计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(6):108−116 [doi: 10.15888/j.cnki.csa.009522]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 BiLSTM-XGBoost 混合模型的储层岩性识别①

杜睿山[1,2], 黄玉朋[1], 孟令东[2], 张轶楠[1], 周长坤[1]

[1](东北石油大学 计算机与信息技术学院, 大庆 163318)
[2](油气藏及地下储库完整性评价黑龙江省重点实验室 (东北石油大学), 大庆 163318)
通信作者: 杜睿山, E-mail: ruishan_du@163.com

摘　要: 储层岩性分类是地质研究基础, 基于数据驱动的机器学习模型虽然能较好地识别储层岩性, 但由于测井数据是特殊的序列数据, 模型很难有效提取数据的空间相关性, 造成模型对储层识别仍存在不足. 针对此问题, 本文结合双向长短期循环神经网络 (bidirectional long short-term memory, BiLSTM) 和极端梯度提升决策树 (extreme gradient boosting decision tree, XGBoost), 提出双向记忆极端梯度提升 (BiLSTM-XGBoost, BiXGB) 模型预测储层岩性. 该模型在传统 XGBoost 基础上融入了 BiLSTM, 大大增强了模型对测井数据的特征提取能力. BiXGB 模型使用 BiLSTM 对测井数据进行特征提取, 将提取到的特征传递给 XGBoost 分类模型进行训练和预测. 将 BiXGB 模型应用于储层岩性数据集时, 模型预测的总体精度达到了 91%. 为了进一步验证模型的准确性和稳定性, 将模型应用于 UCI 公开的 Occupancy 序列数据集, 结果显示模型的预测总体精度也高达 93%. 相较于其他机器学习模型, BiXGB 模型能准确地对序列数据进行分类, 提高了储层岩性的识别精度, 满足了油气勘探的实际需要, 为储层岩性识别提供了新的方法.

关键词: 神经网络; 机器学习; 测井数据; 岩性分类; BiLSTM; XGBoost

引用格式: 杜睿山,黄玉朋,孟令东,张轶楠,周长坤.基于 BiLSTM-XGBoost 混合模型的储层岩性识别.计算机系统应用,2024,33(6):108–116. http://www.c-s-a.org.cn/1003-3254/9522.html

Reservoir Lithology Identification Using Hybrid Model BiLSTM-XGBoost

DU Rui-Shan[1,2], HUANG Yu-Peng[1], MENG Ling-Dong[2], ZHANG Yi-Nan[1], ZHOU Chang-Kun[1]

[1](School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)
[2](Key Laboratory of Oil and Gas Reservoir and Underground Gas Storage Integrity Evaluation (Northeast Petroleum University), Daqing 163318, China)

Abstract: Reservoir lithology classification is the foundation of geological research. Although data-driven machine learning models can effectively identify reservoir lithology, the special nature of well logging data as sequential data makes it difficult for the model to effectively extract the spatial correlation of the data, resulting in limitations in reservoir identification. To address this issue, this study proposes a bidirectional long short-term memory extreme gradient boosting (BiLSTM-XGBoost, BiXGB) model for predicting reservoir lithology by combining bidirectional long short-term memory (BiLSTM) and extreme gradient boosting decision tree (XGBoost). By integrating BiLSTM into the traditional XGBoost, the model significantly enhances the feature extraction capability for well logging data. The BiXGB model utilizes BiLSTM to extract features from well logging data, which are then input into the XGBoost classification model for training and prediction. The BiXGB model achieves an overall prediction accuracy of 91% when applied to a reservoir lithology dataset. To further validate its accuracy and stability, the model is tested on the publicly available UCI

① 基金项目: 黑龙江省自然科学基金 (LH2021F004)
收稿时间: 2023-12-13; 修改时间: 2024-01-10; 采用时间: 2024-01-29; csa 在线出版时间: 2024-05-07
CNKI 网络首发时间: 2024-05-10

Occupancy dataset, achieving an overall prediction accuracy of 93%. Compared to other machine learning models, the BiXGB model accurately classifies sequential data, improving the accuracy of reservoir lithology identification and meeting the practical needs of oil and gas exploration. This provides a new approach for reservoir lithology identification.

Key words: neural network; machine learning; logging data; lithology classification; bidirectional long short-term memory (BiLSTM); extreme gradient boosting decision tree (XGBoost)

1 引言

岩性识别是油气勘探和工程中具有重要意义的储层表征任务, 是储层质量评价 (孔隙度和渗透率) 的基础[1,2], 并可支持相关的地质研究和钻探活动 (沉积模拟、有利区带预测和井规划). 测井作为一种有效的测量手段, 可以从地面地球物理测量中预测地下地层岩性. 测井资料包含丰富的地质信息, 是地层岩性和物性的综合反映[3].

然而, 传统的测井解释严重依赖专业知识和人工经验, 劳动强度大、耗时长, 往往存在专家经验的主观性和不一致性[4]. 由于非常规储层 (碳酸盐岩、致密砂岩或砂砾岩储层[5,6]) 地质条件的复杂性以及测井资料的多样性和数据量的增加, 传统的测井解释方法显示出很大的局限性. 因此, 研究人员正转向更先进的岩性识别方法.

机器学习技术已经被地质工程研究作为解决其面临的复杂且具有挑战性的问题的替代方法, 以实现自动化和提高效率. 随着算法、计算理论和硬件 (图形处理器) 进步, 机器学习在从大量数据中学习复杂模式和关系等方面显示出巨大的优势[7,8]. 用于岩性识别的机器学习算法主要有两类, 无监督学习方法和有监督学习方法. 有监督学习方法主要是使用一组训练数据来学习特征和相应标签之间的关系, 并建立模型来预测以前未见的数据. 在利用测井资料进行岩性分类的研究中, 有监督学习算法比很多无监督学习算法有很大的优势[8,9].

2 相关工作

Bressan 等人[10]将机器学习算法应用于来自国际海洋学计划 (IODP) 的海上威尔斯的多变量数据进行岩性识别. 为了更好地评估考察模型, 他们采用交叉验证用于 4 个机器学习模型的训练. 多层感知器 (multilayer perceptron, MLP) 和随机森林被选为最好的模型. Xie 等人[11]提出了一种用于多类别数据分类的机器学习框架, 以训练具有外部因素的石油科学数据集. 应用无监督机器学习方法来检测离群值, 并且使用 4 个集成模型来对两个数据集进行岩性分类. 最佳准确度在 89%–91% 之间. 潘少伟等人[12]针对传统机器学习模型的不足提出使用 XGBoost 模型对岩性进行预测, 实验结果表明 XGBoost 模型能更好地对岩性进行分类. 并且因 XGBoost 模型因其训练稳定且计算效率高在岩性识别中得到了广泛的应用[13–15]. 同时神经网络在岩性识别方面也具有良好的性能[16].

之前的一些研究证明了机器学习方法在地质岩性识别中的适用性, 但仍存在一些不足. 如神经网络模型的训练需要大量时间并且对硬件要求较高, 导致模型训练成本较高[12]. 当前用于储层岩性识别的机器学习模型容易忽略目标深度处的测井曲线信息, 同时忽略了测井曲线的变化趋势和储层岩性相关性, 导致模型分类效果不佳[17,18].

为解决单一模型精度不足、神经网络模型训练时间长以及机器学习模型在数据特征学习方面的不足等问题, 本文提出双向记忆极端梯度提升 BiXGB 模型. 该模型使用 BiLSTM 网络作为特征提取模块, 更好的提取测井数据的空间特征. 使用 XGBoost 作为模型的最后一层, 通过学习 BiLSTM 提取的数据特征对储层岩性进行分类. 与传统的 BiLSTM 不同, 模型中 BiLSTM 没有全连接层, 减少了计算参数的数量, 不需要将权重从全连接层带回来重新调整前几层中的权重. 优化了模型的训练时间. 通过 BiLSTM 提取测井数据相关信息, 使得 XGBoost 可以学习到更多测井数据特征, 进而提高模型的分类效果. 通过与常用的岩性识别模型进行实验对比. 结果表明本文提出的 BiXGB 模型在岩性识别过程中具有更高的精度和效率, 这可以有效地帮助地质专业人员高效的识别储层岩性.

3 方法原理

3.1 BiLSTM

BiLSTM 对传统长短期记忆神经网络 (long short-

term memory, LSTM) 优化进行了改进[19]. Hochreiter 等人[20]建立了 LSTM 模型, 有效地解决了传统递归神经网络的梯度爆炸或消失问题. 控制机制实现了信息的选择性传递; 在传统循环神经网络 (recurrent neural network, RNN) 的基础上引入遗忘门、输入门和输出门, 使模型在反向传播过程中保持较稳定的误差, 并能在多个时间步长上学习, 从而提高时间序列预测的精度. 其结构如图 1 所示.

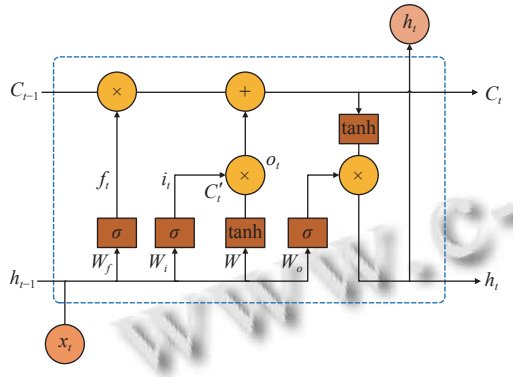


图 1 LSTM 网络结构图

LSTM 模型的总体框架由输入门x_t、单元状态C_t、临时单元状态C'_t、隐层状态h_t、遗忘门f_t、记忆门i_t、时刻t的输出门o_t组成. LSTM 的计算过程可以概括为: 通过遗忘单元状态中的信息并存储新的信息, 传递对后续计算有用的信息, 丢弃无用的信息, 并且在每个时间步h_t输出隐层状态, 其中遗忘、存储和输出由遗忘门f_t、记忆门i_t以及最后时刻的隐层状态h_{t-1}和当前输入x_t计算出的输出门o_t控制.

BiLSTM 由前向 LSTM 层和后向 LSTM 层组成. 两者都会影响输出, 既有利于前向序列信息的输入, 也有利于后向序列信息的输入. 它充分考虑了过去和未来的信息, 有利于进一步提高模型预测的精度. BiLSTM 的结构如图 2 所示.

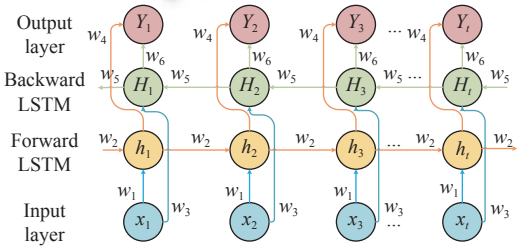


图 2 BiLSTM 网络结构图

图 2 中x_i $(i=1,2,\cdots,t)$对应时间输入的数据, h_i $(i=1,2,\cdots,t)$表示前向迭代的 LSTM 隐藏状态, H_i $(i=1,2,\cdots,t)$表示后向迭代的 LSTM 隐藏状态, Y_i $(i=1,2,\cdots,t)$表示对应的输出数据, w_i $(i=1,2,\cdots,6)$表示每一层的权重, BiLSTM 最终输出计算公式如下所示:

$$h_i = f_1(w_1 x_i + w_2 h_{i-1}) \tag{1}$$

$$H_i = f_2(w_3 x_i + w_5 H_{i+1}) \tag{2}$$

$$Y_i = f_3(w_4 h_i + w_6 H_i) \tag{3}$$

其中, f_1, f_2, f_3对应不同层的激活函数.

3.2 XGBoost

XGBoost 是属于梯度提升决策树 (gradient boosting decision tree, GBDT) 的一种, 对原有 GBDT 方法进行了多重增强, 以提高效率和可扩展性[21]. 这种机器学习方法通常用于数据的分类和回归. XGBoost 是K个分类树或回归树的集合, 每个树都有K_E^i $(i \in 1,\cdots,K)$个节点, 最后的预测结果是每个树的预测得分之和:

$$\hat{y}_i = \varphi(x_i) = \sum_{k=1}^{k} f_k(x_i),\ f_k \in F \tag{4}$$

其中, x_i是训练集数据, y_i是和训练集相对应的类标签, f_k是第 k 棵树的叶子分数, F是所有决策树的 K 评分 (K-score) 集合.

XGBoost 算法的目标函数分为损失函数和正则项两部分. 损失函数描述了目标的预测值与真实值之间的差值. 正则项控制树的复杂性, 防止过拟合. 目标函数公式为:

$$\begin{aligned}\tilde{\ell}^{(t)} &\simeq \sum_{i=1}^{n}\left[g_i f_i(x_i) + \frac{1}{2}h_i f_i^2(x_i)\right] + \Omega(f_t)\\ &= \sum_{i=1}^{n}\left[g_i f_i(x_i) + \frac{1}{2}h_i f_i^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2\\ &= \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right)w_j^2\right] + \gamma T\end{aligned} \tag{5}$$

其中, $I_j = \{i \mid q(x_i) = j\}$表示叶子 t 的实例集.

$$g_i = \frac{\partial l(\hat{y}_i^{(t-1)}, y_i)}{\partial \hat{y}_i^{(t-1)}} \tag{6}$$

$$h_i = \frac{\partial^2 l(\hat{y}_i^{(t-1)}, y_i)}{\partial (\hat{y}_i^{(t-1)})^2} \tag{7}$$

损失函数一阶、二阶梯度统计量可定义为, 对所给树结构$q(x_i)$, 可以计算出叶子j的最优权重w_j^*, 目标

函数的最优解.

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda} \tag{8}$$

$$\tilde{\ell}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T \tag{9}$$

3.3 BiXGB 结构

XGBoost 模型是一种集成学习模型具有良好的分类性能, 可以实现从测井数据中提取的特征进行岩性分类的任务, 具有较好的泛化能力和鲁棒性. 尽管 XGBoost 模型在储层岩性分类方面表现出色, 但由于无法提取测井数据中的序列特征, 因此在进行岩性识别时存在一定的局限性. 而 BiLSTM 能够双向提取测井数据中的序列特征, 更好地捕捉测井数据的正向和反向的依赖关系. 通过对特征信息进行选择性记忆和遗忘, BiLSTM 能够准确地学习测井信息随深度变化趋势和前后相关性. 但由于 BiLSTM 是一种递归神经网络, 需要大量的计算资源和时间来训练和推理. 在处理大规模测井数据时, 导致训练时间过长或者无法满足实时性要求.

因此本文根据 XGBoost 和 BiLSTM 的特点, 提出了一种基于 BiLSTM-XGBoost 混合模型的储层岩性识别模型 BiXGB, 模型的结构如图 3 所示.

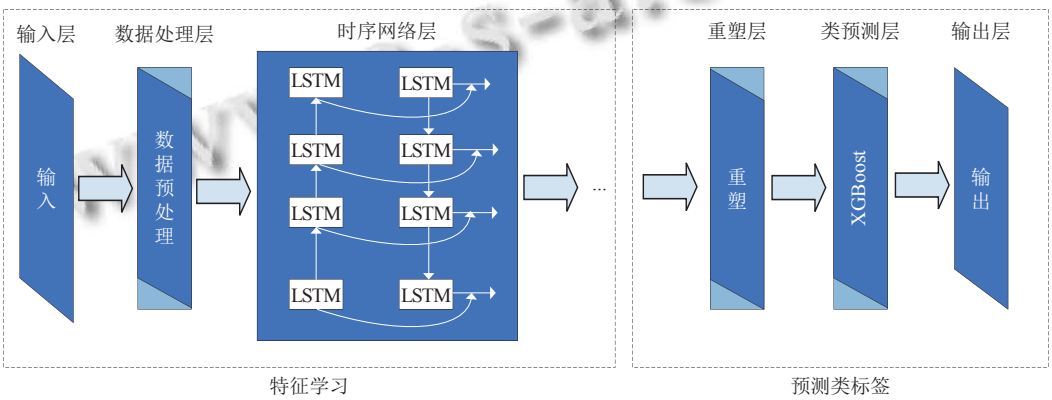


图 3 BiXGB 模型结构图

BiXGB 模型主要分为两部分: 一是数据的特征学习部分, 二是类标签预测部分. 特征学习部分主要包含输入层、数据预处理层以及双向时序网络层. 标签预测部分主要包含重塑层和类预测层.

3.3.1 特征学习

特征学习部分能够从训练数据集中学习到数据的关键特征. 模型的分类精度和有效的特征学习紧密相关.

输入层: 输入层是模型的第 1 层, 负责数据的输入. 数据集 X 由一元组 (x_i, y_i) 组成, 其中 i 数据集的索引. x_i 是 $M\times N$ 的特征矩阵, y_i 是向量 x_i 对应的类标签. 直接输入的数据一般不适合直接进行模型训练, 需要通过数据预处理层对数据进行处理, 转换成符合模型训练的数据格式.

数据预处理层: 在数据预处理层, 主要是对数据集进行归一化标准化处理, 计算公式如下:

$$x_i' = \frac{x_i - u}{\sigma} \tag{10}$$

其中, $u = \sum_{i=1}^{N} x_i$, $\sigma = \sum_{i=1}^{N}(x_i - u)^2$. 将处理后的数据转换为时序网络层需要的三维张量, 形状为 (data_size, time_steps, input_features), 其中 data_size 表示样本数量, time_steps 表示时间步长, input_features 表示特征数量.

双向时序网络层: 主要是从输入数据中提取相关特征. 该层由双向长短期记忆 (BiLSTM) 神经网络和双曲正切 (tanh) 激活函数组成. BiLSTM 具有选择性记忆和遗忘机制, 能够有效学习测井数据随深度变化的趋势以及数据前后相关性. 并且 BiLSTM 可以双向提取测井数据序列信息, 更好的收集测井数据的正向和反向依赖关系. 应用 tanh 激活函数可以引入非线性变换, 增加网络的表达能力和学习能力. 有助于 BiLSTM 网络更好地适应复杂的数据分布和提取特征.

3.3.2 类标签预测

数据通过双向时序网络层进行特征提取所得的特征信息是张量形式. 在输入到 XGBoost 分类器之前, 必

须将张量通过重塑层转换为分类器需要的向量形式.

重塑层: 主要是将双向时序网络层输出的特征张量转换为类预测层所需的向量.

类预测层: 主要是使用 XGBoost 作为分类模块, 通过对重塑层传入的 BiLSTM 提取的特征数据进行学习来预测储层岩性. XGBoost 作为强大的梯度提升算法, 能够有效地捕捉特征之间的非线性关系, 可以充分利用 BiLSTM 提取的测井数据特征.

输出层: 输出层通过类预测层得到分类结果. 然后将测试集输入训练好的模型进行分类, 生成分类结果.

3.4 BiXGB 模型算法

在 BiXGB 算法中, BiLSTM 特征提取模型和 XGBoost 分类模型同时训练. 首先使用 BiLSTM 对输入数据进行特征提取, 然后将提取到的特征传递给 XGBoost 分类模型进行训练和识别. 这种设计不仅减少了参数数量, 简化了算法结构, 并且不需要在全连接层进行反向传播. BiXGB 算法具有自动特征学习能力, 通过使用 BiLSTM 网络进行特征提取, 能够自动学习测井曲线的变化趋势和储层岩性相关性, 提高了算法的泛化能力和适应性. BiXGB 可以在一次前向传播中完成特征学习和分类预测两个任务, 减少了冗余步骤和不必要的计算.

BiXGB 模型训练流程见算法 1. 假设 $X=\{(x_i,y_i)|1\leqslant j\leqslant M\}$, 其中, M 是数据集的大小, y_j 是向量 x_j 的类标签.

算法 1. BiXGB 分类算法

输入: 数据分类数据集 $T=SS\cup QS\cup TS$ (其中 SS 表示训练集, QS 表示验证集, TS 表示测试集).
输出: 测试集分类预测结果.

Step 1. 初始化数据集并将数据集输入模型中, 对数据集进行归一化标准化处理, 然后将处理好的数据转换为三维张量;
Step 2. 对处理好的数据集使用双向时序网络层进行特征提取, 得到训练集特征张量 X_train、验证集特征张量 X_val 和测试集特征张量 X_test;
Step 3. 将提取的数据特征进行重塑得到训练集特征向量 X'_train、验证集特征向量 X'_val 和测试集特征向量 X'_test, 然后为分类预测层初始化数据集;
Step 4. 初始化 XGBoost 分类器;
Step 5. 将训练集和验证集通过重塑层得到的特征向量 X'_train 和 X'_val 输入 XGBoost 分类器模型进行模型训练, 然后对 X'_test 进行分类预测, 得到分类预测结果.

4 实验及分析

4.1 数据集获取

实验首先获取某油田的测井曲线, 使用测井曲线解释软件获得测井数据, 在数据清洗后构建出该油田储层岩性的测井数据, 共有 3 345 条测井数据如表 1 所示. 该地区岩性主要由粗粒砂岩、中粒砂岩、细粒砂岩、粉砂岩、泥质粉砂岩、粉砂质泥岩和泥岩组成, 分别用 0、1、2、3、4、5、6 表示. 用于训练分类模型的输入变量包括声波时差 (AC)、井径 (CAL)、伽马射线 (GR)、钾 (K)、电阻率 (RD)、自然电位 (SP).

表 1 油田测井数据

深度 (m)	AC (μs/m)	CAL (cm)	GR (API)	K (%)	RD (Ω·m)	SP (MV)	标签
2 887.5	0.029 1	0.031 7	0.026 7	0.202 3	0.078 985 77	0.333 754	4
2 888.0	0.032 8	0.033 4	0.030 5	0.203 1	0.076 064	0.333 669	4
2 888.5	0.034 3	0.037 0	0.032 3	0.200 1	0.074 503	0.333 619	4
…	…	…	…	…	…	…	…
4 559.0	0.262 0	0.050 6	0.163 0	0.398 4	0.064 970 463	0.421 846 126	2
4 559.5	0.280 8	0.051 9	0.172 1	0.383 1	0.064 963 157	0.419 508 045	2
4 560.0	0.305 2	0.055 3	0.184 2	0.366 8	0.065 756 153	0.417 479 831	2

为了验证模型对序列数据集的分类效果以及模型的实用性和准确性, 同时选用 UCI 中公开的 Occupancy 数据集[22]进行实验, 该数据集于 2016 年被提出. 使用历史数据训练模型, 然后根据当前的环境参数来预测房间是否会被占用, 与测井数据的相同点是该数据集也是序列数据, 该数据包含 6 个属性, 时间、湿度、光照、二氧化碳以及湿度比, 包含 2 个类别, 分别是房子被占用和房子未被占用.

4.2 模型评估

对于分类模型的性能评估有很多种方式, 本文采用准确率、$F1$ 分数、AUC 值以及召回率等指标对, 模型进行评价[23]. 这些评价指标在计算的时候涉及真、假正例, 真、假负例, 混淆矩阵如表 2 所示.

准确率 (*Accuracy*): 衡量所有预测正确的样本即模型输出结果类别与实际类别一致的样本占所有样本的比例.

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP} \tag{11}$$

表 2 混淆矩阵

真实标签	预测标签	
	正例 (positive)	负例 (negative)
正例 (positive)	*TP*	*FN*
负例 (negative)	*FP*	*TN*

精确度 (*Precision*): 用于衡量预测为正的实例中真正例的样本比例.

$$Precision = \frac{TP}{TP+FP} \tag{12}$$

召回率 (*Recall*): 代表了所有正样本中被正确预测的样本所占的比例, 召回率是评价分类器识别正样本能力的重要指标, 在分类器的性能评价中起到了至关重要的作用.

$$Recall = \frac{TP}{TP+FN} \tag{13}$$

*F*1 值: 是综合性的模型评价指标, 其数值为 *Recall* 和 *Precision* 的加权调和平均值. *F*1 值越大分类效果越好.

$$F1 = 2\times\frac{Precision\times Recall}{Precision+Recall} \tag{14}$$

AUC: AUC 为 ROC 曲线下的面积, 因为 ROC 曲线通常不能指示哪个分类器更好, 而 AUC 是一个值, 模型的 AUC 值越大, 则该模型越好. AUC 值通常在 0.5 和 1 之间浮动, 如果 AUC 小于 0.5, 则表示模型的准确性低于随机结果, 如果 AUC 等于 1, 则表明分类精度为 100%.

4.3 模型训练

实验采用 TensorFlow 深度学习框架, 在 CPU 平台上运行, 运行内存为 16 GB. 首先将数据按照训练集:验证集:测试集为 6:2:2 的比例进行划分, 然后将划分好的数据进行保存. 以保证实验中所有模型的训练均采用相同训练集、验证集和测试集.

实验中神经网络的参数设置, 两个 BiLSTM 神经网络层隐藏层神经元个数均为 256, 激活函数为 tanh, 学习率为5×10^{-4}, Dropout 率为 0.3, 使用交叉熵损失函数.

(1) 岩性数据集分类

为了评估 BiXGB 分类模型在岩性数据分类上的性能, 实验选用支持向量机 (support vector machine, SVM), 随机森林 (random forest, RF)、XGBoost、BiLSTM 与 BiXGB 几种分类方法进行分类效果对比, 实验结果如表 3 所示.

表 3 各模型在岩性数据集上实验结果比较

模型	*Accuracy*	AUC	*Recall*	*Precision*	*F*1	mAP	Time (s)
SVM	0.822	0.952	0.766	0.857	0.801	0.884	1.046
RF	0.841	0.964	0.752	0.889	0.790	0.901	1.864
XGBoost	0.874	0.970	0.790	0.913	0.831	0.915	3.766
BiLSTM	0.862	0.976	0.834	0.850	0.834	0.927	1 180.93
BiXGB	**0.910**	**0.982**	**0.858**	**0.930**	**0.890**	**0.953**	**67.407**

各模型在储层岩性数据集上实验结果的 ROC 曲线图如图 4 所示.

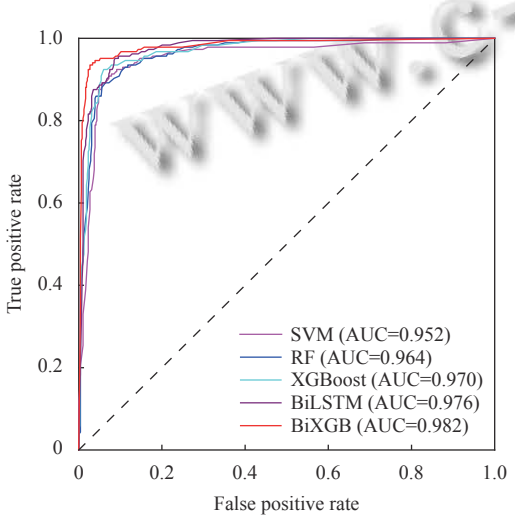


图 4 各模型在 Occupancy 数据集中的 ROC 曲线图

从表 3 和图 4 分析可知, BiXGB 模型在岩性数据集下与 XGBoost 相比分类效果显著提高. 与 BiLSTM 相比本模型的训练时间更短, 分类效果也显著提高. 这表明使用 BiLSTM 进行数据的特征提取并用机器学习方法进行分类的方法可行性.

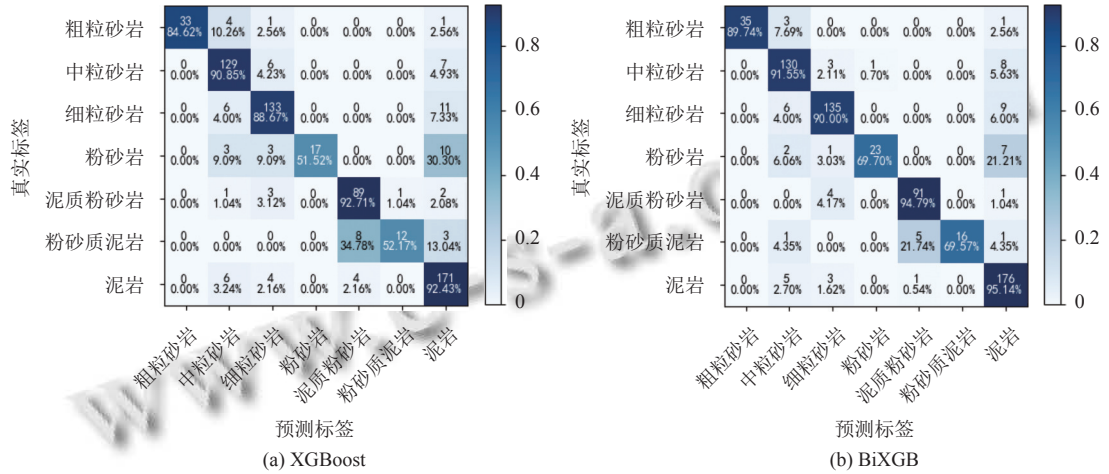
相较于传统的 BiLSTM 模型, BiXGB 模型去除了全连接层, 减少了参数数量, 从而降低了模型的复杂性. 此外, XGBoost 作为模型的最后一层, 能够有效地处理大规模数据集, 提升了整体模型的训练和推理速度. 由于去除了全连接层, BiXGB 模型的训练时间相对较短, 提高了模型的实时性能. 在实际应用中, 地质专业人员可能需要对大量的测井数据进行岩性分类, 因此快速的训练过程可以节省时间和资源.

与传统的机器学习方法相比, 本模型的分类效果更好, 关键在于 BiLSTM 可以很好地提高学习的长期

依赖性, 更好地提取数据特征. 并且 XGBoost 具有很好的数据分类能力以及较好的实时性, 本模型结合两个算法的优点从而有效地提高了模型在测井数据上的储层岩性分类准确度.

在此仅给出 XGBoost 和 BiXGB 模型的混淆矩阵用于对比模型在各种岩性上的识别准确率, 两个模型在岩性数据上的识别结果的混淆矩阵如图 5 所示. 由图 5 混淆矩阵可知, 本模型与 XGBoost 相比对储层各种岩性的识别准确率都得到了提升.

部分岩性识别效果如图 6 所示. 从图 6 中可以看出相较于单一的模型, 本文提出 BiXGB 模型对该研究区岩性的识别结果更符合真实研究区岩性分布情况.



(a) XGBoost (b) BiXGB

图 5 XGBoost 和 BiXGB 模型混淆矩阵

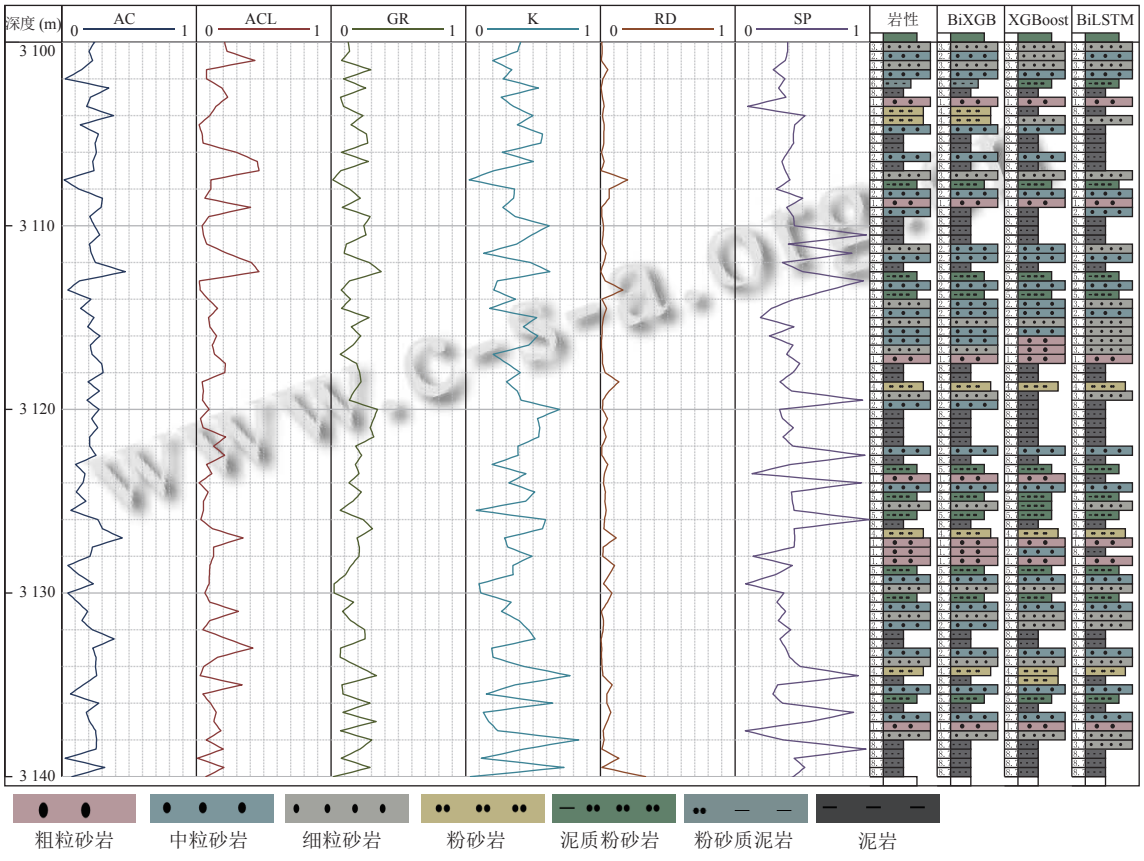


图 6 测试集岩性识别结果

(2) Occupancy 数据集分类

为了验证模型的实用性和准确性, 研究中将 BiXGB 模型应用于 Occupancy 数据的分类问题中, 实验中同样选用支持向量机 (SVM), 随机森林 (RF)、XGBoost、BiLSTM 与 BiXGB 几种分类方法进行分类效果对比实验结果如表 4 所示.

表 4 各模型在 Occupancy 数据集上实验结果比较

模型	*Accuracy*	AUC	*Recall*	*Precision*	*F*1	mAP	Time (s)
SVM	0.751	0.970	0.581	0.608	0.575	0.945	1.294
RF	0.848	0.971	0.852	0.802	0.813	0.951	1.787
XGBoost	0.918	0.990	0.783	0.811	0.791	0.963	1.367
BiLSTM	0.902	0.983	0.860	0.848	0.852	0.971	110.416
BiXGB	**0.930**	**0.991**	**0.860**	**0.940**	**0.890**	**0.973**	**15.891**

各模型在 Occupancy 数据集上实验结果的 ROC 曲线图如图 7 所示.

由表 4 和图 7 分析可知, 对于 Occupancy 数据集中的分类问题 BiXGB 模型也表现出了不错的分类效果, 结果表明本模型具有较好的准确性和稳定性.

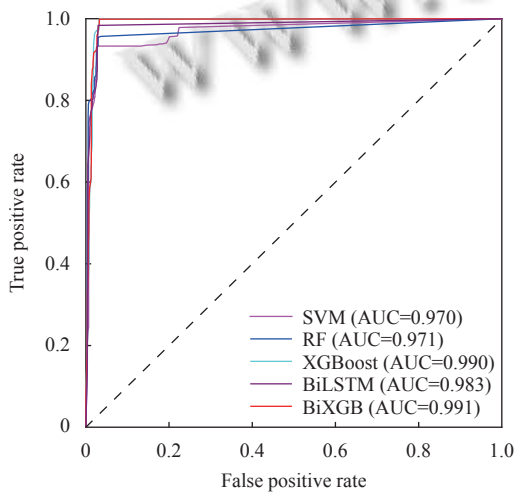


图 7 各模型在 Occupancy 数据集中的 ROC 曲线图

5 结论

本文针对在储层岩性识别中单一模型精度不足、神经网络模型训练时间长以及机器学习模型在数据特征学习方面的不足等问题, 提出了一种基于 BiLSTM-XGBoost 混合模型的储层岩性识别模型 BiXGB. 该模型通过在传统 XGBoost 的基础上引入 BiLSTM 作为特征提取模块, 提升了模型的特征提取能力, 使得模型在训练时可以学习到更丰富的测井数据特征. 在储层岩性数据集上测试精度达到了 91.0%. 相较于传统的岩性识别模型拥有更高的准确性和稳定性. BiXGB 模型为油气勘探领域提供了更高效、准确的储层岩性识别方法, 能够更好地满足油气勘探的实际需要.

在未来的研究中, 将进一步去解决储层岩性类别不均衡和测井数据量少导致某些岩性识别准确率不高的问题, 同时, 将研究一种智能优化算法, 用于优化模型参数, 进一步提升模型在储层岩性分类任务上的表现.

参考文献

1 Liu JJ, Liu JC. An intelligent approach for reservoir quality evaluation in tight sandstone reservoir using gradient boosting decision tree algorithm —A case study of the Yanchang Formation, mid-eastern Ordos Basin, China. Marine and Petroleum Geology, 2021, 126: 104939. [doi: 10.1016/j.marpetgeo.2021.104939]

2 Gu YF, Bao ZD, Song XM, *et al*. Complex lithology prediction using probabilistic neural network improved by continuous restricted Boltzmann machine and particle swarm optimization. Journal of Petroleum Science and Engineering, 2019, 179: 966–978. [doi: 10.1016/j.petrol.2019.05.032]

3 Xie YX, Zhu CY, Hu RS, *et al*. A coarse-to-fine approach for intelligent logging lithology identification with extremely randomized trees. Mathematical Geosciences, 2021, 53(5): 859–876. [doi: 10.1007/s11004-020-09885-y]

4 Liu HN, Wu YP, Cao YC, *et al*. Well logging based lithology identification model establishment under data drift: A transfer learning method. Sensors, 2020, 20(13): 3643. [doi: 10.3390/s20133643]

5 Lu XC, Sun D, Xie XY, *et al*. Microfacies characteristics and reservoir potential of Triassic Baikouquan Formation, northern Mahu sag, Junggar Basin, NW China. Journal of Natural Gas Geoscience, 2019, 4(1): 47–62. [doi: 10.1016/j.jnggs.2019.03.001]

6 Zhao ZX, He YB, Huang X, *et al*. Study on fracture characteristics and controlling factors of tight sandstone reservoir: A case study on the Huagang formation in the Xihu depression, East China Sea Shelf Basin, China. Lithosphere, 2021, 2021(S1): 3310886.

7 Bergen KJ, Johnson PA, De Hoop MV, *et al*. Machine learning for data-driven discovery in solid Earth geoscience. Science, 2019, 363(6433): eaau0323. [doi: 10.1126/science.aau0323]
8 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. 计算机工程与应用, 2018, 54(24): 260–265. [doi: 10.3778/j.issn.1002-8331.1709-0084]
9 Singh H, Seol Y, Myshakin EM. Automated well-log processing and lithology classification by identifying optimal features through unsupervised and supervised machine-learning algorithms. SPE Journal, 2020, 25(5): 2778–2800. [doi: 10.2118/202477-PA]
10 Bressan TS, De Souza MK, Girelli TJ, *et al*. Evaluation of machine learning methods for lithology classification using geophysical data. Computers & Geosciences, 2020, 139: 104475.
11 Xie YX, Zhu CY, Lu Y, *et al*. Towards optimization of boosting models for formation lithology identification. Mathematical Problems in Engineering, 2019, 2019: 5309852.
12 潘少伟, 王朝阳, 张允, 等. 基于长短期记忆神经网络补全测井曲线和混合优化 XGBoost 的岩性识别. 中国石油大学学报 (自然科学版), 2022, 46(3): 62–71.
13 Zhou KB, Zhang JY, Ren YS, *et al*. A gradient boosting decision tree algorithm combining synthetic minority oversampling technique for lithology identification. Geophysics, 2020, 85(4): WA147–WA158. [doi: 10.1190/geo2019-0429.1]
14 Dev VA, Eden MR. Formation lithology classification using scalable gradient boosted decision trees. Computers & Chemical Engineering, 2019, 128: 392–404.
15 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. 石油地球物理勘探, 2019, 54(2): 447–455.
16 尹生阳, 曾维, 王胜, 等. 基于声波信号的岩性智能分类方法. 吉林大学学报 (地球科学版), 2022, 52(6): 2060–2070.
17 周雪晴, 张占松, 朱林奇, 等. 基于双向长短期记忆网络的流体高精度识别新方法. 中国石油大学学报 (自然科学版), 2021, 45(1): 69–76.
18 王庆凯. 基于长短期记忆网络和时空序列模型的岩性识别方法研究 [硕士学位论文]. 秦皇岛: 燕山大学, 2022.
19 De Baets L, Ruyssinck J, Peiffer T, *et al*. Positive blood culture detection in time series data using a BiLSTM network. arXiv:1612.00962, 2016.
20 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]
21 Chen TQ, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco: ACM, 2016. 785–794.
22 Sayed AN, Himeur Y, Bensaali F. Deep and transfer learning for building occupancy detection: A review and comparative analysis. Engineering Applications of Artificial Intelligence, 2022, 115: 105254. [doi: 10.1016/j.engappai.2022.105254]
23 Alyasin EI, Ata O, Mohammedqasim H. Novel hybrid classification model for multi-class imbalanced lithology dataset. Optik, 2022, 270: 170047. [doi: 10.1016/j.ijleo.2022.170047]

(校对责编: 孙君艳)