DOI:10.3969/j.issn.1672-9854.2024.02.007 文章编号:1672-9854(2024)-02-0188-09

基于ADASYN-GS-XGBOOST混合模型的火山岩测井岩性识别

宋梓豪[1],巩红雨[2],冉爱华[2],杨鹏辉[2],刘迪仁[1]
1 长江大学油气资源与勘探技术教育部重点实验室;2 中国石油华北油田公司第三采油厂

摘 要 火山岩的形成环境复杂,有些地区的火山岩可能只发育两三种岩石类型,这会导致不同岩性取心资料的代表性严重失衡。针对现有的测井岩性识别方法在处理类间不均衡样本时出现效果较差的问题,提出基于ADASYN-GS-XGBOOST混合模型的火山岩岩性识别方法。首先通过ADASYN过采样算法对不均衡样本进行处理得到新的样本集,再以XGBOOST算法作为基分类器对样本进行分类,并利用网格搜索法(GS)对模型进行参数优化,以此建立ADASYN-GS-XGBOOST混合岩性识别模型。将该混合模型训练后的结果与K近邻、朴素贝叶斯、随机森林、XGBOOST及SMOTE-GS-XGBOOST等算法的岩性识别结果进行对比,表明基于ADASYN-GS-XGBOOST算法建立的模型识别效果最好。该方法克服了已有岩性识别方法无法有效解决不均衡样本的问题,极大地提升了火山岩岩性识别的准确率。

关键词 ADASYN算法;XGBOOST算法;混合模型;火山岩;测井;岩性识别
中图分类号:P631.8 **文献标识码**:A
引用:宋梓豪,巩红雨,冉爱华,等. 基于ADASYN-GS-XGBOOST混合模型的火山岩测井岩性识别[J]. 海相油气地质,2024, 29(2): 188-196.
SONG Zihao, GONG Hongyu, RAN Aihua, et al. Lithology logging identification of volcanic rock based on ADASYN-GS-XGBOOST hybrid model[J]. Marine origin petroleum geology, 2024, 29(2): 188-196.

0 前 言

随着油气勘探的难度增大,以及油气资源的需求量日益增多,在碎屑岩、碳酸盐岩等储层中寻找油气藏已无法满足当今需求,因此,针对火山岩油气储层的研究和勘探开发就显得尤为重要[1-7]。岩性的识别是火山岩油气资源勘探开发的重要内容。利用钻井取心是识别火山岩岩性成功率最高的方法,但受限于取心成本昂贵及取心率不高,往往无法获得全井段的岩心,因此,在勘探开发过程中借助测井资料对目的层进行辅助解释是研究火山岩储层的重要手段。但火山岩复杂的沉积环境导致岩性测井参数划分界限模糊,而且测井曲线信息往往出现大量的冗余,这对岩性识别过程造成一定的干扰。除此之外,有些地区的钻井取心样本可能只出现两三种岩性,造成了样本数据集的不均衡问题,给火山岩岩性的精准划分带来巨大的困难。

近些年来,随着机器学习技术的快速发展,针对多分类不均衡样品问题的解决思路有很多种,但应用于测井岩性识别领域的研究相对较少。针对岩性样本数据不均衡的问题,常应用于均衡数据集的有过采样和欠采样等算法,其中SMOTE算法作为一种经典的过采样算法被前人广泛使用。李雄飞等[8]提出不均衡数据挖掘算法PCBoost,通过数据合成的方法逐步增加少数类的合成样例,并通过修正被子分类器错分的合成样例,减少了不恰当的人工合成样例对集成分类器的影响;王光宇等[9]提出一种基于近邻清除算法(NM)和过采样算法(SMOTE)相结合形成的NM-SMOTE算法,实现了对多数类样本欠采样、同时对少数类样本过采样;罗仁泽等[10]

收稿日期:2023-10-18;改回日期:2024-03-25
本文受国家重点研发计划项目"深地资源勘探开发"(编号:2018YFC060330502)资助
第一作者:宋梓豪,硕士研究生,主要从事地球物理测井技术研究及应用。通信地址:430100湖北省武汉市蔡甸区大学路111号;E-mail:1298646764@qq.com
通信作者:刘迪仁,博士,教授,主要从事电法测井正反演、煤层气和复杂储层测井评价及光纤传感技术等方面的理论和应用研究。通信地址:430100 湖北省武汉市蔡甸区大学路111号;E-mail: liudr@yangtzeu.edu.cn

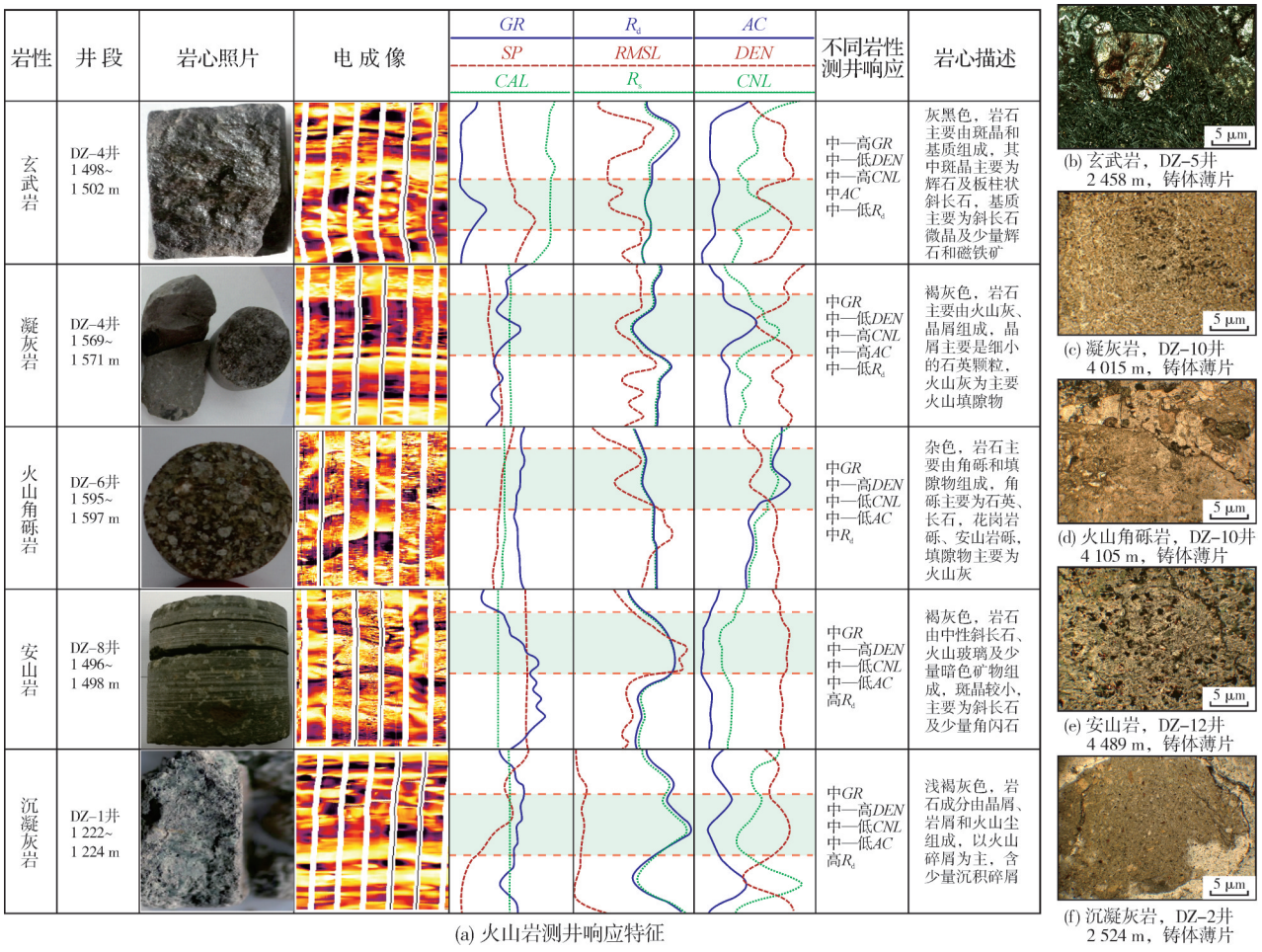
将K-means聚类分析算法和SMOTE算法结合，有效降低少数类样本被误分的风险。综合来说，上述方法只考虑了单纯通过增加少数类样本来均衡样本数量，并没有考虑到数据集各类别之间的不均衡特征，因此易导致少数类样本产生过拟合，降低了模型对样本的泛化能力。

为了解决上述问题，本文利用自适应合成采样算法（ADASYN）来对火山岩不均衡样本进行采样，提出一种基于ADASYN-GS-XGBOOST混合模型的火山岩岩性识别方法。以DZ区块的实际测井资料为基础，总结出不同类型火山岩测井响应特征，并根据敏感参数公式对不同测井曲线的敏感程度进行排名；同时依据ADASYN算法对5类火山岩非均衡样本进行处理，实现样本均衡化；最后采用XGBOOST（极限梯度提升树）算法作为基分类器建立火山岩岩性识别模型，并将识别结果与K近邻、朴素贝叶斯、随机森林、XGBOOST等算法进行对比和分析，验证了ADASYN-GS-XGBOOST混合模型应用于测井岩性识别的可行性和优越性。

1 火山岩类型及测井响应特征

1.1 岩性类型和电性特征

本文的研究数据来源于DZ区块的火山岩储层。DZ区块海拔位置相对较高，多发育隆起，经过长期的风化、溶蚀以及后期的构造运动，火山岩裂缝以及次生孔隙较发育，油气资源丰富。根据岩心资料和薄片鉴定结果（图1），识别出研究区主要发育凝灰岩、火山角砾岩、安山岩、玄武岩、沉凝灰岩5种火山岩类型。除此之外，研究区还发育流纹质凝灰岩、凝灰质砂岩、安山质角砾岩等，但由于样本较少，本文不作为研究对象。不同岩石类型的含量如图2所示。



(a) 火山岩测井响应特征

(b) 玄武岩，DZ-5井 2 458 m，铸体薄片

(c) 凝灰岩，DZ-10井 4 015 m，铸体薄片

(d) 火山角砾岩，DZ-10井 4 105 m，铸体薄片

(e) 安山岩，DZ-12井 4 489 m，铸体薄片

(f) 沉凝灰岩，DZ-2井 2 524 m，铸体薄片

图1 DZ区块部分代表井段火山岩岩性及电性特征

Fig. 1 Lithological and electrical characteristics of volcanic rocks in some representative well sections of DZ block

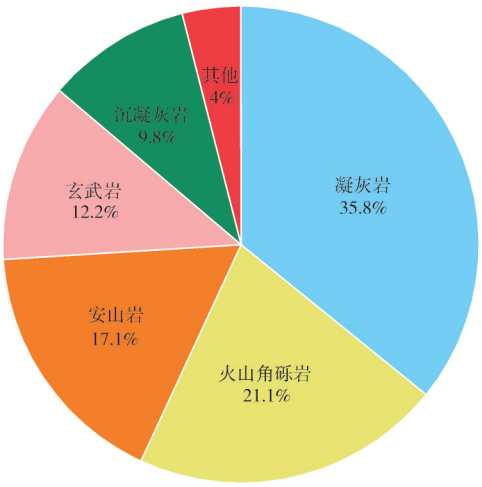


图2 DZ区块钻井取心段火山岩岩性统计饼状图

Fig. 2 Statistical pie chart of volcanic rock lithology in drilling core sections of DZ block

1.2 敏感参数选择

岩性识别模型的预测精度与所选取输入曲线的类型直接相关，故优选出对火山岩岩性敏感的测井曲线尤为重要。孙兴刚等[11]提出流体敏感参数公式来定义不同岩石物理量对流体的敏感程度，为流体敏感参数的优选提供了思路。以此为基础，通过统计不同测井曲线的响应特征的变化来构建岩性敏感参数公式：

$$LS=\left|\frac{\overline{A_m}-(\overline{A_1}+\ldots\overline{A_n})/n}{[\overline{A_m}]}\right| \tag{1}$$

式(1)中：$\overline{A_m}$表示凝灰岩测井曲线A的平均测井响应值(由于研究区块凝灰岩的测井响应特征不明显，故以其为量纲标准)；$\overline{A_1}\cdots\overline{A_n}$表示其他类火山岩测井曲线$A$的平均测井响应值；$LS$表示岩性敏感程度，$LS$值越大，表示测井曲线$A$对岩性的敏感程度越高。

根据敏感参数公式计算出火山岩取心样本中7种测井曲线的敏感系数，结果如图3所示：敏感程度排名由高到低依次为深侧向电阻率(R_t)、自然伽马(GR)、光电截面指数(P_e)、补偿中子(CNL)、补偿密度(DEN)、声波时差(AC)、井径(CAL)。由于CAL数据趋于单一值，对模型帮助不大，故选择R_t、GR、P_e、CNL、DEN、AC曲线作为岩性识别模型的输入曲线。

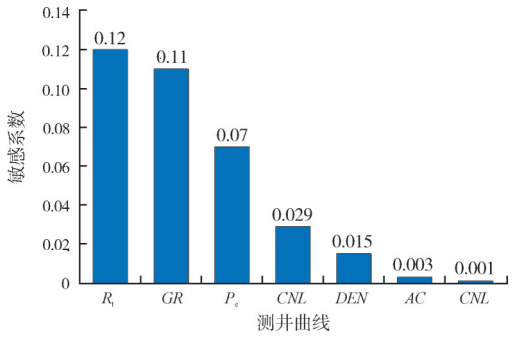


图3 DZ区块测井曲线敏感系数统计直方图

Fig. 3 Statistical histogram of sensitivity coefficient of logging curve in DZ block

1.3 火山岩测井响应特征分析

对所优选的6类测井曲线的响应特征信息进行提取，建立不同类型火山岩的测井响应特征统计库(表1)和特征分布小提琴图(图4)。由表1和图4可知，自然伽马曲线和中子曲线在区分玄武岩和火山角砾岩时的效果较好，光电截面指数在区分火山角砾岩和安山岩时也具有良好的效果，但单一的测井曲线只能简单地对2类或3类岩性进行区分，而实际地层中的火山岩岩性往往十分复杂，只利用单一测井曲线识别岩性效果较差。

为了研究二维测井曲线交会图识别火山岩的效果，优选出对岩性比较敏感的GR、CNL、DEN以及P_e参数作为指示曲线，建立GR-CNL、GR-P_e、GR-DEN、P_e-CNL、P_e-DEN以及CNL-DEN的交会图(图5)，对火山岩岩性进行分析。由图5可知，玄武岩具有高密度、低自然伽马的响应特征，GR-DEN交会图对玄武岩识别效果最好；火山角砾岩具有低光电截面指

表1 DZ区块火山岩测井响应特征统计

Table 1 Statistics of logging response of volcanic rock in DZ block

岩性	GR/API	P_e/(b·e^{-1})	R_t/(Ω·m)	AC/(μs·m^{-1})	CNL/%	DEN/(g·cm^{-3})
玄武岩	8.3~39.2	3.18~5.02	98.6~319.4	165.6~264.2	12.1~21.7	2.37~2.75
凝灰岩	19.8~79.1	3.68~4.40	45.5~461.6	166.3~292.3	12.3~36.2	2.26~2.66
火山角砾岩	42.2~82.3	2.66~3.54	124.2~439.2	171.3~291.3	21.4~37.3	2.29~2.69
安山岩	23.6~46.7	3.99~5.21	130.4~566.3	174.1~231.2	11.3~25.7	2.35~2.72
沉凝灰岩	31.2~77.5	4.21~6.31	31.2~118.3	202.2~309.1	16.1~30.1	2.34~2.59

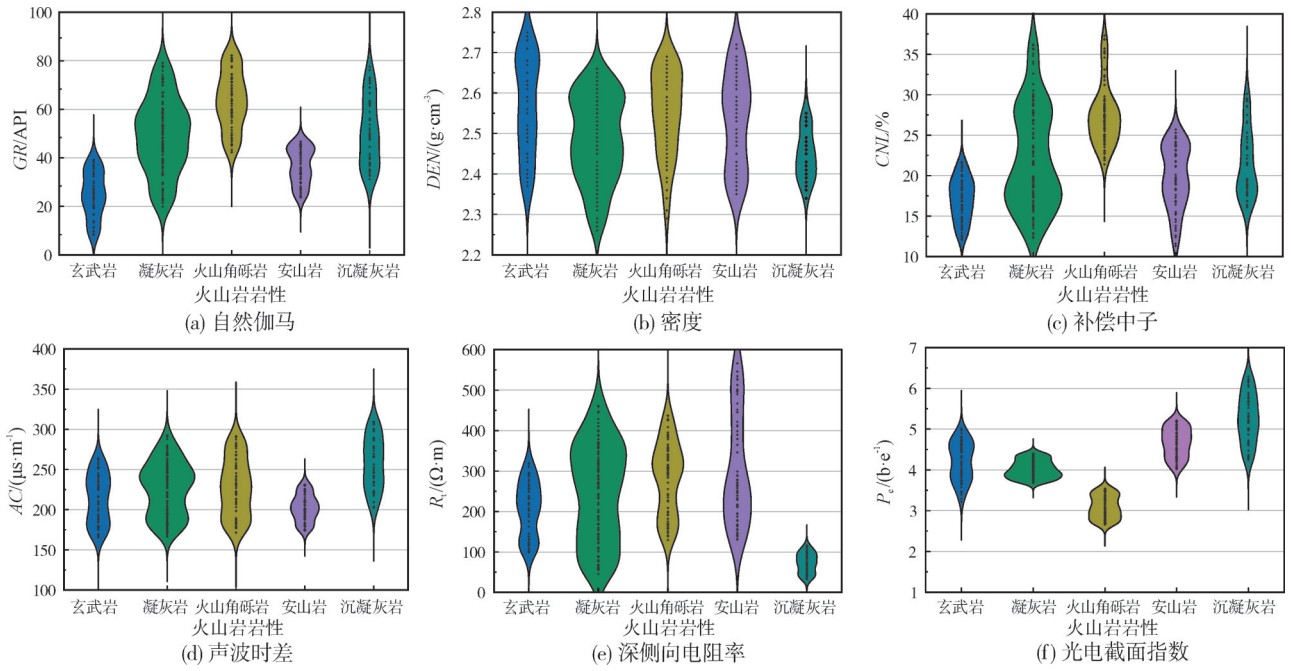


图4　DZ区块火山岩测井响应值的特征分布小提琴图

Fig. 4　Violin diagram showing the characteristic distribution of the logging response values of volcanic rock in DZ block

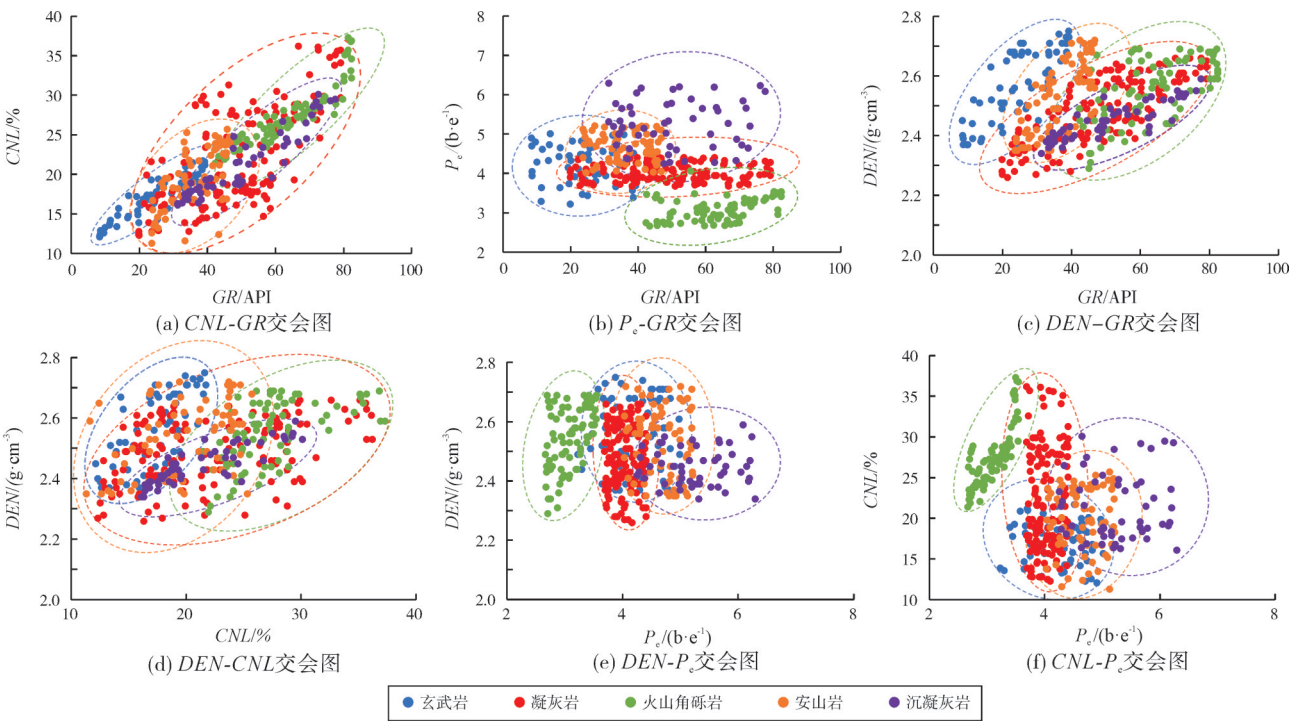


图5　DZ区块测井响应值交会图

Fig. 5　Crossplots of logging response values in DZ block

数的响应特征，P_e系列的交会图对火山角砾岩的识别效果均较好；沉凝灰岩具有高光电截面指数、高自然伽马的响应特征，GR-P_e交会图对沉凝灰岩识别效果最好；安山岩和凝灰岩相比于其他岩性的测井响应特征并不明显，其取值空间具有较大的重叠部分。总之，交会图对部分岩性具有较好的识别效果，但安山岩和凝灰岩利用交会图识别的效果较差。测井曲线交会图法在实际岩性识别过程中耗

时长且无法对复杂岩性进行高效识别，因此需要利用机器学习快速准确地识别火山岩岩性。

2 算法原理

2.1 ADASYN原理

自适应合成采样算法（ADASYN）是基于过采样算法（SMOTE）所改进的算法[12-13]。ADASYN算法的基本原理是通过少数类样本周围的分布密度来生成新的合成样本，相比于SMOTE算法的优点是能够自适应地生成合成样本，从而准确地反映样本的分布情况。

对于训练集 $T=\{(x_1,y_1),(x_2,y_2)\cdots(x_i,y_i)\}$，$i=1,\cdots,m$，定义 m_s 和 m_l 分别为少数类样本和多数类样本的数量。ADASYN算法步骤如下：

（1）计算少数类样本和多数类样本间的不平衡度α：

$$\alpha=\frac{m_s}{m_l} \tag{2}$$

若 $\alpha<\alpha_{th}$（α_{th} 是不平衡度的最大阈值），则需要对少数类样本进行新样本合成，反之则无需合成新样本。

（2）计算少数类样本所需合成样本的总数 G：

$$G=(m_l-m_s)\times\beta \tag{3}$$

式(3)中：$\beta\in[0,1]$，表示控制样本的均衡水平，当 $\beta=1$ 时，多数类样本数量等于少数类样本。

（3）对于少数类的每个样本 x_i，利用欧几里得距离找出 K 个最近邻样本，并计算出 K 近邻样本中多数类的占比 r_i。

（4）对 r_i 进行标准化处理：

$$\rho_i=r_i\times\left(\sum_{i=1}^{m_s}r_i\right)^{-1} \tag{4}$$

式(4)中：ρ_i 为概率密度，$\sum_i\rho_i=1$。

（5）根据样本的分布密度，计算出每个少数类样本所需生成新样本的数目 g_i：

$$g_i=\rho_i\times G \tag{5}$$

（6）根据SMOTE算法合成出每个少数类样本所需合成的新样本，直至新样本数量满足 g_i。SMOTE的计算公式如下：

$$S_i=x_i+(x_{zi}-x_i)\times\varepsilon \tag{6}$$

式(6)中：S_i 为合成样本，x_{zi} 表示的是 x_i 的 K 个近邻样本中任一一个少数类样本，$(x_{zi}-x_i)$ 为 n 维空间的差向量，ε 为[0，1]之间的随机数。

2.2 网格搜索法

网格搜索法（GS）是机器学习中常用的一种穷举遍历算法。在给模型调参过程中，先给定超参数一个较广的搜索范围，来寻找全局最优值可能的位置，然后通过改变步长逐渐缩小搜索范围，并利用K折验证法对每个参数组合进行验证，从而找到最优的模型参数组合。

2.3 XGBOOST原理

极限梯度提升树（XGBOOST），是基于梯度提升树（GBDT）的改进算法。相对于GBDT算法，XGBOOST算法能够进行并行化运算，有效地提高了模型的运算速度，并且在目标函数中增加了正则项，减少了目标函数的复杂程度，避免模型出现过拟合的现象[14]。其算法训练过程如图6所示。

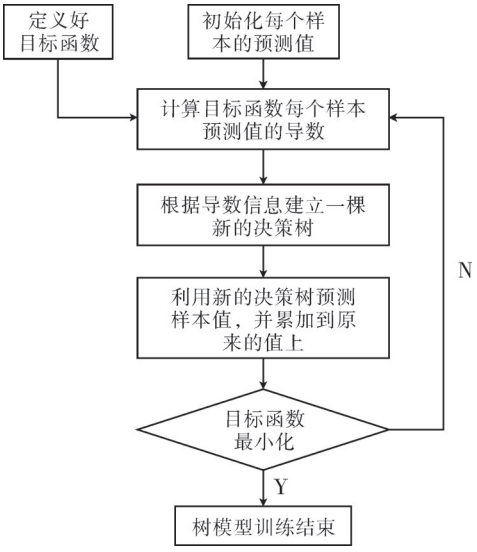


图6 XGBOOST算法的流程图

Fig. 6 Flow chart of XGBOOST algorithm

2.4 ADASYN-GS-XGBOOST混合模型构建

在建立XGBOOST岩性识别模型时，分类器存在将样本判别为多数类的倾向，故存在多数类样本预测准确率较高、少数类样本预测准确率较低的现象，这大大影响了分类器的性能。针对上述问题，本文采用ADASYN算法对少数类样本进行过采样，使得样本数量达到均衡，并基于XGBOOST分类器提出ADASYN-GS-XGBOOST混合模型。其模型框架如图7所示。

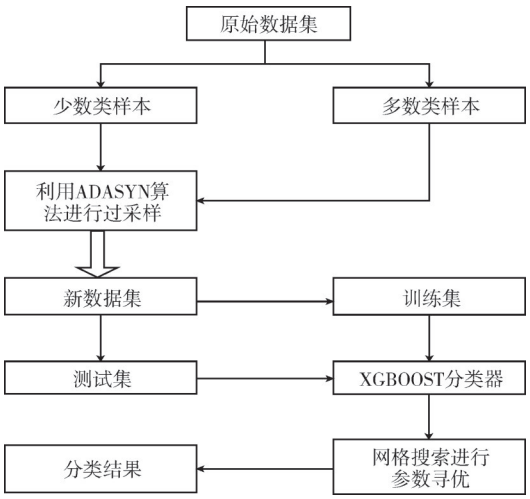


图7 ADASYN-GS-XGBOOST混合模型的流程图

Fig. 7 Flow chart of ADASYN-GS-XGBOOST hybrid model

3 基于ADASYN-GS-XGBOOST混合模型的岩性识别

3.1 数据预处理

DZ区块共有13口取心井，选择其中11口井作为样本井，2口井作为验证井。首先，将样本井中所有井段的火山岩岩性划为玄武岩、安山岩、火山角砾岩、凝灰岩、沉凝灰岩5类，分别用0、1、2、3、4表示，并且结合岩心数据完成研究区火山岩测井曲线岩性标签的确定，共获得410个岩性标签样本。因为不同测井曲线参数量纲及数量级不一致，因此，在过采样前需对标签样本数据进行归一化处理。其次，通过Python软件实现ADASYN过采样，使处理后少数类样本数量与多数类样本数量达到均衡。处理后的结果如表2所示。

表2 不同岩性样本经处理后的样本统计

Table 2 Sample statistics of different lithology samples after processing

岩性	类别	样本数	
		处理前	ADASYN处理后
玄武岩	0	52	155
凝灰岩	1	153	153
火山角砾岩	2	90	157
安山岩	3	73	149
沉凝灰岩	4	42	154
总计		410	768

3.2 基于网格搜索法的参数优化

影响XGBOOST模型的参数有多种，主要有：迭代次数t、决策树的最大深度D_{max}、学习率eta、惩罚项系数g、最小样本权重总和W_{min}、训练每棵树时使用的数据占全部训练集的比例S以及使用的特征占全部特征的比例C。将预处理后的岩性标签样本数据按7:3的比列划分为训练集和测试集，并利用网格搜索法对XGBOOST模型中的t、D_{max}、eta、g、W_{min}、S以及C参数按照特定的步长进行寻优。经过多次调试，模型最优参数如表3所示。

表3 ADASYN-GS-XGBOOST算法的参数设置

Table 3 Parameter settings of ADASYN-GS-XGBOOST algorithm

参 数	取值范围	步长	最优参数
决策树的最大深度(D_{max})	6~10	1	9
学习率(eta)	0.1~0.3	0.1	0.1
迭代次数(t)	10~500	10	250
惩罚项系数(g)	0~1	0.1	0
最小样本权重总和(W_{min})	0~10	1	1
使用的数据占全部训练集的比例(S)	0~1	0.1	0.8
使用的特征占全部特征的比例(C)	0~1	0.1	0.9

3.3 应用结果与对比

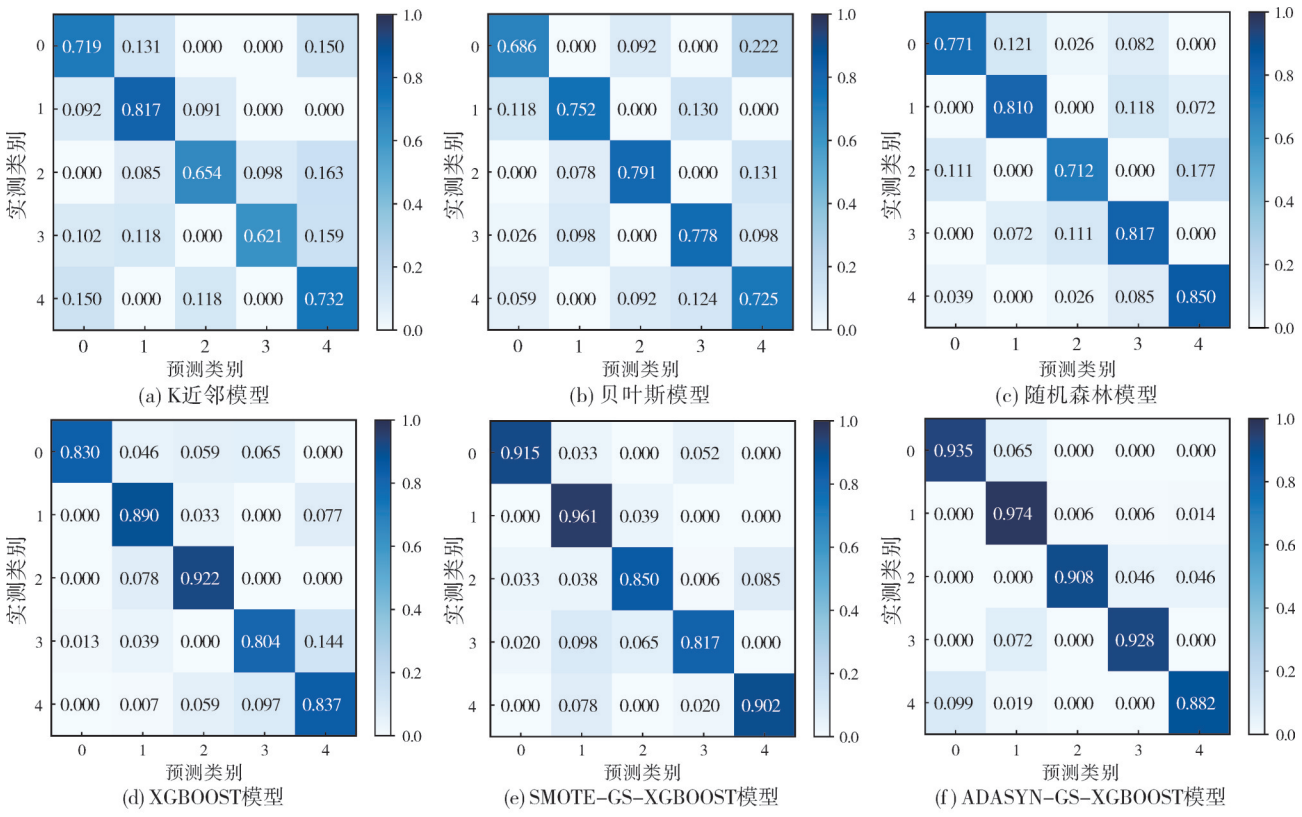
在确立了混合模型的最优参数组合后，引入混淆矩阵[15]中的准确率作为岩性识别模型的综合评价指标。将所得到的火山岩岩性的分类预测结果与K近邻、朴素贝叶斯、随机森林、XGBOOST及SMOTE-GS-XGBOOST算法在测试集上的预测结果进行对比（表4）。从表4可以看出：K近邻、朴素贝叶斯、随机森林及XGBOOST对火山岩的识别准确率分别为75%、79%、80%、83%；SMOTE-GS-XGBOOST模型的岩性识别准确率为86%；ADASYN-GS-XGBOOST模型的岩性识别准确率达到92%，相比XGBOOST模型提高9%，相比SMOTE-GS-XGBOOST模型提高6%。图8是基于不同岩性识别模型所建立的混淆矩阵图，图中横轴代表每种岩性被预测的个数，纵轴代表测试集中每种岩性的真实个数，对角线代表每种岩性被正确分类的比例。结果表明ADASYN-GS-XGBOOST模型相对于其他5种模型预测正确的比例最高，分类效果也更加理想。

为了进一步验证ADASYN-GS-XGBOOST模型对火山岩岩性的识别效果，采用DT-4井中的1 460~1 560 m深度段以及DT-6井中的960~1 040 m深度

表4 不同模型的岩性识别准确率统计

Table 4 Statistics on the accuracy of lithology identification for different models

模 型	玄武岩	凝灰岩	火山角砾岩	安山岩	沉凝灰岩	平均值
K近邻	0.77	0.67	0.78	0.80	0.74	0.75
朴素贝叶斯	0.72	0.84	0.64	0.88	0.88	0.79
随机森林	0.81	0.83	0.77	0.81	0.79	0.80
XGBOOST	0.73	0.89	0.92	0.81	0.80	0.83
SMOTE-GS-XGBOOST	0.80	0.86	0.88	0.87	0.89	0.86
ADASYN-GS-XGBOOST	0.95	0.97	0.96	0.85	0.88	0.92



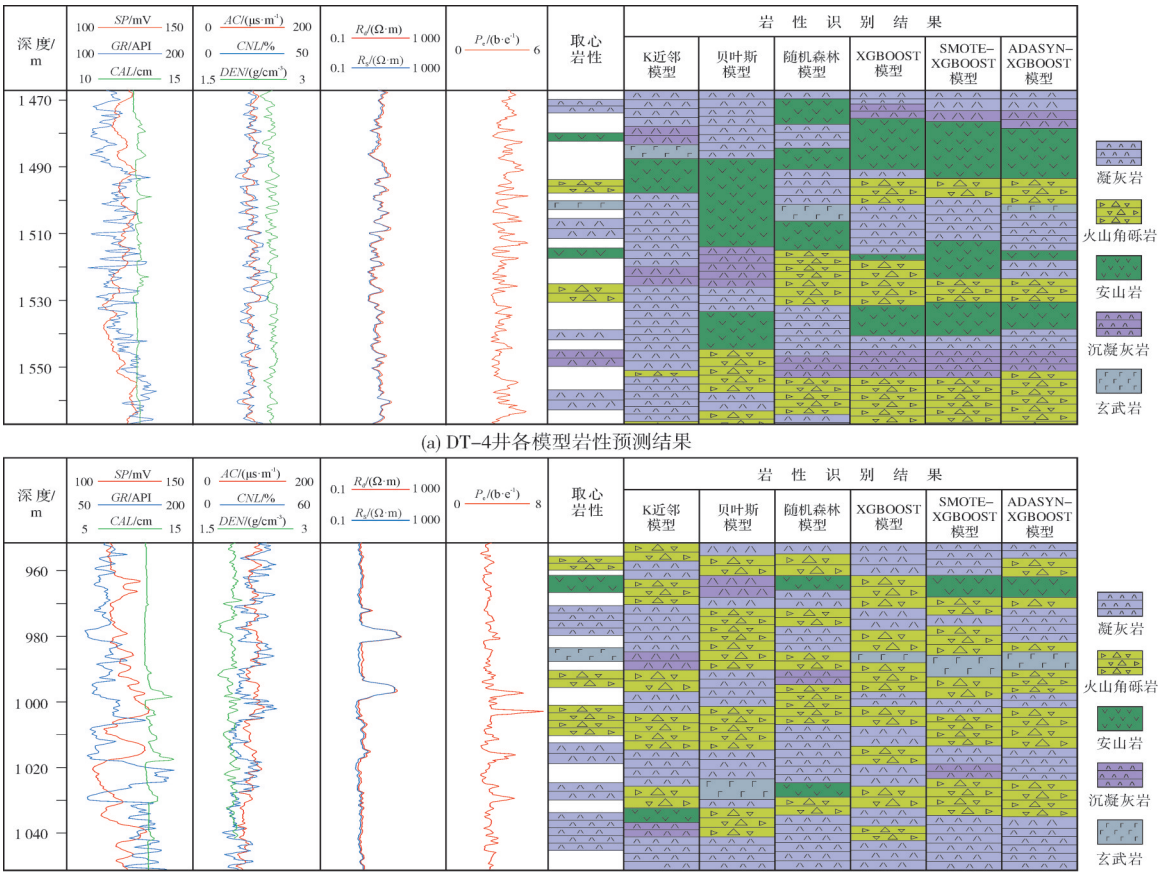
注：横坐标、纵坐标上的数字分别表示火山岩岩性：0-玄武岩，1-安山岩，2-火山角砾岩，3-凝灰岩，4-沉凝灰岩。

图8 不同火山岩岩性识别模型的混淆矩阵

Fig. 8 Confusion matrix of different lithology identification models of volcanic rocks

段作为岩性预测验证井段(图9)。图9中不同模型的岩性识别结果对比显示，ADASYN-GS-XGBOOST模型的识别总符合率最高，达到90%以上。其中，DT-4井取心段总厚度为42.9 m(图9a)，识别岩性与取心岩性相符的厚度为38.7 m，符合率为90.2%；DT-6井取心段总厚度为57.3 m(图9b)，识别岩性与取心岩性相符的厚度为51.8 m，符合率为90.4%。

在岩性识别过程中，由于部分凝灰岩和火山角砾岩分别含有少量的安山岩岩屑和凝灰质，导致其测井曲线特征比较相近，因此K近邻、贝叶斯和随机森林模型对于凝灰岩和火山角砾岩的识别容易出现误判，识别准确率均小于85%，而ADASYN-GS-XGBOOST对于凝灰岩和火山岩角砾岩的识别效果最好，识别准确率均达到了95%以上。ADASYN算法通过选择性增强特征相似类别的样本数量，实现了局部平衡，从而消除了传统SMOTE算法盲目均衡样本的弊端，并且解决了同类模型面对复杂储层难以分类的问题，进一步提高了模型的识别准确性。



(a) DT-4井各模型岩性预测结果

(b) DT-6井各模型岩性预测结果

图9　DT区块不同模型火山岩岩性识别效果验证

Fig. 9　Verification of lithology identification effect of volcanic rocks for different models in DT block

4　结　论

(1)针对研究区块不同岩性取心样本的数量不均衡问题,提出基于ADASYN-GS-XGBOOST混合模型的火山岩岩性识别方法。首先,根据DT区块的火山岩测井响应特征的变化建立岩性敏感参数公式,实现了对模型特征值的重要性排序,并优选出R_t、GR、P_e、CNL、DEN、AC曲线作为模型的输入参数;其次,采用ADASYN算法依据样本分布权重增加少数类样本数据量,降低样本不均衡度;最后,利用XGBOOST算法对火山岩岩性进行分类并通过网格搜索法对模型进行参数寻优得到最优岩性识别模型。

(2)ADASYN-GS-XGBOOST混合模型的识别准确率相比传统的机器学习方法至少提高了10%,且在DT-4井、DT-6井两口验证井上的总体符合率达到了90%以上,验证了该模型应用于不均衡样本情况下的火山岩测井岩性识别的可行性和有效性,可为此类问题的解决提供方法参考。

参 考 文 献

[1] 孙龙德, 邹才能, 朱如凯, 等. 中国深层油气形成、分布与潜力分析[J]. 石油勘探与开发, 2013, 40(6): 641-649.
SUN Longde, ZOU Caineng, ZHU Rukai, et al. Formation, distribution and potential of deep hydrocarbon resources in China [J]. Petroleum exploration and development, 2013, 40(6): 641-649.

[2] 唐华风, 王璞珺, 边伟华, 等. 火山岩储层地质研究回顾[J]. 石油学报, 2020, 41(12): 1744-1773.
TANG Huafeng, WANG Pujun, BIAN Weihua, et al. Review of volcanic reservoir geology [J]. Acta petrolei sinica, 2020, 41 (12): 1744-1773.

[3] 束景锐, 肖敦清, 苏俊青. 黄骅坳陷枣北地区沙三段火山岩油藏储层特征[J]. 特种油气藏, 1997, 4(3): 1-5.
SHU Jingrui, XIAO Dunqing, SU Junqing. Reservoir characteristics of Shasan-Section volcanic reservoir in Zaobei area of Huanghua Depression[J]. Special oil & gas reservoirs, 1997, 4(3): 1-5.

[4] 徐正顺, 王渝明, 庞彦明, 等. 大庆徐深气田火山岩气藏储集层识别与评价[J]. 石油勘探与开发, 2006, 33(5): 521-531.
XU Zhengshun, WANG Yuming, PANG Yanming, et al. Identification and evaluation of Xushen volcanic gas reservoirs in Daqing[J]. Petroleum exploration and development, 2006, 33 (5): 521-531.

[5] 李宁, 乔德新, 李庆峰, 等. 火山岩测井解释理论与应用[J].

石油勘探与开发, 2009, 36(6): 683-692.
LI Ning, QIAO Dexin, LI Qingfeng, et al. Theory on logging interpretation of igneous rocks and its application[J]. Petroleum exploration and development, 2009, 36(6): 683-692.
[6] FAN Cunhui, QIN Qirong, LIANG Feng, et al. Fractures in volcanic reservoir: a case study of Zhongguai uplift in northwestern margin of Junggar Basin, China[J]. Earth sciences research journal, 2018, 22(3): 169-174.
[7] ZHANG Xilong, XIA Yanqing, ZHANG Yan, et al. Volcanic reservoir characteristics and hydrocarbon genesis of Jiamuhe Formation in Jinlong 2 wellblock, Junggar Basin[J]. Petroleum science and technology, 2018, 36(19): 1516-1523.
[8] 李雄飞, 李军, 董元方, 等. 一种新的不平衡数据学习算法PCBoost[J]. 计算机学报, 2012, 35(2): 202-209.
LI Xiongfei, LI Jun, DONG Yuanfang, et al. A new learning algorithm for imbalanced Data-PCBoost[J]. Chinese journal of computers, 2012, 35(2): 202-209.
[9] 王光宇, 宋建国, 徐飞, 等. 不平衡样本集随机森林岩性预测方法[J]. 石油地球物理勘探, 2021, 56(4): 679-687.
WANG Guangyu, SONG Jianguo, XU Fei, et al. Lithology prediction method of random forest with unbalanced sample set[J]. Oil geophysical prospecting, 2021, 56(4): 679-687.
[10] 罗仁泽, 庹娟娟, 倪华玲, 等. 基于改进集成学习的测井岩性识别方法研究[J]. 石油物探, 2023, 62(2): 212-224.
LUO Renze, TUO Juanjuan, NI Hualing, et al. Logging lithology identification method based on improved ensemble learning[J]. Geophysical prospecting for petroleum, 2023, 62(2): 212-224.
[11] 孙兴刚, 魏文, 李红梅. 岩石物理参数的流体敏感性分析[J]. 油气藏评价与开发, 2012, 2(1): 37-40, 49.
SUN Xinggang, WEI Wen, LI Hongmei. Fluid sensitivity analysis of petrophysical parameters[J]. Reservoir evaluation and development, 2012, 2(1): 37-40, 49.
[12] 陈虹, 赵建智, 肖成龙, 等. 改进ADASYN-SDA的入侵检测模型研究[J]. 计算机工程与应用, 2020, 56(2): 97-105.
CHEN Hong, ZHAO Jianzhi, XIAO Chenglong, et al. Research on improved intrusion detection model of ADASYN-SDA[J]. Computer engineering and applications, 2020, 56(2): 97-105.
[13] 汪万敏, 智路平. 基于ADASYN-SFS-RF的欺诈检测模型泛化性能提升及可解释性研究[J]. 计算机应用研究, 2022, 39(12): 3605-3613.
WANG Wanmin, ZHI Luping. Fraud detection model generalization performance improvement and interpretability study based on ADASYN-SFS-RF[J]. Application research of computers, 2022, 39(12): 3605-3613.
[14] CHEN T Y, CHEN L C, CHEN Y M. Mining location-based service data for feature construction in retail store recommendation[C]//Advances in Data Mining. Applications and Theoretical Aspects. Cham: Springer International Publishing, 2017: 68-77.
[15] 孔英会, 景美丽. 基于混淆矩阵和集成学习的分类方法研究[J]. 计算机工程与科学, 2012, 34(6): 111-117.
KONG Yinghui, JING Meili. Research of the classification method based on confusion matrixes and ensemble learning[J]. Computer engineering and science, 2012, 34(6): 111-117.

编辑:黄革萍

Lithology logging identification of volcanic rock based on ADASYN-GS-XGBOOST hybrid model

SONG Zihao[1], GONG Hongyu[2], RAN Aihua[2], YANG Penghui[2], LIU Diren[1]
1. *Key Laboratory of Exploration Technologies for Oil and Gas Resources, Ministry of Education, Yangtze University*;
2. *Third Oil Production Plant of PetroChina Huabei Oilfield Company*

Abstract: The forming environment of volcanic rocks is complex, and the lithology of volcanic rocks in a certain area may be mainly composed of two or three types, which leads to serious imbalance of core data of different lithology. The existing lithology identification methods are not effective in dealing with unbalanced samples among classes. To solve these problems, a volcanic rock lithology identification method based on ADASYN-GS-XGBOOST hybrid model is proposed. The unbalanced samples are processed by ADASYN oversampling algorithm to obtain a new sample set, and then XGBOOST is used as the base classifier to classify the samples. The ADASYN-GS-XGBOOST hybrid lithology identification model is established by using Grid Search to optimize the parameters of the model. The results of the hybrid model training are compared with those of K nearest neighbor, naive Bayes, random forest, XGBOOST and SMOTE-GS-XGBOOST algorithms. The results show that the model based on ADASYN-GS-XGBOOST algorithm has the best identification effect. This method overcomes the problem that existing lithology identification methods can not effectively solve the problem of unbalanced samples, and greatly improves the accuracy of lithology identification of volcanic rocks.

Key words: ADASYN algorithm; XGBOOST algorithm; hybrid model; volcanic rocks; logging; lithology identification

SONG Zihao, First author: Master Candidate at Yangtze University, mainly engaged in geophysical logging technology and application. Add: No. 111 Daxue Rd., Caidian District, Wuhan, Hubei 430100, China. E-mail: 1298646764@qq.com
LIU Diren, Corresponding author: PhD, Professor, mainly engaged in the theoretical and applied research of forward and inversion of electric logging, well logging evaluation of coalbed methane and complex reservoirs and optical fiber sensing technology. Add: No. 111 Daxue Rd., Caidian District, Wuhan, Hubei 430100, China. E-mail: liudr@yangtzeu.edu.cn