–45.

[7] 赵彤彤,张春雷,张春雨,等. 基于模糊熵的 KNN 分类模型在岩性识别中的应用[J]. 计算机工程与应用,2018,54(24):260–265.
Zhao Tongtong, Zhang Chunlei, Zhang Chunyu, et al. Application of KNN classification model based on fuzzy entropy in lithology recognition [J]. Computer Engineering and Applications, 2018, 54(24):260–265.

[8] 张梓童,张春雷,张艳,等. 数据空间结构性及在 KNN 算法中的应用[J]. 数学的实践与认知,2019,49(1):195–202.
Zhang Zitong, Zhang Chunlei, Zhang Yan, et al. Data space structure and application in KNN algorithm [J]. Mathematics in Practice and Theory, 2019, 49(1):195–202.

[9] 陈玉林,李戈理,杨智新,等. 基于 KNN 算法识别合水地区长 7 储层岩性岩相[J]. 测井技术,2020,44(2):182–185.
Chen Yulin, Li Geli, Yang Zhixin, et al. Identification of lithology and lithofacies of Chang 7 reservoir in Heshui area by KNN algorithm [J]. Well Logging Technology, 2020, 44(2):182–185.

[10] 赵杰,李春华. 基于神经网络的两种岩性识别方法的研究[J]. 现代电子技术,2009,32(23):136–138.
Zhao Jie, Li Chunhua. Research on two lithology identification methods based on neural network [J]. Modern Electronics Technique, 2009, 32(23):136–139.

[11] 陈刚. PNN 在煤田随钻测井岩性判识应用研究[J]. 地质与资源,2018,27(1):103–106.

- Chen Gang. Application of PNN in the lithology identification of logging while drilling in coal field[J]. *Geology and Resources*, 2018, 27(1): 103–106.
- [12] 陈刚, 汪凯斌, 薛必辞, 等. 随钻测井中岩性识别方法的对比及应用[J]. *煤田地质与勘探*, 2018, 46(1): 165–169.
- Chen Gang, Wang Kaibin, Xue Bici, et al. Comparison and application of LWD lithology identification method[J]. *Coal Geology & Exploration*, 2018, 46(1): 165–169.
- [13] 李政宏, 刘永福, 张立强, 等. 数据挖掘方法在测井岩性识别中的应用[J]. *断块油气藏*, 2019, 26(6): 713–718.
- Li Zhenghong, Liu Yongfu, Zhang Liqiang, et al. Application of data mining method in lithology identification using well log[J]. *Fault-block Oil & Gas Field*, 2019, 26(6): 713–718.
- [14] 林香亮, 朱建伟, 刘光涛, 等. 基于 PCA-SVM 的砂砾岩岩性识别[J]. *长江大学学报(自然科学版)*, 2020, 17(1): 21–26.
- Lin Xiangliang, Zhu Jianwei, Liu Guangtao, et al. Lithologic identification of glutenite based on PCA-SVM[J]. *Journal of Yangtze University (Natural Science Edition)*, 2020, 17(1): 21–26.
- [15] 石锁, 余继峰, 曹慧涛, 等. 基于高斯核 SVM 的储层岩性识别——以东濮凹陷上古生界碎屑岩为例[J]. *中国科技论文*, 2020, 15(1): 112–118.
- Shi Suo, Yu Jifeng, Cao Huitao, et al. Reservoir lithology identification based on SVM using radial basis function: an example of Upper Paleozoic elastic rocks in Dongpu sag[J]. *China Science Paper*, 2020, 15(1): 112–118.
- [16] Chen T, Guestrin C. XGboost: A scalable tree boosting system[C]// *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016: 785–794.
- [17] Zhou K, Zhang J, Ren Y, et al. A gradient boosting decision tree algorithm combining synthetic minority over-sampling technique for lithology identification[J]. *Geophysics*, 2020, 85: 1–52.
- [18] Dev V A, Eden M R. Formation lithology classification using scalable gradient boosted decision trees[J]. *Computers & Chemical Engineering*, 2019, 128: 392–404.
- [19] 闫星宇, 顾汉明, 肖逸飞, 等. XGBoost 算法在致密砂岩气层测井解释中的应用[J]. *石油地球物理勘探*, 2019, 54(2): 447–455.
- Yan Xingyu, Gu Hanming, Xiao Yifei, et al. XGBoost algorithm applied in the interpretation of tight-sand gas reservoir on well logging data[J]. *Oil Geophysical Prospecting*, 2019, 54(2): 447–455.
- [20] 杨维, 李歧强. 粒子群优化算法综述[J]. *中国工程科学*, 2004, 6(5): 87–94.
- Yang Wei, Li Qiqiang. Survey on particle swarm optimization algorithm[J]. *Engineering Science*, 2004, 6(5): 87–94.
- [21] 刘建华. 粒子群算法的基本理论及其改进研究[D]. 长沙: 中南大学, 2009.
- Liu Jianhua. The research of basic theory and improvement on particle swarm optimization [D]. Changsha: Center South University, 2009.
- [22] 温阳东, 李龙剑. 基于 LDIW-PSO 算法的 BP 神经网络在压力传感器中的应用[J]. *化工自动化及仪表*, 2014, 41(9): 1031–1034.
- Wen Yangdong, Li Longjian. Application of LDIW-PSO algorithm-based BP neural network in pressure sensor[J]. *Control and Instruments in Chemical Industry*, 2014, 41(9): 1031–1034.
- [23] Tang T B, Murray A F. Adaptive sensor modeling and classification using a continuous restricted Boltzmann machine (CRBM) [J]. *Neurocomputing*, 2006, 70(7–9): 1198–1206.
- [24] Huang H B, Li R X, Yang M L, et al. Evaluation of vehicle interior sound quality using a continuous restricted Boltzmann machine-based DBN[J]. *Mechanical Systems & Signal Processing*, 2017, 84: 245–267.
- [25] Chen Y, Lu L, Li X. Application of continuous restricted Boltzmann machine to identify multivariate geochemical anomaly[J]. *Journal of Geochemical Exploration*, 2014, 140(4): 56–63.
- [26] 黎盼. 低渗透砂岩储层微观孔隙结构表征及生产特征分析[D]. 西安: 西北大学, 2019.
- Li Pan. Microscopic pore structure characterization and production characteristic analysis of low-permeability sandstone reservoir: a case study on Chang 4 + 5 and Chang 6 reservoir of T area in Jiyuan Oilfield [D]. Xi'an: Northwest University, 2019.
- [27] 赵冰瑶. 姬塬油田王盘山地区长 4 + 5 储层成岩相微观孔隙结构及渗流特征研究[D]. 西安: 西北大学, 2018.
- Zhao Bingyao. The study on microscopic pore structure and percolation characteristics of diagenetic facies in Chang 4 + 5 reservoir of Wangpanshan area, Jiyuan Oilfield [D]. Xi'an: Northwest University, 2018.
- [28] 翟利华. 姬塬油田西部长 4 + 5 油层组沉积微相与单砂体研究[D]. 北京: 中国石油大学(北京), 2018.
- Zhai Lihua. Study on sedimentary microfacies and single sandbodies on Chang 4 + 5 member in the western Jiyuan Oilfield [D]. Beijing: China University of Petroleum(Beijing), 2018.
- [29] 沈佳男. 鄂尔多斯盆地姬塬油田西部长 4 + 5 油层组储层评价[D]. 北京: 中国石油大学(北京), 2018.
- Shen Jianan. Reservoir evaluation of Chang 4 + 5 member in west Jiyuan Oilfield in Ordos Basin [D]. Beijing: China University of Petroleum(Beijing), 2018.

(编辑 张玉银)