【环境与能源 / Environment and Energy】

克拉玛依油田七区砂砾岩油藏智能岩性识别

陆吉[1,2]，林伯韬[1,2]，史璨[1,2]，张家豪[1,2]

1）中国石油大学(北京)石油工程学院，北京102249；2）中国石油大学(北京)人工智能学院，北京102249

摘　要：中国新疆克拉玛依油田七区砂砾岩油藏岩性多样，隔夹层发育，常规取心识别方法成本高，在非取心段识别精度低，导致储层划分困难．为实现岩性的快速准确识别，根据地质资料将研究区内岩性划分为泥岩、粗砂岩、中-细砂岩、粗砾岩、中-细砾岩和煤层6种．基于敏感性分析，优选测井参数提取主成分，建立岩性识别图版，识别准确率达81.37%．针对不均衡样本导致的少数类识别率低的问题，提出一种基于k均值聚类人工少数类过采样(k-means synthetic minority oversampling technique, KMSMOTE)与随机森林结合的智能岩性识别模型，通过对少数类样本过采样提升识别精度，该模型的识别准确率达到92.94%．将图版法和KMSMOTE-随机森林应用于邻井进行岩性识别并对比分析结果发现，KMSMOTE-随机森林识别准确率为95.71%，优于图版法的82.91%．同时，对各类岩性的识别准确率均高于传统的随机森林模型，证明KMSMOTE和随机森林结合的智能岩性识别模型在不均衡岩性样本识别问题上具有较好的适用性，泛化能力强，能够快速、准确地识别地层岩性．研究结果为不均衡岩性样本识别提供了智能化新思路．

关键词：地球物理；油田开发；测井；随机森林；岩性识别；主成分分析；不均衡数据集；砂砾岩储层

中图分类号：TE121　　**文献标志码：**A　　**doi：**10.3724/SP.J.1249.2023.03361

Intelligent lithologic identification of sandy conglomerate reservoirs in District No.7 of Karamay oilfield

LU Ji[1,2], LIN Botao[1,2], SHI Can[1,2], and ZHANG Jiahao[1,2]

1) College of Petroleum Engineering, China University of Petroleum (Beijing), Beijing 102249, P. R. China

2) College of Artificial Intelligence, China University of Petroleum (Beijing), Beijing 102249, P. R. China

Abstract: The sandy conglomerate reservoirs in Karamay are characterized by diverse lithology and interlayers. The cost of the conventional coring methods is high, and the identification accuracy in non-coring section is low, which leads to difficulty in reservoir classification. In order to achieve rapid and accurate identification of lithology, the lithology of the target area is classified into mudstone, coarse sandstone, medium-fine sandstone, coarse conglomerate, medium-fine conglomerate, and coal seam based on geological data. Firstly, the principal component analysis method is adopted to establish the lithology identification cross plot based on sensitivity analysis of well log data with an accuracy rate of 81.37%. Secondly, a lithology identification model is proposed based on the combination of k-means synthetic minority oversampling technique (KMSMOTE) and random forest to improve minority identification accuracy. The model improves the identification accuracy to achieve the accuracy of about 92.94% for oversampling minority samples. The two methods are applied to adjacent wells for comparative analysis of lithology identification. The accuracy of the KMSMOTE random forest is 95.71%, better than that of the cross plot method of 82.91%. The accuracy of minority sample identification is higher than that of the traditional random forest

Received: 2022-04-16; **Accepted:** 2022-08-04; **Online** (CNKI)**:** 2022-11-19

Foundation: National Natural Science Foundation of China (42277122); Science Foundation of China University of Petroleum (Beijing) (2462020YXZZ030)

Corresponding author: Professor LIN Botao. E-mail: linb_cupb@163.com

Citation: LU Ji, LIN Botao, SHI Can, et al. Intelligent lithologic identification of sandy conglomerate reservoirs in District No. 7 of Karamay oilfield [J]. Journal of Shenzhen University Science and Engineering, 2023, 40(3): 361-369. (in Chinese)

http://journal.szu.edu.cn

model, with excellent applicability and generalization ability on imbalanced lithology identification. The research results provide a new intelligent idea for imbalanced lithology sample identification.

Key words: geophysics; oil field development; well logging; random forest; lithology identification; principal component analysis; imbalance dataset; sandy conglomerate

岩性识别是油藏描述、开发、评价以及裂缝监测等工作的重要研究内容，能够揭示岩石地球物理性质与规律[1-2]．中国新疆克拉玛依砂砾岩油藏地层岩性多样、隔夹层发育，准确预测岩性是储层特征描述及地质建模的基础[3-5]．因此，开展岩性识别研究至关重要．

目前，传统方法主要通过交会图版[6-7]和聚类分析[8-9]进行岩性识别．然而，常规交会图版法未考虑数据可分性，识别结果受人为因素影响较大．此外，由于不同粒径的砂岩或砾岩测井参数差异较小，常规交会图版法往往难以依据单一参数针对两种岩性开展次级识别；聚类方法对可分性较低的数据同样出现较多误判．随着机器学习理论的发展，支持向量机、神经网络等多种方法迅速兴起并成为测井数据分析的重要手段．MOU 等[10-12]优选测井参数结合支持向量机进行岩性识别，结果证明该方法识别效果高于多元回归分析方法；武中原等[13-15]则考虑岩石沉积序列基于长短期记忆网络进行岩性识别，相比贝叶斯概率模型岩性识别准确率提升了 12%．上述两类方法在岩性识别中取得了一定的应用效果，但支持向量机的模型性能受核函数和惩罚因子影响较大，神经网络对测井数据的噪声较为敏感，且容易陷入局部最小，影响最终结果．

随机森林作为一种集成学习方法能够有效减少单一学习器造成的拟合问题，同时受数据异常值及噪声影响较小．TROTT 等[16]利用随机森林结合地球化学数据对区域岩性进行识别，与传统图版法相比识别率准确率高、耗时短；AO 等[17]提出一种概率随机森林模型并将其应用于岩性识别，结果表明该方法在薄层识别中表现出较大的优势．然而在机器学习模型预测过程中，受测井层段及地质特征等因素影响，岩性训练样本会呈现不均衡分布，导致模型对少数类的识别准确率下降．

综上，针对图版法受数据可分性影响问题，本研究采用主成分分析(principal component analysis，PCA)对测井数据进行降维提取，提高数据可分性，结合常规测井资料敏感性分析，建立岩性识别改进图版，有效识别砂砾岩次级岩性；同时，提出一种针对不均衡岩性样本的智能识别方法，基于 KMSMOTE 算法对少数类样本进行过采样均衡，将均衡后的样本集用于随机森林训练，构建岩性智能识别模型，实现不均衡岩性数据的快速准确分类．

1 油藏概况

克拉玛依七区油藏作位于准噶尔盆地西北缘克北断裂带内[18]，主要发育砂砾岩型油藏，油藏厚度大，平均埋深为 650～1 500 m．区域构造位于白碱滩南断裂上盘，受克乌断裂带和白碱滩断裂带共同控制作用，为西北向东南倾斜的单斜构造，顺倾斜方向逐渐变陡；其北中部地层倾角较缓，东南部下陷带倾角大，平均倾角 30°；区块东侧发育 5137 井断裂、BJ1162 井断裂和 G113 井断裂(图 1)．沉积自下而上经历了辫状河-辫状河三角洲的演化过程．主力油层 J_1b_4 和 $J_1b_5^1$ 层为河道沉积，发育河床滞留和心滩微相，J_1b_4 在七区中部发育泛滥平原微相，区域内主要发育泥岩、中-细砂岩、粗砂岩、中-细砾岩、粗砾岩和煤层，岩性纵向变化快，且粒径多样，隔夹层较为发育．

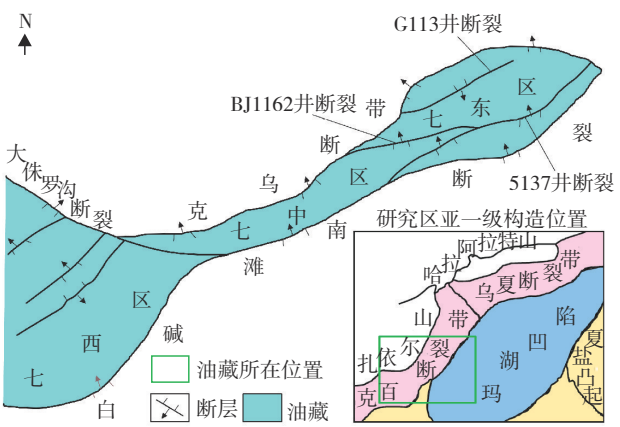


图 1 研究区域构造位置

Fig. 1 Geographical and structural position of research area.

2 原理与方法

由于成像测井数据获取成本较高，本研究采用常规测井数据针对克拉玛依油田开展岩性识别工作，分别为自然伽马(natural gamma ray，GR)、

http://journal.szu.edu.cn

自然电位(spontaneous potential，SP)、井径(borehole diameter，CALI)、地层真电阻率(true formation resistivity，RT)、侵入带电阻率(invaded zone resistivity，RI)、冲洗带电阻率(flushed zone formation resistivity，RXO)、声波时差(acoustic，AC)、密度(density，DEN)和补偿中子孔隙度(compensated neutron logging，CNL)，测井数据柱状图请扫描论文末页右下角二维码查看图S1.

基于该区块的测井数据及地质解释资料，分析不同岩性测井数据变化趋势和敏感性差异，优选岩性变化敏感参数. 考虑常规测井数据重叠区间，通过主成分变换增强数据可分性，建立基于交会图版的岩性智能识别方法. 此外，对测井数据进行过采样均衡，通过随机森林算法结合敏感参数构建岩性智能识别模型，对比两种方法应用效果. 其中，智能识别的流程如图2.

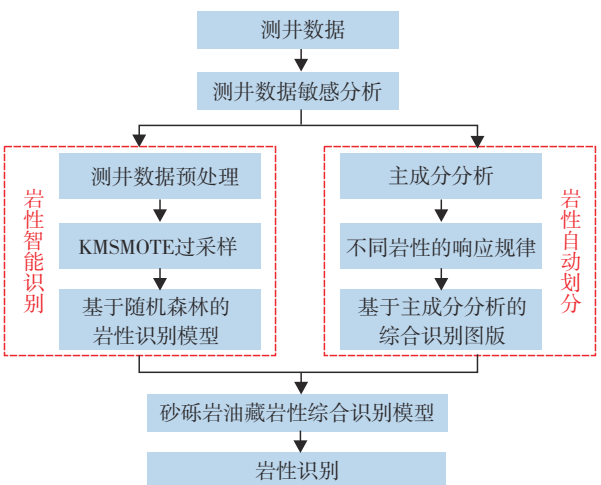


图2　智能岩性识别流程

Fig. 2　Flow chart of lithology identification.

2.1　PCA

PCA是数理统计中一种常用的方法[19]，通过对原始数据进行正交变换，将原始高维数据映射到低维超平面上得到主成分矩阵$\boldsymbol{F}$，变换公式为

$$\boldsymbol{F}=\boldsymbol{U}^{\mathrm{T}}\boldsymbol{X}=\left[\boldsymbol{F}_1,\boldsymbol{F}_2,\cdots,\boldsymbol{F}_p\right] \tag{1}$$

其中，$\boldsymbol{X}$为原始数据矩阵；$\boldsymbol{U}$为正交变换矩阵；$\boldsymbol{F}_i$为主成分列向量，$i=1,2,\cdots,p$，p为原始数据特征轴数.

主成分分析通过正交变换得到一系列不相关的综合变量，各变量方差大小反应了主成分所包含的信息量. 当方差达到某一数值时，认为该主成分能够表征原始数据. 因此，选取一定数量的主成分，可在最大程度保留原始数据特征的情况下达到数据降维.

2.2　随机森林

随机森林是一种基于决策树集成的有监督学习方法，它利用决策树组对数据集的不同子样本进行拟合，并通过对拟合结果进行均值计算来提高预测精度和控制过拟合问题. 对于分类问题，随机森林对于bagging抽样得到的k个训练集进行训练，得到子结果$\{t_1,t_2,t_3,\cdots,t_k\}$，使用投票或平均统计分析的形式给出最终分类结果[20]为

$$T(x)=\begin{cases}\arg\max\sum_{h=1}^{k}I(t_h(x)), & \text{投票法}\\ \frac{1}{k}\sum_{h=1}^{k}t_h(x), & \text{平均法}\end{cases} \tag{2}$$

其中，$T(x)$为样本x的分类结果；$t_h(x)$为样本x在第h个训练集中的分类结果；I为概率决策函数.

2.3　k均值聚类人工少数类过采样算法

人工少数类过采样(synthetic minority oversampling technique，SMOTE)算法是一种合成过采样方法，通过向数据集随机添加合成少数类样本来提高模型的训练效果[21]. 该方法能一定程度上均衡数据分布，然而对所有少数类样本插值，某些少数类异常样本会使得原有数据分类边界被破坏，对分类造成影响，见图3(a). 图3(a)中蓝点表示少数类样本；红点表示多数类样本；黑圈白点表示合成样本；黑圈蓝点表示少数类异常样本；黑色虚线表示原决策边界；绿色实线表示合成样本后决策边界. 针对这一问题，文献[22]提出了改进的k均值聚类人工少数类过采样(k-means synthetic minority oversampling technique，KMSMOTE)算法，该方法将k-means聚类与SMOTE算法结合，首先对数据集聚类

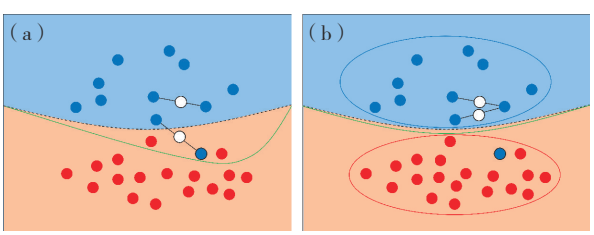


图3　(a) SMOTE算法和(b)KMSMOTE算法过采样示意图

Fig. 3　Oversampling plot of (a) SMOTE and (b) KMSMOTE. Blue dots represent the minority class of samples, red dots represent the majority class of samples, white dots with black egde represent synthetic samples, blue dots with black edge represent the abnormal samples, black dotted lines represent the original decision boundary, the green solid line represents the decision boundary after the oversampling.

http://journal.szu.edu.cn

形成k个簇，再选取每簇的簇心和簇内其余点合成新样本，避免分类边界模糊，使得合成样本仍符合原始数据的分布特征，见图3(b)．其中，红、蓝色椭圆表示聚类簇．合成少数样本S^j_{new}计算公式为

$$S^j_{new} = c_j + R(0,1)(S^j_0 - c_j) \tag{3}$$

其中，j为聚类簇的序号；c_j为第j簇的簇心；S^j_0为以c_j为簇心的原始少数样本；$R(0,1)$表示生成0～1之间的某个随机数．

3 岩性识别

3.1 测井参数敏感性分析及优选

测井数据的变化范围能够表征不同岩性的测井响应特征．选取研究区域内3口井的9条测井曲线，共计4 560条数据绘制交会图版，并利用置信椭圆进行参数敏感性分析，结果如图4．其中，g为自然伽马(这里采用美国石油协会单位(American Petroleum Institute units，API)，即放射性层中部测井仪记录的计数率与非放射性层中部记录的计数速率之差的1/200)；r为地层真电阻；a为声波时差；ρ为地层密度；c为补偿中子孔隙度；v为侵入带电阻；o为冲洗带电阻；p为自然电位；l为井径．测井数据分布表请扫描论文末页右下角二维码查看表S2．

图4中置信椭圆表示90%的点所落到的区域，椭圆越窄，说明2种测井曲线相关性越高；椭圆越分散，说明测井参数对岩性的判别越敏感．为进一步说明测井参数敏感性，本研究定义样本比表示置信椭圆的离散程度，

$$D_i = \frac{N_i}{M_i} \quad (i = 1, 2, \cdots, n) \tag{4}$$

其中，D_i为第i个岩性的置信椭圆离散度；M_i为该置信椭圆内总样本数；N_i为第i个置信椭圆内第i个岩性的样本数．D_i值越大代表椭圆数据离散度越高，即测井参数越敏感，数据区分度越高，当D_i=

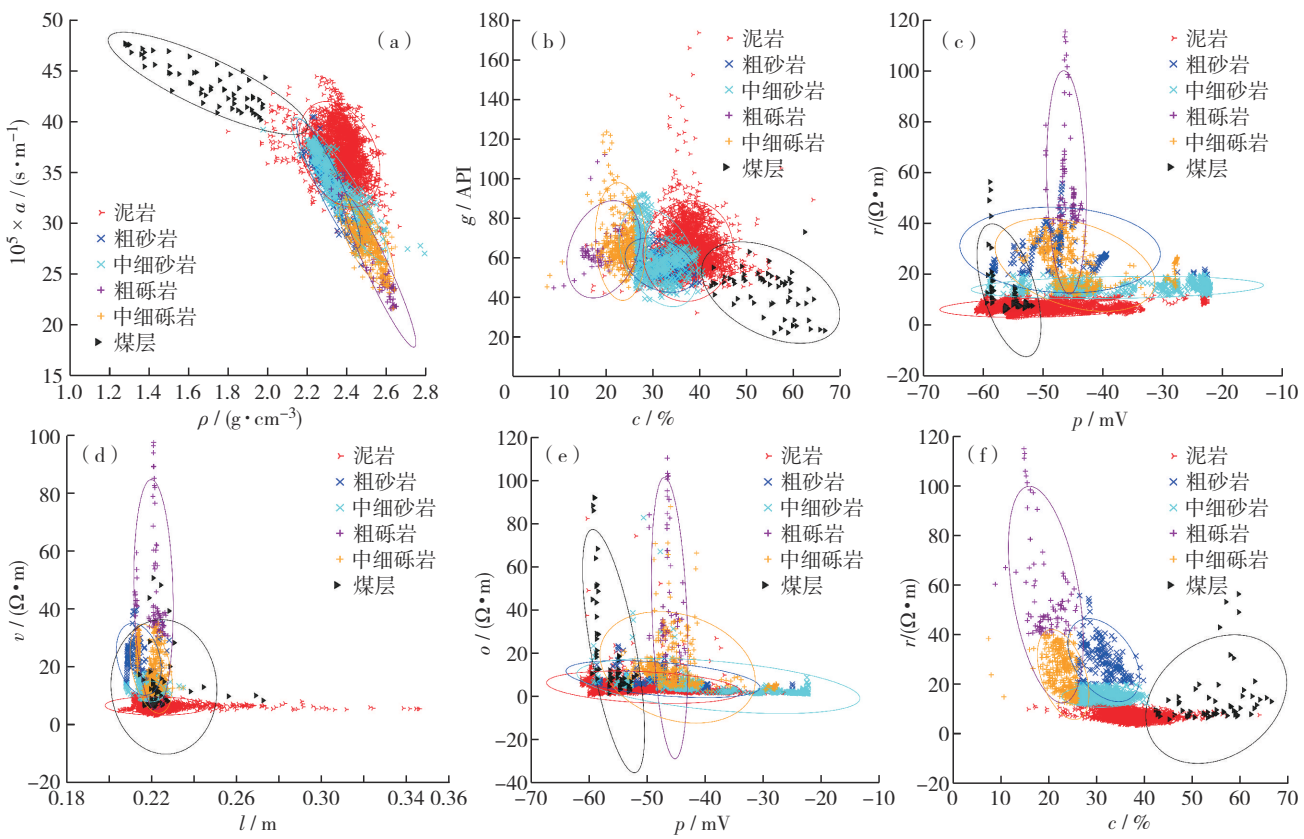


图4 二维测井参数置信椭圆交会(90%置信区间) (a)声波时差-密度交会；(b)自然伽马-中子交会；(c)地层真电阻-自然电位交会；(d)侵入带电阻-井径交会；(e)冲洗带电阻-自然电位交会；(f)地层真电阻-自然伽马交会

Fig. 4 Two-dimensional logging response parameter confidence ellipse intersection graph (90% confidence interval). (a) Cross plot of acoustic and density, (b) cross plot of natural gamma ray and compensated neutron logging, (c) cross plot of true formation resistivity and spontaneous potential, (d) cross plot of invaded zone resistivity and borehole diameter, (e) cross plot of flushed zone formation resistivity and spontaneous potential, and (f) cross plot of true formation resistivity and natural gamma ray.

http://journal.szu.edu.cn

1 时，表示第 i 椭圆内仅有第 i 类岩性，离散度最高，此时测井参数最敏感.

本研究规定当 $D_i > 60\%$ 时，认为该测井参数对该类岩性敏感. 图 5 为置信椭圆离散度热力图. 由图 5 可知，研究区内声波时差、密度、中子、自然伽马和地层真电阻率对各类岩性变化的敏感性强，且区分度较高，井径、自然电位、侵入带电阻率和冲洗带电阻率区分度较差. 井径主要受施工钻头直径控制，采用同一钻头钻进必然导致井径分布较为集中. 自然电位易受到所测岩层厚度的影响，无法表征隔层和夹层，使得数据可分性较差. 侵入带电阻率和冲洗带电阻率则受钻井液侵入影响，区分度较差，无法作为岩性识别依据. 因此，选取自然伽马、地层真电阻率、声波时差、密度和中子参数进一步分析.

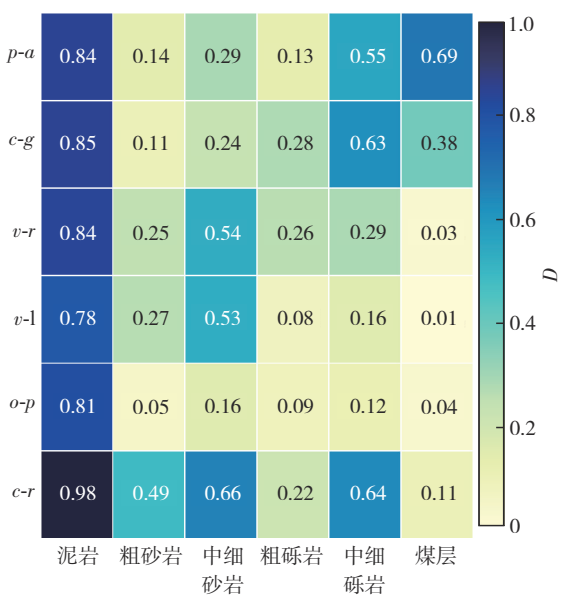


图 5　置信椭圆离散度热力图

Fig. 5　Heatmap of confidence ellipse dispersion.

由图 4 可知，煤层呈现低密度、高中子和高声波时差的分布特点，泥岩则在密度和声波均表现为中高值，砂岩中子和地层真电阻表现为中低值，砾岩则呈现高密度和低中子的分布特点. 结合图 5 可知，由于泥岩参数分布集中，因此，各测井参数均较为敏感，较容易识别. 同时煤层主要通过密度和声波时差进行识别，因此，可直接通过优选后的测井参数识别，两类岩性的识别判据见表 1.

利用常规测井参数交会图能够区分泥岩、煤层，但对砂岩、砾岩的次级识别分效果较差，为精确识别中-细砂岩、粗砂岩、中-细砾岩和粗砾岩，

表 1　煤层和泥岩识别判据

Table 1　Identification criterion of coal seam and mudstone

岩性	a/(s·m^{-1})	ρ/(g·cm^{-3})	r/(Ω·m)	c/%
煤层	40 ~ 50	1. 25 ~ 2. 00	11 ~ 60	45 ~ 70
泥岩	26 ~ 46	2. 00 ~ 2. 50	3 ~ 11	20 ~ 45

需要对测井参数进行主成分提取，增强数据区分度. 基于测井参数敏感性分析，选取声波时差、密度、中子、自然伽马和地层真电阻率提取主成分，主成分方差大小代表原始数据信息量. 本研究规定当主成分方差大于 85% 时认为该主成分可代表原始数据特征，通过主成分分析得到方差大于 85% 的主成分 F_1、F_2 和 F_3 分别为

$$\begin{cases} F_1 = 0.1025g + 0.4346r - 0.5656a + \\ \qquad 0.4222\rho - 0.5499c \\ F_2 = 0.8246g - 0.4061r - 0.0792a - \\ \qquad 0.3832\rho - 0.0454c \\ F_3 = 0.5511g + 0.5270r + 0.0865a - \\ \qquad 0.6384\rho - 0.0598c \end{cases} \tag{5}$$

各成分的系数正负表示主成分与该参数的正负相关关系，其绝对值大小代表相关程度. 由式(5)可知，F_1 与自然伽马、地层真电阻和密度呈正相关，与声波时差、中子呈负相关，并且与自然伽马的相关度相对较小；F_2 只与自然伽马呈正相关，与其他参数皆呈负相关，与中子相关性相对较低；F_3 与自然伽马、地层真电阻和声波时差呈正相关，与密度、中子呈负相关，并且同样与中子的相关度不高.

由图 6 和图 7 交会图可知，主成分分析有效提高了数据可分性，4 种岩性的区分度较高，因此确定砂岩和砾岩主成分细分识别判据如表 2.

基于以上分析，建立常规测井数据和主成分综合识别图版，识别流程图请扫描论文末页右下角二维码查看图 S3.

选取 30% 的数据作为测试集，利用综合识别图版对该井的岩性进行分类，得到岩性识别准确率如表 3.

由表 3 可知，主成分综合图版识别平均准确率较高，达到 80% 以上，对泥岩、煤层、粗砾岩和中细砾岩的识别较为准确，但粗砂岩和中细砂岩识别结果与实际岩性有一定差异，原因为该井段非储层岩性发育较多，而泥岩的自然伽马较高，导致主成分 F_2、F_3 与自然伽马参数的相关性高，而中细砂岩和粗砂岩的自然伽马数据区间差别较小，使得通过

http://journal.szu.edu.cn

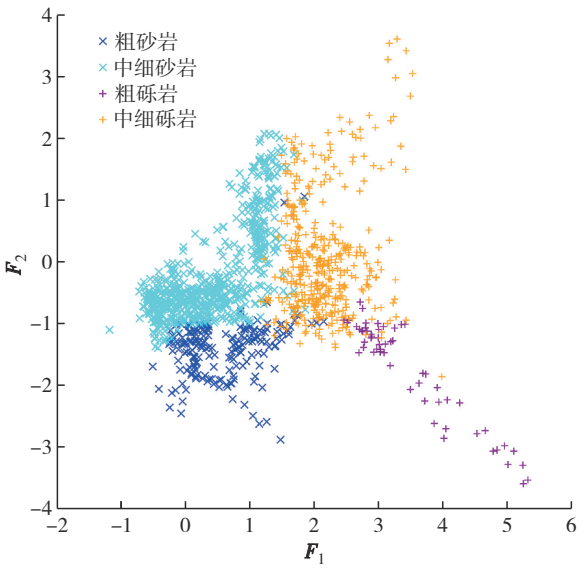


图6 主成分F_1和F_2交会

Fig. 6 The cross plot of principal components of F_1 and F_2.

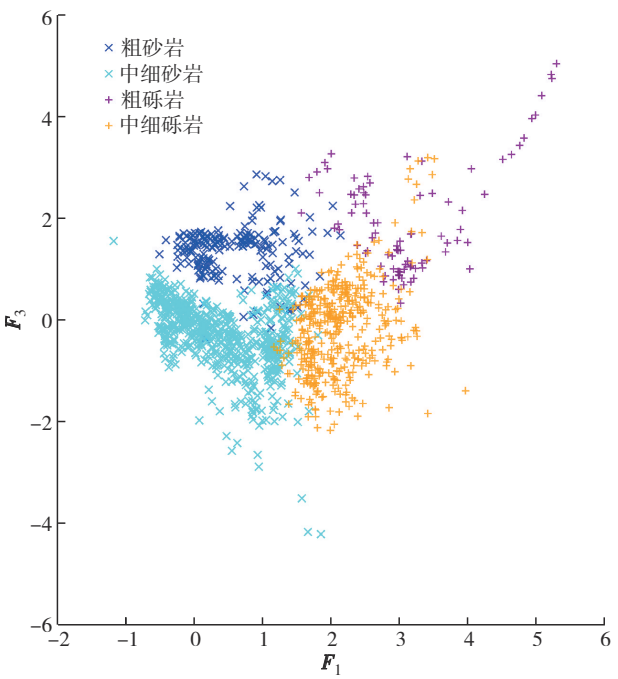


图7 主成分F_1和F_3交会

Fig. 7 The cross plot of principal components F_1 and F_3.

表2 砂岩和砾岩细分识别判据

Table 2 Identification criterion of sandstone and conglomerate

岩性	F_1	F_2	F_3
粗砂岩	−1.0 ~ 1.4	−3.0 ~ −1.1	0.8 ~ 3.0
中细砂岩	−1.0 ~ 1.4	−1.1 ~ 2.5	−4.0 ~ 0.8
粗砾岩	1.4 ~ 3.4	−4.0 ~ −1.2	0.8 ~ 4.5
中细砾岩	2.5 ~ 6.0	−1.2 ~ 3.5	−2.0 ~ 0.8

表3 综合图版识别结果

Table 3 Lithologic identification by cross plot %

岩 性	识别准确率
泥岩	90.50
粗砂岩	65.96
中细砂岩	62.70
粗砾岩	86.67
中细砾岩	94.29
煤层	88.14
平均	81.37

F_2和F_3主成分标准识别的结果误差较大，同时，对于图版中各岩性识别参数阈值设置也是影响模型精度的因素之一.

3.2 基于随机森林的智能岩性识别

智能岩性识别模型以对岩性变化敏感的五条测井曲线为依据，采用随机森林算法对研究区域内测井数据进行岩性划分，使用 Python 语言并基于 sklearn 库实现.

通过 GridSearchCV 对模型各项参数进行遍历评估，选取最优模型参数，模型参数请扫描论文末页右下角二维码查看表 S4.

利用随机森林模型对测试集进行岩性识别，随机森林平均准确为 90.02%(表 4)，相比于主成分综合图版准确率提升了 9.66%，证明了基于随机森林进行岩性识别的可行性.

表4 随机森林识别结果

Table 4 Lithologic identification by random forest %

岩 性	识别准确率
泥岩	95.69
粗砂岩	87.83
中细砂岩	96.19
粗砾岩	84.77
中细砾岩	95.64
煤层	79.97
平均	90.02

由于在多分类问题中会出现真假分类的情况，为进一步分析识别结果，本研究采用宏平均(macro-average，MA)方法评价模型实际识别准确率 A 为

$$A = \frac{1}{n}\sum_{w=1}^{n}\left(\frac{2P_1}{2P_1 + N_0 + P_0}\right)_w \tag{6}$$

http://journal.szu.edu.cn

其中，n 为分类类别总数；$w = 1, 2, \cdots, n$；P_1 为真正例；P_0 为假正例；N_0 为假反例．在测试集中随机森林的宏平均准确率达到 86.13%，而主成分图版的宏平均准确率为 75.10%，模型训练和预测中均未出现过拟合和欠拟合问题．对比综合图版和随机森林的结果可知，随机森林在岩性识别问题上表现确实优于图版法，具有实际应用价值．

在单项岩性分类问题中，由于该井段泥岩层较为发育，煤层和粗砾岩较少，并且在其他岩层中多夹杂泥岩薄层互层，使得粗砾岩和煤层的识别准确率低于其他岩性．因此，本研究采用 KMSMOTE 算法对不均衡样本进行过采样，原数据集中泥岩占比高达 69.17%，因此，将泥岩视为多数类样本，对其余岩性进行均衡采样，均衡前后样本分布图请扫描论文末页右下角二维码查看图 S5．

采用均衡样本训练后，随机森林平均准确率达到 92.94%，宏平均达到 88.86%（表 5）．单项岩性识别准确率均高于不均衡样本训练的随机森林准确率，其中，粗砾岩和煤层识别准确率均达到 90% 以上，说明 KMSMOTE 算法对少数类样本过采样的过程中很好地保证了分类边界，有效提高了不平衡岩性样本的识别准确率．

表 5　KMSMOTE-随机森林识别结果

Table 5　Lithology identification by KMSMOTE and random forest combined　%

岩　性	识别准确率
泥岩	95.69
粗砂岩	90.13
中细砂岩	96.27
粗砾岩	89.64
中细砾岩	95.87
煤层	90.03
平均	92.94
宏平均	88.86

表 6 为随机森林模型中变量重要性排序，地层电阻和中子重要性较高，声波时差和密度对分类的影响次之，而自然伽马的重要性最低．

图 8 为不同岩性的自然伽马箱型图．由图 8 可知不同岩性自然伽马区间跨度大，各岩性数值重叠区间较多，对岩性过渡时的识别结果造成影响．因此，在随机森林模型中通过降低自然伽马的重要性来平衡数据分布的影响，保证分类结果为全局最优解．

表 6　随机森林变量重要性

Table 6　Parameter importance sequence of random forest

重要性排序	重要性分数
地层真电阻	0.455 5
中子	0.211 9
声波时差	0.179 1
密度	0.139 6
自然伽马	0.013 4

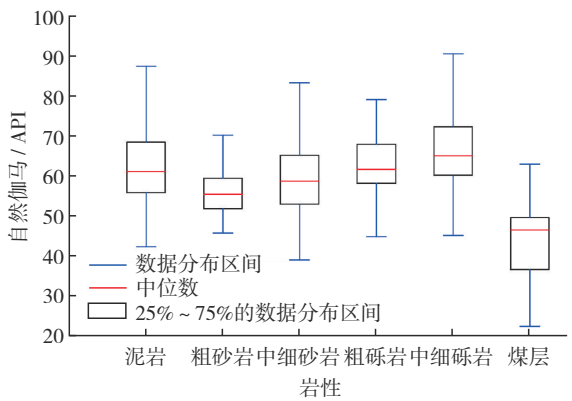


图 8　不同岩性的自然伽马箱型图

Fig. 8　The GR boxplot of different lithologies.

4　应用实例

为进一步验证模型的泛化能力，选取研究区域内一口邻井开展智能岩性识别．将综合图版法和随机森林模型的识别结果对比分析，综合图版平均识别准确率达到 82.91%，随机森林平均准确率达到 91.79%，KMSMOTE-随机森林识别平均准确率达到 95.71%，结果见图 9．

综合图版对砂岩识别出现偏差，原因是主成分分析虽能提升数据可分性，但仍有部分粗和中细砂岩的测井数据较为接近，使得该岩性出现误判．而采用 KMSMOTE-随机森林法对砂岩识别准确率较高，同时粗砾岩和煤层识别率均高于单一随机森林方法，说明该模型对于不均衡岩性样本仍能保持良好的性能．

综上所述，KMSMOTE-随机森林的模型在识别中表现出较强的泛化能力，能够较为准确地识别隔层和夹层岩性．与综合图版识别结果相比，KMSMOTE-随机森林的模型受数据区分度影响较小，与实际岩性吻合度较高，结果见图 10．

http://journal.szu.edu.cn

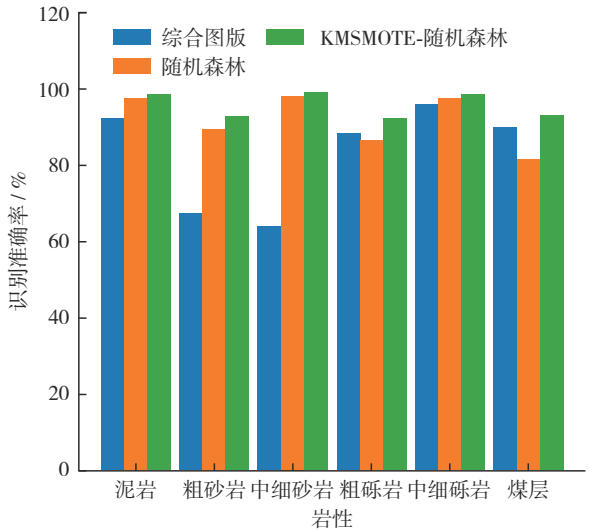


图 9 邻井岩性识别结果

Fig. 9 Lithology identification of an adjacent well.

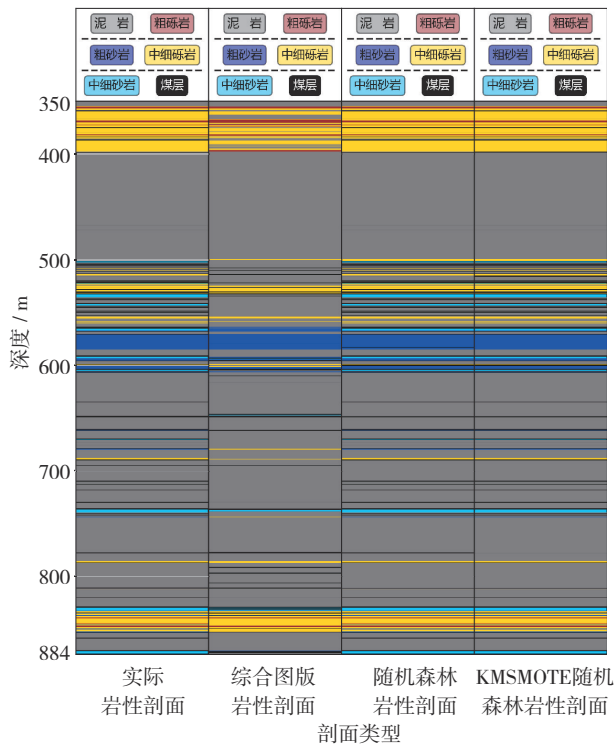


图 10 邻井岩性剖面类型识别结果

Fig. 10 Lithology identification of an adjacent well.

5 结 论

1）主成分分析能够增强数据可分性，结合测井数据敏感性分析优选 5 种测井响应参数，建立克拉玛依七区砂砾岩油藏岩性识别图版，有效识别砂砾岩次级岩性.

2）基于 KMSMOTE 和随机森林结合的智能岩性识别方法，可通过对少数类样本过采样合成新数据集来提升模型训练效果. 经实际数据验证，与单一随机森林模型相比，采用该方法可有效提高不均衡样本中少数类的识别精度，与实际岩性剖面的吻合度较高，实现了岩性的快速准确划分.

3）KMSMOTE 算法有效平衡了样本的不均衡分布，对岩性识别问题有很好的适应性. 未来的改进研究将针对优化聚类簇数选取和提高算法速度两方面开展工作，力求进一步优化模型的识别效果.

基金项目：国家自然科学基金资助项目(42277122)；中国石油大学(北京)科研基金资助项目(2462020YXZZ030)

作者简介：陆吉(1999—)，中国石油大学（北京）油气井工程博士研究生. 研究方向：储层改造评价.
E-mail: luji514@sina.com

引　　文：陆吉，林伯韬，史璨，等. 克拉玛依油田七区砂砾岩油藏智能岩性识别[J]. 深圳大学学报理工版，2023，40(3)：361-369.

参考文献 / References：

[1] REN Quan, ZHANG Hongbing, ZHANG Dailu, et al. A framework of active learning and semi-supervised learning for lithology identification based on improved naive Bayes [J]. Expert Systems with Applications, 2022, 202: 117278.

[2] REN Quan, ZHANG Hongbing, ZHANG Dailu, et al. Lithology identification using principal component analysis and particle swarm optimization fuzzy decision tree [J]. Journal of Petroleum Science and Engineering, 2022, 220: 111233.

[3] PEI Jianya, ZHANG Yunfeng. Prediction of reservoir fracture parameters based on the multi-layer perceptron machine-learning method: a case study of ordovician and cambrian carbonate rocks in Nanpu Sag, Bohai Bay Basin, China [J]. Processes, 2022, 10(11): 2445.

[4] OSAROGIAGBON A U, KHAN F, VENKATESAN R, et al. Review and analysis of supervised machine learning algorithms for hazardous events in drilling operations [J]. Process Safety and Environmental Protection, 2021, 147: 367-384.

[5] 马俊修，石胜男，陈进，等. 基于机器学习的玛湖地区水平井压裂设计优化[J]. 深圳大学学报理工版，2021，38(6)：621-627.
MA Junxiu, SHI Shengnan, CHEN Jin, et al. Optimization of fracture design for horizontal wells in Mahu region based on machine learning [J]. Journal of Shenzhen University Science and Engineering, 2021, 38(6):621-627. (in Chinese)

http://journal.szu.edu.cn

[6] 姜振海. 基于分频智能反演的储层预测研究——以大庆黑鱼泡南部探区为例[J]. 地质科学，2021，56(4)：1052-1061.
JIANG Zhenhai. Research on reservoir prediction based on frequency-divided intelligent inversion: a case study from southern exploration area of Heiyupao in Daqing [J]. Chinese Journal of Geology, 2021, 56(4): 1052-1061. (in Chinese)

[7] KONATÉ A A, MA Houalin, PAN Heping, et al. Analysis of situ elemental concentration log data for lithology and mineralogy exploration: a case study [J]. Results in Geophysical Sciences, 2021, 8: 100030.

[8] REN Quan, ZHANG Hongbing, ZHANG Dailu, et al. A novel hybrid method of lithology identification based on *k*-means++ algorithm and fuzzy decision tree [J]. Journal of Petroleum Science and Engineering, 2022, 208: 109681.

[9] MEDJA USSALU J L M, BASSREI A. Well log analysis for lithology and fluid contacts in Rovuma Basin-Mozambique: application of cluster and discriminant analyses [J]. Journal of African Earth Sciences, 2022, 185: 104419.

[10] DAN Mou, WANG Zhuwen, TAN Xili, et al. A variational inequality approach with SVM optimization algorithm for identifying mineral lithology [J]. Journal of Applied Geophysics, 2022, 204: 104747.

[11] ZHANG He. Research on identification model of element logging shale formation based on IPSO-SVM [J]. Petroleum, 2022, 8(2): 185-191.

[12] SHAYEGANPOUR S, TANGESTANI M H, GORSEVSKI P V. Machine learning and multi-sensor data fusion for mapping lithology: a case study of Kowli-kosh area, SW Iran [J]. Advances in Space Research, 2021, 68(10): 3992-4015.

[13] 武中原，张欣，张春雷，等. 基于LSTM循环神经网络的岩性识别方法[J]. 岩性油气藏，2021，33(3)：120-128.
WU Zhongyuan, ZHANG Xin, ZHANG Chunlei, et al. Lithology identification based on LSTM recurrent neural network [J]. Lithologic Reservoirs, 2021, 33(3): 120-128. (in Chinese)

[14] LIU Zaobao, LI Long, FANG Xingli, et al. Hard-rock tunnel lithology prediction with TBM construction big data using a global-attention-mechanism-based LSTM network [J]. Automation in Construction, 2021, 125: 103647.

[15] 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[J]. 中国石油大学学报自然科学版，2022，46(3)：62-71.
PAN Shaowei, WANG Chaoyang, ZHANG Yun, et al. Lithology identification based on LSTM neural networks completing log and hybrid optimized XGBoost [J]. Journal of China University of Petroleum Edition of Natural Science, 2022, 46(3): 62-71. (in Chinese)

[16] TROTT M L, LEYBOURNE M, HALL L, et al. Random forest rock type classification with integration of geochemical and photographic data [J]. Applied Computing and Geosciences, 2022, 15: 100090.

[17] AO Yile, ZHU Liping, GUO Shuang, et al. Probabilistic logging lithology characterization with random forest probability estimation [J]. Computers and Geosciences, 2020, 144: 104556.

[18] 王振宇，林伯韬，于会永，等. 克拉玛依油田七区八道湾组砂砾岩油藏地应力特征[J]. 新疆石油地质，2020，41(3)：314-320.
WANG Zhenyu, LIN Botao, YU Huiyong, et al. Characteristics of *in-situ* stress in sandy conglomerate reservoir of badaowan formation in district No. 7, Karamay Oilfield [J]. Xinjiang Petroleum Geology, 2020, 41(3): 314-320. (in Chinese)

[19] SAIDI F, KHETARI S, YAHIA I S, et al. The use of principal component analysis (PCA) and partial least square (PLS) for designing new hard inverse perovskites materials [J]. Computational Condensed Matter, 2022, 31: e00667.

[20] 王凯. 基于改进随机森林算法的P2P贷前信用风险评估方法研究[D]. 南京：南京邮电大学，2020.
WANG Kai. Research on P2P pre-loan credit risk assessment method based on improved random forest algorithm [D]. Nanjing: Nanjing University of Posts and Telecommunications, 2020. (in Chinese)

[21] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique [J]. Journal of artificial intelligence research, 2002, 16: 321-357.

[22] DOUZAS G, BACAO F, LAST F. Improving imbalanced learning through a heuristic oversampling method based on *k*-means and SMOTE [J]. Information Sciences, 2018, 465: 1-20.

【中文责编：晨兮；英文责编：天澜】

补充材料

http://journal.szu.edu.cn