Advances in Geosciences 地球科学前沿, 2020, 10(11), 1118-1136
Published Online November 2020 in Hans. http://www.hanspub.org/journal/ag
https://doi.org/10.12677/ag.2020.1011110

Hans汉斯

复杂火山岩油藏岩性智能识别及预测

孔垂显[1*]，王志章[2#]，冯赫青[3]，魏周城[1]，刑雅文[3]，王伟方[2]，陈文浩[2]，杨　笑[1]

[1]中国石油新疆油田分公司勘探开发研究院，新疆 克拉玛依
[2]中国石油大学(北京)油气资源与探测国家重点实验室，北京
[3]中国石油华北油田公司，河北 沧州
Email: #978301193@qq.com

收稿日期：2020年11月9日；录用日期：2020年11月23日；发布日期：2020年11月30日

摘　要

目前火山岩油气藏正在引起广泛关注，其复杂程度超过其他油藏类型。针对二连盆地、准噶尔盆地复杂火山岩油气藏，其岩性具有复杂多变的特点，常规方法难以准确识别的问题，本文提出利用机器学习的方法对研究区岩性进行智能识别，获得了良好效果。研究中，在分析研究区火山岩储层地质特点基础上，根据取芯描述、薄片分析、成像测井等信息，分析不同岩性的测井响应特征。并根据测井信息，构造M，N两个对火山岩岩性极为敏感的参数，确定了识别岩性的8个敏感特征参数为：GR，DT，RHOB，CNL，RT，RI，M，N。根据测井特征参数和岩性标签，利用机器学习中的决策树、随机森林、梯度提升树、贝叶斯四种不同方法，建立了四种岩性识别预测模型。对不同模型进行了对比评价研究，优选出最优的随机森林岩性识别模型，岩性识别的准确率达到0.9以上，为火山岩油气藏评价奠定了基础。

关键词

复杂火山岩，岩性特征，机器学习，智能识别

Intelligent Recognition and Prediction of Lithology of Complex Volcanic Reservoir

Chuixian Kong[1*], Zhizhang Wang[2#], Heqing Feng[3], Zhoucheng Wei[1], Yawen Xing[3], Weifang Wang[2], Wenhao Chen[2], Xiao Yang[1]

[1]PetroChina Xinjiang Oilfield Branch Exploration and Development Research Institute, Karamay Xinjiang
[2]China University of Petroleum (Beijing) State Key Laboratory of Oil and Gas Resources and Exploration, Beijing
[3]PetroChina North China Oilfield Company, Cangzhou Hebei
Email: #978301193@qq.com

Received: Nov. 9th, 2020; accepted: Nov. 23rd, 2020; published: Nov. 30th, 2020

*第一作者。
#通讯作者。

文章引用: 孔垂显, 王志章, 冯赫青, 魏周城, 刑雅文, 王伟方, 陈文浩, 杨笑. 复杂火山岩油藏岩性智能识别及预测[J]. 地球科学前沿, 2020, 10(11): 1118-1136. DOI: 10.12677/ag.2020.1011110

Abstract

At present, volcanic rock oil and gas reservoirs are attracting widespread attention, and their complexity exceeds other reservoir types. Aiming at the complex volcanic reservoirs in Erlian Basin and Junggar Basin, their lithology is complex and changeable, and conventional methods are difficult to accurately identify the problem. This paper proposes to use machine learning methods to intelligently identify the lithology in the study area, and obtain good results. In the study, based on the analysis of the geological characteristics of the volcanic reservoir in the study area, the logging response characteristics of different lithologies were analyzed based on information such as coring description, thin section analysis, and imaging logging. And according to the logging information, the two parameters of structure M and N that are extremely sensitive to volcanic rock lithology, eight sensitive characteristic parameters for identifying lithology are determined: GR, DT, RHOB, CNL, RT, RI, M, N. According to logging feature parameters and lithology tags, four different methods of decision tree, random forest, gradient boosting tree, and Bayesian in machine learning are used to establish four lithology recognition and prediction models. Different models were compared and evaluated, and the best random forest lithology recognition model was selected. The accuracy of lithology recognition was above 0.9, which laid the foundation for the evaluation of volcanic reservoirs.

Keywords

Complex Volcanic Rock, Lithological Characteristics, Machine Learning, Intelligent Recognition

1. 引言

1957 年中国火山岩油气藏在准噶尔盆地第一次被成功发现，到目前为止，该油气藏已勘探 50 年以上，在准噶尔盆地等 11 个含油气盆地中找到了火山岩油气藏存在[1]。能够准确认识火山岩岩性是研究火山岩油藏的重要基础，但火山岩岩性复杂，矿物成分复杂多变，测井响应特征不明显，岩性识别难度极大。

前人在这方面做了很多的研究。Sanyal 等人(1980)利用声波时差，中子测井曲线的交会图识别岩性[2]。Benoit 等(1980)总结了声波时差测井、中子测井等常规测井在花岗岩，玄武岩等上的响应特征[3]。范宜仁等(1998)在新疆准噶尔盆地，根据交会图技术有效识别火山岩岩性[4] [5]。张莹(2007)根据成像测井资料识别火山岩。谭伏霖，王志章等人(2010)鉴于取芯样本有限，采取样本扩充法识别火山岩岩性[6] [7] [8] [9]。由于火山岩岩石类型多，仅根据单一的测井曲线很难将岩性准确识别出来，近几年许多人利用综合的信息来识别岩性，主要利用地质结合测井和数学算法方法来识别岩性。田艳等(2010)根据逐步分析法和 fisher 判别的方法进行岩性识别[10]。程国建等(2010)为了提高测井在岩性识别的准确率，将粒子群优化算法(PSO)与最小二乘支持向量机相结合对实际测井资料进行岩性识别。鞠武等(2012)利用有序聚类分析方法识别岩性。李建国等(2015)利用深度神经网络结合测井资料进行岩性的识别。牟丹等(2015)基于最小二乘支持向量机识别岩性[11] [12] [13] [14] [15]。

本文综合利用取芯，薄片，成像资料解释的岩性标签标定测井资料，形成测井和岩性标签的样本库。将样本库分为训练集(占 70%)和测试集(占 30%)两部分。结合机器学习中的决策树，随机森林，梯度提升

树和贝叶斯方法，利用训练集建立 4 种火山岩岩性识别模型，利用测试集评价这 4 种模型的稳定性，选择最优模型。结果表明，随机森林和梯度提升树方法模型最优，可作为该研究区利用常规测井曲线识别火山岩岩性的有效方法。

2. 原理方法

2.1. 决策树

决策树是一种描述对实例进行分类的树形结构，决策树由结点和有向边组成。决策树是属性与值之间的一种映射关系。决策树算法过程包括特征的选择、树的生成和树剪枝的过程，根据特征选择的方法不一样，形成了不同的算法，主要包括 ID3 算法，C4.5 算法和 CATR 算法。

ID3 是决策树算法的一种，是在原始决策树算法的基础上实现的。采用的分治的思想，其主要特点是在结点上选择特征时采用的是信息增益的方法。信息熵是衡量随机变量出现的期望值，熵值越大，代表信息的不确定性越大。一般选择熵值大的特征属性作为结点的划分[16]。计算过程是：在开始计算输入的所有属性，根据输入的所有特征属性，根据公式计算这些特征属性的信息增益，选择最大的作为分枝结点。在每个分支结点都进行此运算，最后将每个类别分到不同的叶子节点中，形成一棵决策树。当有未知样本输入时，每经过一个结点判断路径走向，最终预测类别。C4.5 算法它是根据 ID3 算法改进，继承优势补齐缺点。其主要不同之处在于利用信息增益率来进行特征属性的选择。对这一点的改进主要是避免了进行特征属性选择时会偏向多值的特征属性的缺点。另一个优点是加入了树的剪枝功能，从而防止过拟合。此外还可以处理缺省值。CATR 算法与前面两种算法不同的是，利用基尼指数值来分枝不同特征属性，并且该方法主要采用的是二分，即每个枝点都只分成两个结点，最终形成一颗二叉树。

决策树方法最重要的一个环节是确定特征属性划分方案，这个划分方案的选择主要是依据信息论的理论。假设训练数据集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，其中 $x_i=\left(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(n)}\right)$ 为输入实例，n 为特征个数，$y_i\in\{1,2,\cdots,K\}$ 为类标记，$i=1,2,\cdots,N$ 为样本容量，A 表示某一特征属性。设 X 是一个取有限个值的离散随机变量，其概率分布为

$$P(X=x_i)=p_i,\ i=1,2,\cdots,n \tag{1}$$

其中信息量(Intormation)和熵(Entropy)的定义为

$$\text{Information}=-\log_2 p_i \tag{2}$$

$$\text{Entropy}=-\sum p_i\log_2 p_i \tag{3}$$

信息增益：$G(D,A)=\text{Entropy}(D)-\sum\frac{D_A}{D}\text{Entropy}(D_A)$ (4)

信息增益率：$G_R(D,A)=\frac{G(D,A)}{\text{Entropy}(D)}$ (5)

基尼指数：$\text{Gini}(D)=1-\sum p_i^2$ (6)

假设用特征 A 将数据集 D 分为 D_1，D_2，则划分的基尼指数为：

$$\text{Gini}(D)=\frac{D_1}{D}\times\text{Gini}(D_1)+\frac{D_2}{D}\times\text{Gini}(D_2) \tag{7}$$

对于决策树还有另一个种类，是回归决策树。在机器学习中每一类方法即可以用来分类，判别某一类的类别，也可以用于回归计算，计算某一个值，主要解决回归问题。在回归决策树中与分类决策树最大不同点是判断结点分支的依据不是选用信息熵的理论，而是用最小化均方差来判断选择划分特征属性。

2.2. 随机森林方法

随机森林属于组合分类器之一，由许多决策树叠加组成。分配给每棵树的样本是从数据集中随机抽取，抽完后放回，抽取数据是有放回的抽样[17]。每抽取一次数据，就建立一颗决策树模型，重复这样的操作，最后所有的决策树就形成了一片森林，即随机森林(图 1)。

该方法与决策树方法有很多相同的地方，都是属于树形结构，且每个结点的特征属性划分方法基本一致，都采用 ID3，C4.5，和 CATR 等算法。其主要不同点是在决策树的基础上，引入了两个不同的随机条件，第一次随机是从数据集中随机的抽取训练数据集(Random subset)，该抽样方法是有放回的抽样，每次抽取一次则形成一颗决策树(图 1)。在决策树里，随机森林该方法采用了第二次随机条件，即对抽取的训练集中的数据的特征属性集合 n 中采取随机选取 $S(S\leq n)$个特征属性集合，根据这 S 个特征进行结点划分和决策树的生成。该方式使得生成的决策树具有随机性。正因为引入了这两个随机的特征，使得该方法相比决策树而言具有更好的效果[18]。该方法优点主要：第一，有对训练数据不容易产生过拟合的现象。第二，在不同属性中可以判断属性的重要程度，根据选择需要可以选折几个重要的特征属性参数。第三，该方法对噪声、异常值等数据具有不敏感的性质。

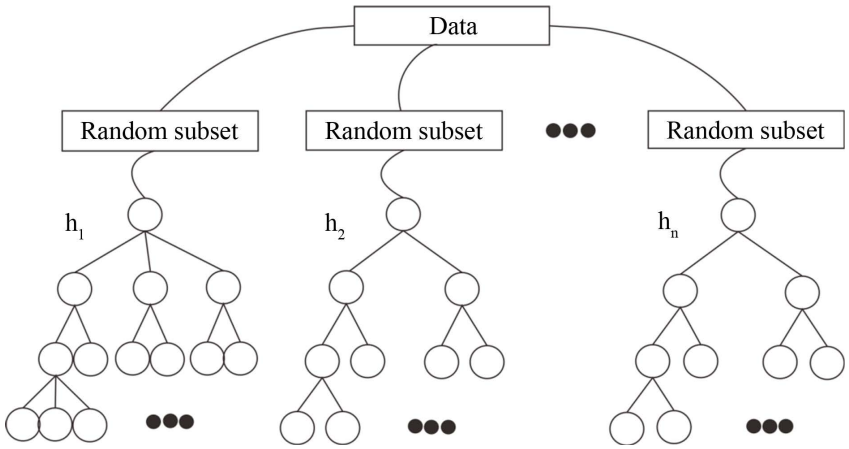


Figure 1. Schematic diagram of random forest
图 1. 随机森林示意图

2.3. 梯度提升树(GBDT)方法

梯度提升树是一种迭代的回归决策树算法，由多颗决策树组成，将所有树的结论累加起来就是最终的结论[19]。该方法是一个应用很广泛的算法，主要是用于分类和回归两方面，其次是特征的选择和创造新特征。该方法学习的过程中，因为是属于加法模型，所以无论是分类还是回归，都应用的是回归模型的决策树。在各结点分支时，不再是利用信息熵理论判断，而是选择最小化均方差来判断结点分支。均方差最小的作为分支判断依据。

该方法主要有两个不同特点。第一种是提升树，即残差方式生成的树。在所给的数据中学习时，最开始是先生成一颗回归决策树。然后根据这颗树的结果与真实值做差值，得到了残差，然后根据每个数据的残差的值再生成一颗决策树。到第 m 步的决策树模型表达式为 $f_m(x)$(公式 8)。其中参数 $\varnothing_m$ 主要依据公式(1.9)，计算残差得到。以此类推，最后得到一个误差较小的树，利用前向分布算法最终把所有的树结合起来即是该方法的分类或回归的决策函数模型(公式 10)。

$$f_m(x)=f_{m-1}(x)+T(x;\varnothing_m) \tag{8}$$

$T(x;\varnothing_m)$是第 m 次的决策树；$f_{m-1}(x)$为前$m-1$步生成的模型；$f_m(x)$是当前生成的模型，$\varnothing_m$为决策树参数。

$$\varnothing_m = \arg\min \sum_{i=1}^{N} L\left(y_i, f_{m-1}(x) + T(x;\varnothing_m)\right) \tag{9}$$

$$f_M(x) = \sum_{m=1}^{M} T(x;\varnothing_m) \tag{10}$$

第二种是梯度提升树，也是本文所应用的树。该方法也是每一颗树学习前 N-1 棵树的残差，直到最后得到分类或回归误差最小的树。与第一种不同的是，该方法利用的是梯度下降的方向求解误差值，这样求解的是局部最优解。而做残差求解是求的全局最优，但仍然选择梯度下降的方向求解，主要是因为该方法灵活，可以方便求解分类和回归问题，并且只要能求导的损失函数都可以使用，可以防止过拟合。对于平方差损失，其二者方法求解方法相同。

梯度提升树的算法过程为：

输入训练集 $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$，损失函数 $L(y, F(x))$，迭代的次数为 M。

1) 使用常量值初始化模型 $F_0(x) = \arg\min \sum_{i=1}^{N} L(y_i, \gamma)$ 其中 γ 是常数。

2) 迭代训练模型

For m = 1:M

计算伪残差：$r_{im} = -\left[\frac{\left(L(y_i, F(x_i))\right)'}{F(x_i)}\right]_{F(x)=F_{m-1}(x)}$ (11)

其中 $i = 1, 2, \cdots, n$

将残差 r_{im} 拟合到回归树 $h_m(x)$中，即 $h_m(x)$的训练集为$\{(x_i, r_{im})\}$

计算乘数 γ

$$\gamma_m = \arg\min \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right) \tag{12}$$

更新模型

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{13}$$

优点：1) 能够允许不同特征组合拥有不同的判别式；2) 对非线性数据有着很好的处理效果；3) 能解决数据缺失问题；4) 可以自动进行特征选择[20]。

2.4. 朴素贝叶斯方法

在机器学习中主要有两种模型，一种是判别模型，另一种是生成模型。前三种是判别模型，贝叶斯是属于生成模型，从数据中学习到联合概率分布，再求出条件概率分布。朴素贝叶斯法分类主要是根据已知的数据学习计算出先验概率，在根据条件独立性假设计算条件概率，最后计算后验概率，对未知数据集进行预测。

设输入的 χ 属于 Rn 为 n 维向量的集合，输出的类标记集合为 $l = \{c_1, c_2, \cdots, c_k\}$，输入的向量 x 属于 χ，输出的不同类 y 属于 l，X 为空间 χ 上的随机向量，Y 为空间 l 上的类别标签。训练数据集为：

$$D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$$

在训练数据集中计算先验概率和条件概率。最后得到后验概率。

先验概率分布：

$$P\left(Y=c_k\right),\ k=1,2,\cdots,K \tag{14}$$

条件概率分布：

$$P\left(X=x\mid Y=c_k\right)=P\left(X^1=x^1,\cdots,X^n=x^n\mid Y=c_k\right),\ k=1,2,\cdots,K \tag{15}$$

条件独立性假设：

$$P\left(X=x\mid Y=c_k\right)=P\left(X^1=x^1,\cdots,X^n=x^n\mid Y=c_k\right)=\prod_{j=1}^{n}P\left(X^j=x^j\mid Y=c_k\right) \tag{16}$$

根据已经计算的先验概率和条件概率，即可根据贝叶斯理论公式构造出预测模型。对于未知数据 x，加载到前面模型，计算出后验概率 $P\left(Y=c_k\mid X=x\right)$，选择后验概率最大的为 x 的输出类。后验概率计算公式为：

$$P\left(Y=c_k\mid X=x\right)=\frac{P\left(X=x\mid Y=c_k\right)P\left(Y=c_k\right)}{\sum_k^K P\left(X=x\mid Y=c_k\right)P\left(Y=c_k\right)} \tag{17}$$

则朴素贝叶斯法分类的基本公式：

$$y=f\left(x\right)=\arg\max\frac{P\left(Y=c_k\right)\prod_j^n P\left(X^{(j)}=x^{(j)}\mid Y=c_k\right)}{\sum_k^K P\left(Y=c_k\right)\prod_j^n P\left(X^{(j)}=x^{(j)}\mid Y=c_k\right)} \tag{18}$$

上式中分母对所有 c_k 都是相同的，所以得到分类器为，

$$y=f\left(x\right)=\arg\max P\left(Y=c_k\right)P\left(X^{(j)}=x^{(j)}\mid Y=c_k\right) \tag{19}$$

贝叶斯算法的缺点是建立在样本属性独立性假设的基础上，所以如果样本属性有关联时其效果不是很好[21]。

3. 研究思路及流程

以关键井，作为重要标签井，利用常规测井和非常规测井资料、取芯资料、通过交会图来判断不同测井开展岩性特征分析，选取对岩性敏感的常规测井系列，如选取“GR”、“SP”、“CAL”、“AC”、“RT”等，以取心井岩心描述组为标签数据。

建立岩性知识库组成训练集，利用机器学习的决策树、随机森林、梯度提升树、贝叶斯等算法，根据优选的测井参数和机器学习建立岩性分类模型，对所建立的各种算法模型进行评价，优选出较好的算法模型，再对未知井段的岩性利用该模型进行预测。见图 2、图 3 流程如：

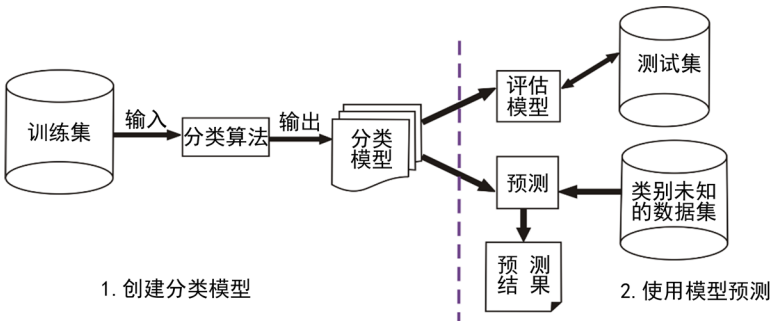


Figure 2. Flow chart of machine learning
图 2. 机器学习流程图

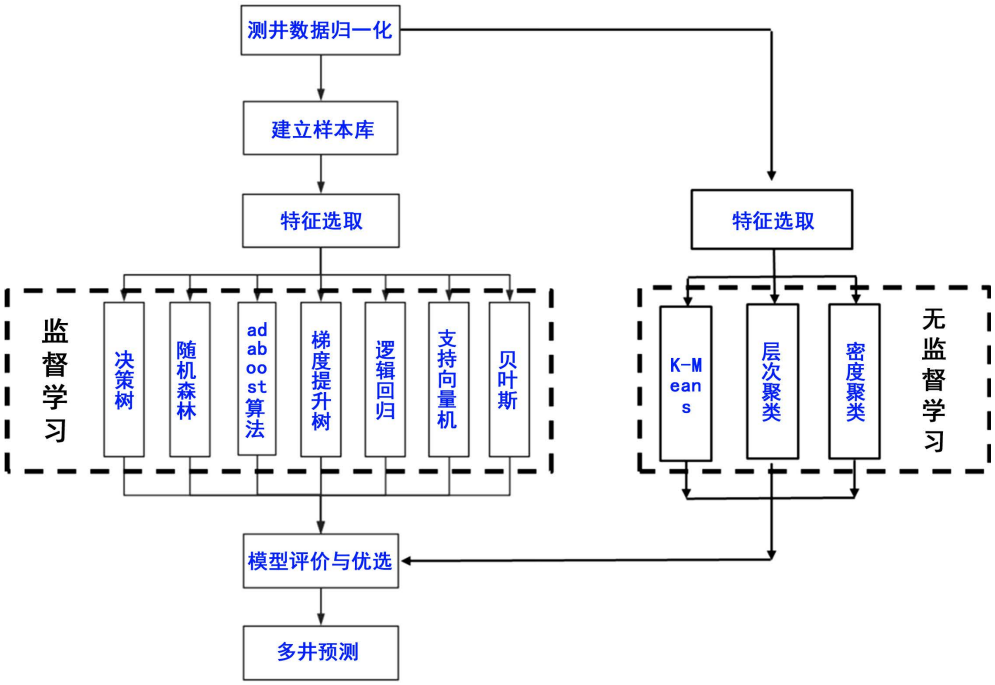


Figure 3. Flow chart of intelligent identification and prediction of complex volcanic rocks
图 3. 复杂火山岩岩性智能识别及预测流程图

4. 应用实例

金龙 2 井区块位于准噶尔盆地西北缘中拐凸起东斜坡带。目的层位于二叠系佳木河组的火山岩地层，岩性复杂，储集空间类型多。根据研究区 8 口取芯井数据，薄片数据，实验分析化验数据可知，研究区火山岩岩性主要分为熔岩类和火山碎屑岩类，熔岩类包括玄武岩，安山岩，英安岩和流纹岩 4 种。火山碎屑岩类包括熔结角砾岩，火山角砾岩和凝灰岩 3 种。

4.1. 火山岩岩性常规测井响应特征

测井的响应主要是应为地下岩石的组分、流体类型、裂缝等综合的影响。其主要的影响程度最高的是岩石的类型。不同的岩石类型其岩石的化学成分、矿物成分等有很大不同，导致测井有不同的变化特征。对于电阻率测井，火山岩岩石越致密电阻率越高，熔岩类的电阻率高于火山碎屑岩类[22] [23] [24]。熔岩中英安岩裂缝发育较少，致密，电阻率很高，安山岩，流纹岩和玄武岩等由于裂缝和气孔的因素影响导致电阻率会偏低。

自然伽马测井的响应特征与岩石的放射性有关，而岩石的放射性与岩石的矿物成分有关。在火山岩中，岩石的放射性与岩石属于基性岩，酸性岩等有关。从基性到酸性逐渐增加[25] [26]。在该研究区目的层内，玄武岩自然伽马小于 26 API，安山岩在 45 API 左右，英安岩和流纹岩高伽马，大于 65 API，在相同类岩石中，放射性的大小还与岩石的结构有关。从熔岩到火山碎屑岩，放射性增加。凝灰岩自然伽马在 60 API 左右，熔结角砾岩低伽马，25 API 左右，火山角砾岩中伽马，50 API 左右。

中子测井主要反映的是地层的孔隙情况。该地区目的层中，凝灰岩本身相对火山岩而言，不是很致密，存在孔隙，测的中子值也比较大，在 25%左右。火山角砾岩和熔结角砾岩也存在较大的原生孔隙，测的中子值约 20%左右，是该区主要储层。安山岩由于裂缝比较发育中子值约 17%左右。玄武岩由于发育杏仁构造，中子测井值较高，约 25%。流纹岩和英安岩发育致密，中子测井值较低，约 10%左右[27]。

密度测井一般而言，由基性到酸性的火山岩，该测井值会逐渐变小[28]。流纹岩为酸性的喷出岩，密度较低。安山岩为中性的喷出岩，密度相对流纹岩较高，为 2.54 g/cm^3 左右。玄武岩为基性的喷出岩，密度最高为 2.7 g/cm^3 左右。在同类岩石中，火山碎屑岩的密度则低于熔岩。

声波时差测井在火山岩石中，火山碎屑岩的致密程度低于熔岩类，其声波时差会高于熔岩[29] [30]。火山角砾岩和熔结角砾岩声波时差在 65 μs/ft 左右。凝灰岩的致密程度最低，声波时差最高，约 80 μs/ft。英安岩最致密，声波时差最低约 57 μs/ft 左右。玄武岩，流纹岩和安山岩声波时差约 60 μs/ft。

成像测井反映周围岩石电阻率的变化，由一种渐变的色板(黑－棕－黄－白)对电阻率由低到高进行刻度，常用来识别结构和构造。一般分为静态平衡图像和动态加强图。静态图采用统一配色，展现井段电阻率的改变。动态图是采用分段配色，突出局部相对微电阻率的变化(颜色由深到浅，电阻率由小到大)。所以利用静态图可以区分沉积岩和火成岩[31] [32]。利用 FMI 动态图可以识别火山碎屑岩(角砾岩和凝灰岩)和具有流纹构造的流纹岩。利用动态图和常规测井曲线可以识别英安岩，安山岩和不具流纹构造的流纹岩。该方法特点是能通过辨别岩石结构和构造来区分岩性。英安岩 FMI 图像呈块状构造，火山角砾岩呈角砾结构，流纹岩呈流纹构造。凝灰岩呈凝灰结构。见图 4。

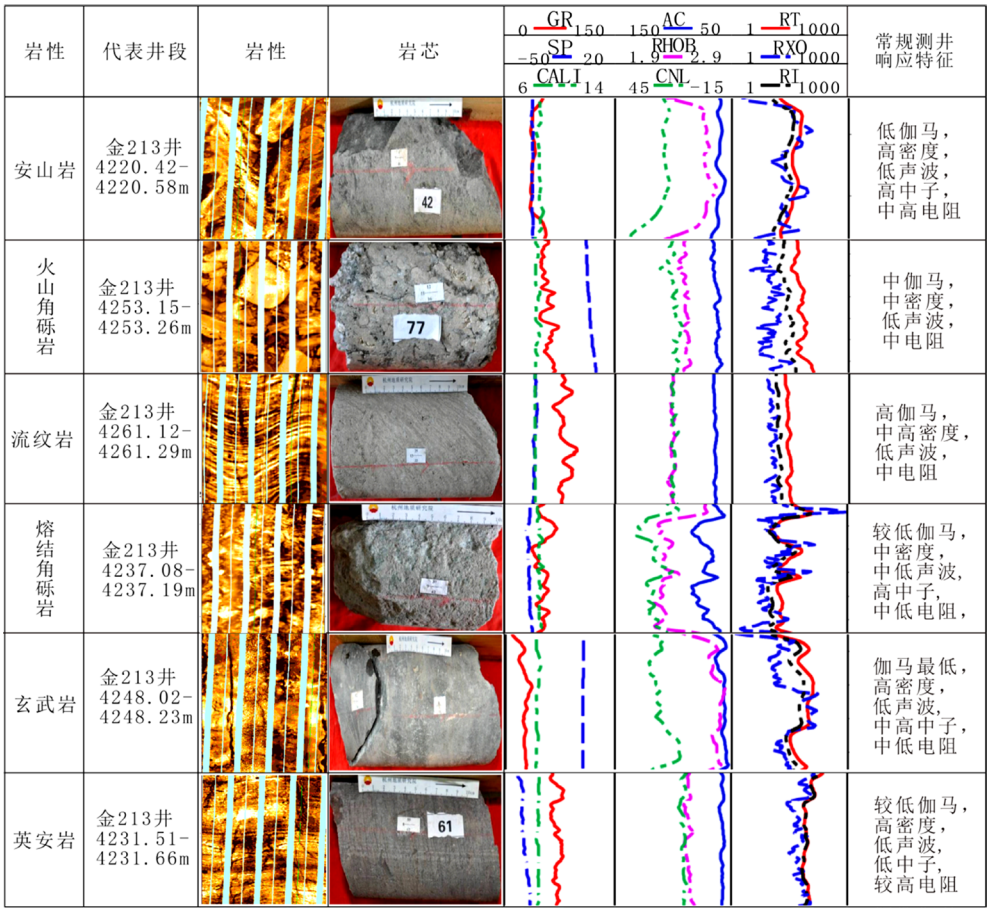


Figure 4. Log response characteristics of volcanic rock
图 4. 火山岩测井响应特征

4.2. 交会图分析

前人在识别火山岩的识别中，主要利用交会分析法，在交会图上描述不同岩性的特征。该方法优点是

简单直观。缺点是利用特征太过单一，不具有代表性，错误率较高。本文采用机器学习的方法的优点是可以融合前人所有研究的方法特点。本文采用交会图分析不同岩性特征，提取敏感特征[33] [34] [35] [36]。

特征选取的个数对岩性预测的结果影响非常大。假如只有一个特征，即在一维空间中，如图 5，在对角上面的 GR，DT，CNL，RHOB，RT，RI，即岩性在一维空间中的响应分布，很难把各类岩性区分开。在二维空间上，即选择两个特征的交会图分析，相比一维特征，岩性的分类效果较好。

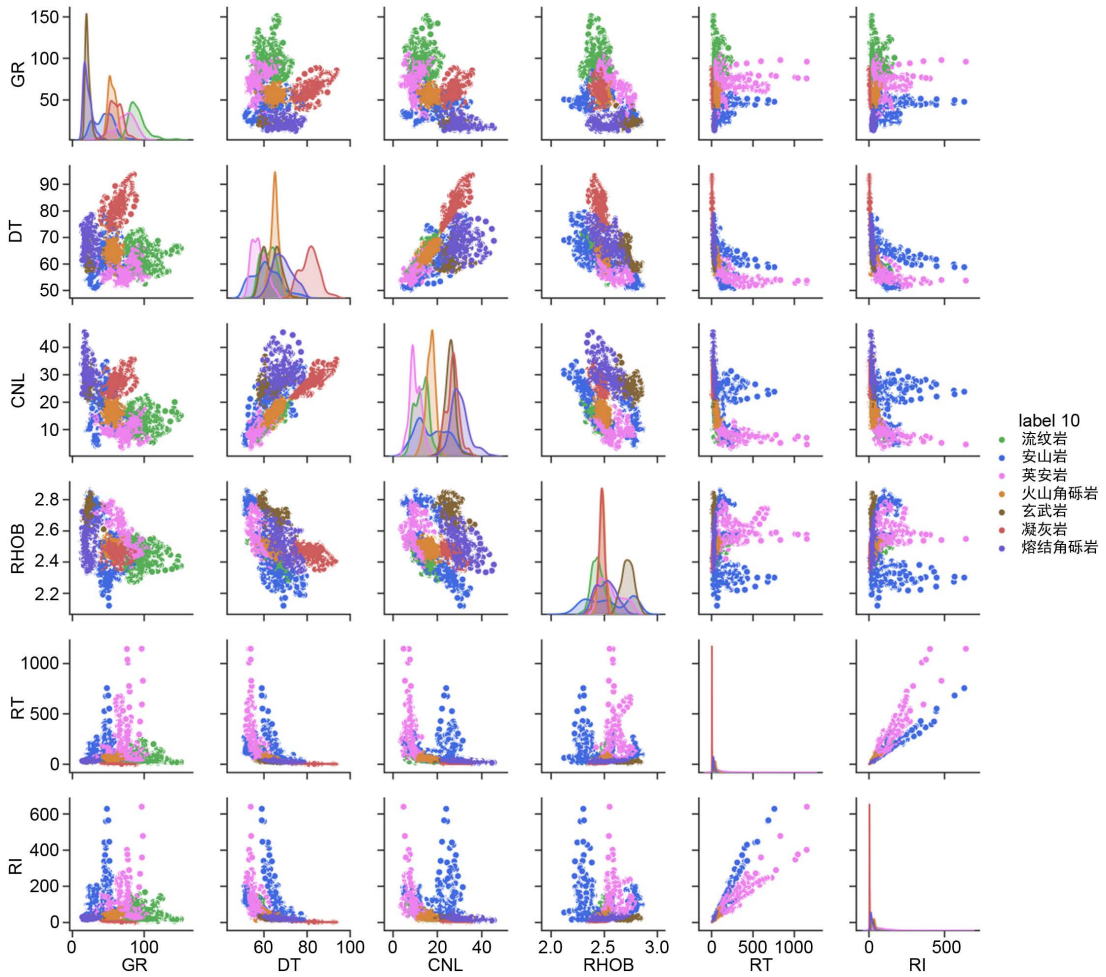


Figure 5. Logging intersection
图 5. 测井交会图

4.3. M-N 交会图

岩性识别前人除了利用常规的测井曲线交会图，还有其中确定岩性模型M-N交会图，主要应用中子、密度和声波三种孔隙度测井。该交会图主要目的是消除孔隙度的影响因素，而突出岩性的作用(如图 6)。M，N 定义为：

$$\mathrm{M}=\frac{\Delta t_f-\Delta t}{\rho_b-\rho_f}\times 0.01 \tag{20}$$

$$\mathrm{N}=\frac{\phi_{Nf}-\phi_N}{\rho_b-\rho_f} \tag{21}$$

式中：$\Delta t_f, \Delta t$ 为声波时差的流体值和测井值；

ρ_b, ρ_f 为密度的流体值和测井值；

ϕ_{Nf}, ϕ_N 为中子的流体值和测井值。

在火山岩中最主要的是裂缝对 M、N 的影响。当地层中的储集空间主要是裂缝时对 M、N 的值有影响。当储层中发育裂缝时，且裂缝为高角度或垂直裂缝时，声波时差不受裂缝的影响，而密度测井值降低，因此导致 M 值将增大，而对于 N 值结果趋于保持不变。所以，当地层中发育高角度裂缝时，数据点向 M 值增大的方向移动，N 值基本不变。当地层发育低角度缝时，声波时差测井值明显增大，此时 M、N 值不受裂缝的影响。假如裂缝很发育，声波时差出现“周波跳跃”现象时，M 值会变小。

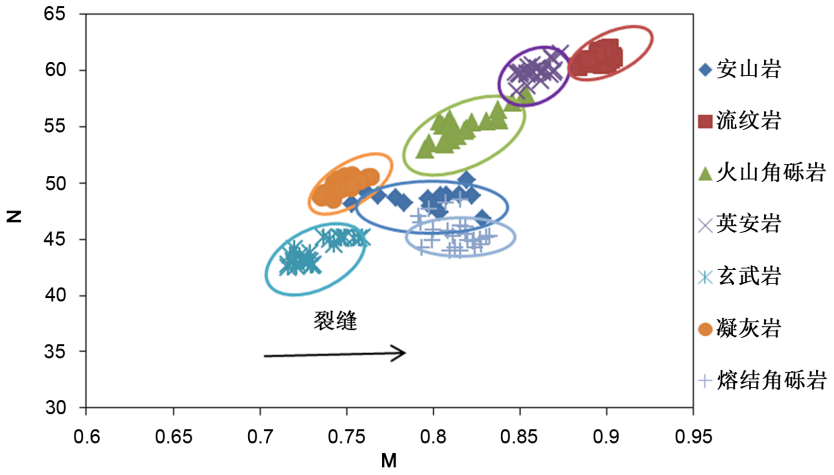


Figure 6. M-N intersection diagram
图 6. M-N 交会图

4.4. 机器学习岩性识别

利用机器学习建立岩性分类及预测模型，具体包括：样本库的建立；特征选择；测井特征数据的归一化；训练模型；评价模型以及岩性预测。

1) 样本库建立

根据取芯资料、测井数据和成像测井数据，分析测井数据与岩性的对应关系，生成样本空间。

2) 特征选择

前面已经具体分析了测井响应特征，优选岩性敏感的测井曲线，即 GR，DT，RHOB，CNL，RT，RI。同时构造两个新的特征，即 M 和 N。最终确定的岩性敏感特征为：GR，N，RI，M，CNL，DT，RT，RHOB 8 种特征，其占用的权重比依次为 0.22, 0.18, 0.16, 0.13, 0.11, 0.08, 0.07, 0.05。

3) 测井数据归一化

测井数据的预处理对推广到多井的岩性识别具有相当重要作用，因为相同的岩性数据利用不同规格仪器测量时，数值上会存在一定的系统偏差。数据归一化处理可以消除量纲影响，降低不同测井仪器测量绝对值误差导致的影响，利于后期建立的模型对全区井段都能有很好的预测效果。因此将前面选取的 8 种特征全部归一化到[0, 1]中。对于 GR，DT，RHOB，CNL，M，N 采用线性变量的归一化，即：

$$Y = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{22}$$

对于电阻 RT，RI 采用先取对数，在归一化，即：

$$Y = \frac{\log X - (\log X)_{\min}}{(\log X)_{\max} - (\log X)_{\min}} \tag{23}$$

如何选取变量的最大值和最小值变的格外重要，如果直接采用该变量的最大值，那么一些极端值会参与运算，干扰主体数据的归一化。为准确获得最大值和最小值，采用累积概率曲线一次导(斜率)的方法来获取特征的最大和最小。

根据各特征的数据，绘制累积概率分布曲线，曲线形态一般是从小到大数据的累积百分比变化很小，到达一定临界值时，数据变化开始变大，将这个临界值作为特征曲线的最小值，到达最后，数据变化慢慢变小，到达一个临界值后数据变化趋于稳定，把这个临界值作为该特征曲线的最大值，如下图 7 所示，在对 JIN202 井 DT 进行归一化时，根据两个的临界值取得最大值为 121 us/ft，最小值为 57 us/ft。原始 DT 数据分布范围为[48, 157]，剔除极端值后数据分布范围为[57, 121]，极端值占据数据范围比例为 41.3%，极端值占数据量比例为 3%。

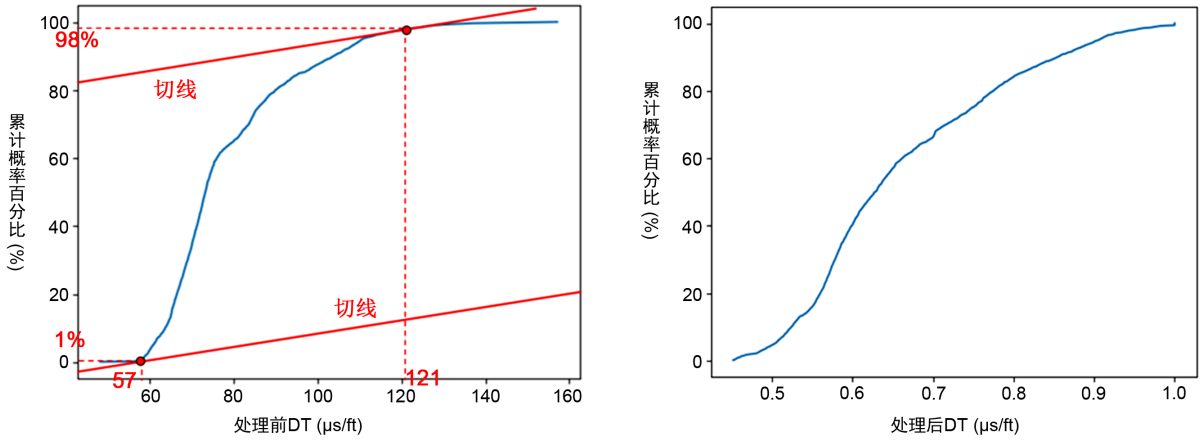
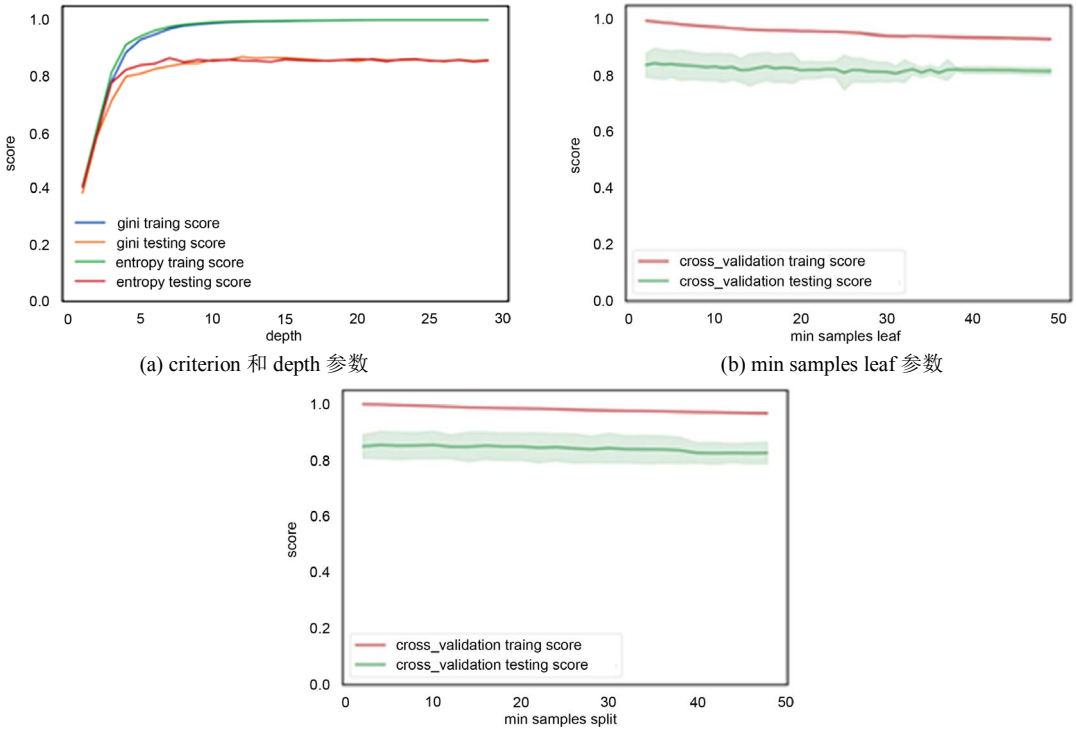


Figure 7. The cumulative probability curve of JIN202 DT before and after normalization
图 7. JIN202 DT 归一化前后累积概率曲线

4) 参数优选

根据前面建立的标签样本数据库，各类岩性标签总数为 4745 个，将样本数据随机切分为训练集和测试集，训练集占 70%，数量为 3321 个岩性样本点，测试集占 30%，1154 个岩性样本点。训练集用来训练模型，测试集用来检验模型的适用性，最终用该模型推广到全区。利用机器学习算法从数据集中学习到良好并稳定的模型，对于参数的选取至关重要。在参数选择的过程中，为了防止过拟合现象，本文采用交叉验证的方式，将数据集采用 5 折交叉验证。参数类型中，criterion 指定决策树结点切分质量评价准则。min_samples_split 参数指定每个内部节点(非叶节点)包含的最少样本数。min_samples_leaf 指定每个叶节点包含的最小的样本数。learning_rate 为每个学习器的学习率，就是权重的缩减系数或步长，取值范围为[0, 1]。n_estimators：即弱学习器的最大迭代次数，或者是最大的弱学习器的个数。一般来说 n_estimators 太小，容易欠拟合，n_estimators 太大，容易过拟合，一般选择一个较好的适中值，调参过程中与 learning_rate 一起考虑。subsample 抽取子样本的比例，取值为(0,1)，如果小于 1，则有部分样本去做 GBDT 的决策树拟合，小于 1 的比例可以减少方差，防止过拟合，但会增加拟合偏度，一般选择在[0.5, 0.8]。图 8、图 9、图 10、图 11 展现了决策树，随机森林和梯度提升树的各个参数优化。表 1 列出了不同方法最后的搜索结果。

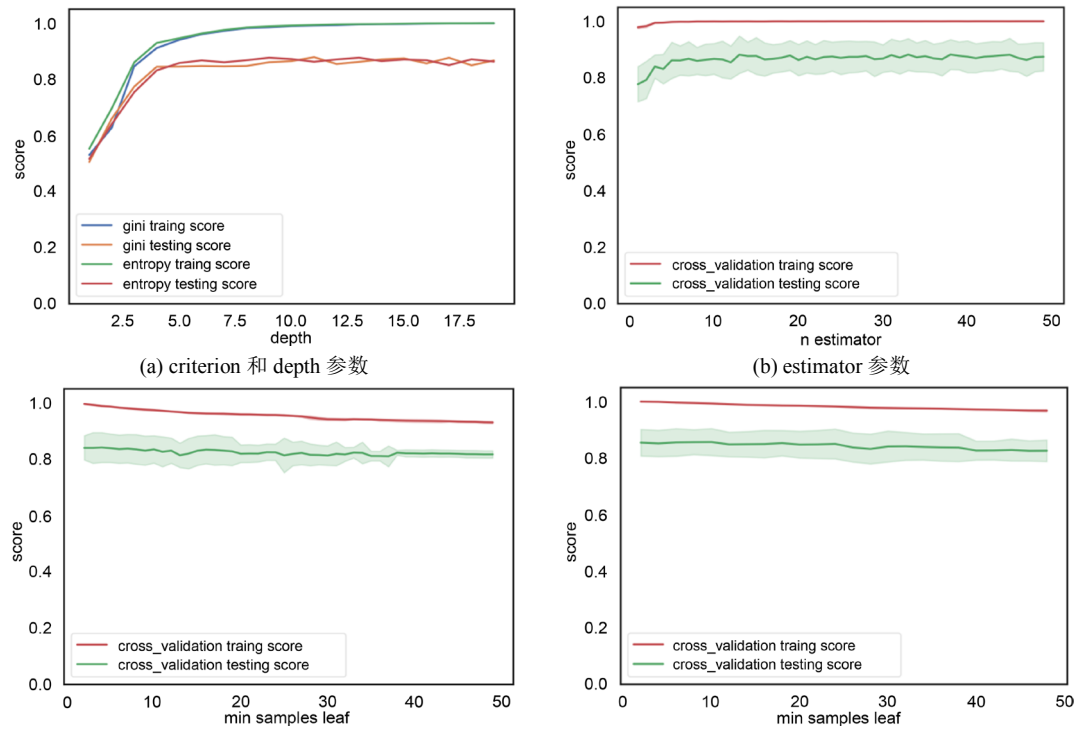


(a) criterion 和 depth 参数

(b) min samples leaf 参数

(c) min samples split 参数

Figure 8. Decision tree parameter optimization
图 8. 决策树参数优选



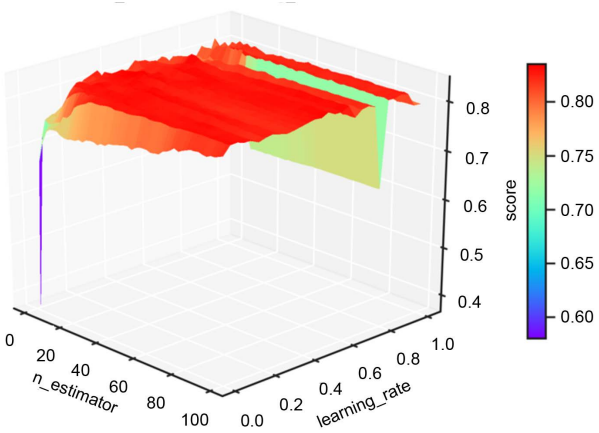
(a) criterion 和 depth 参数

(b) estimator 参数

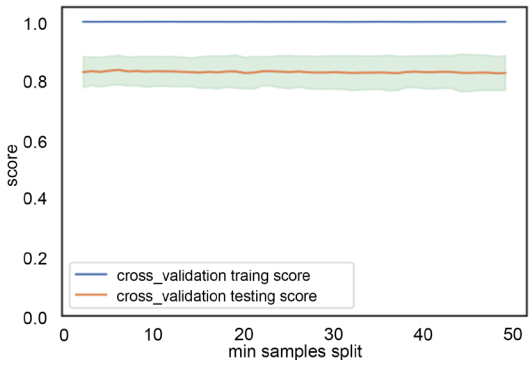
(c) min samples leaf 参数

(d) min samples split 参数

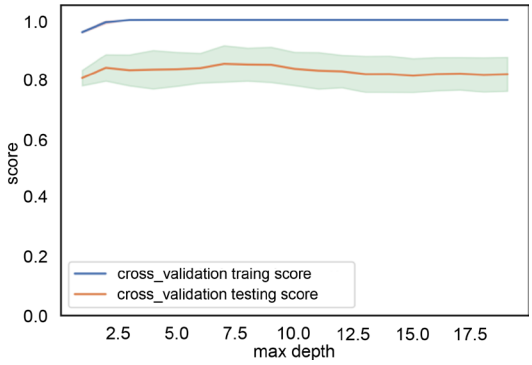
Figure 9. Random forest parameter optimization
图 9. 随机森林参数优选



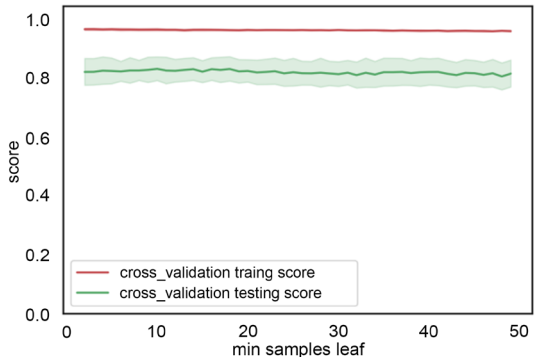
(a) estimator 和 learning rate 参数



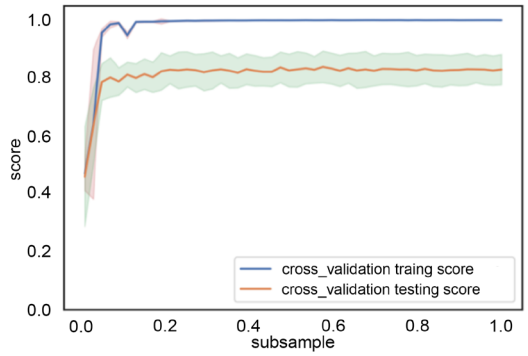
(a) depth 参数



(b) min samples split 参数



(c) subsample 参数



(d) min samples leaf 参数

Figure 10. Gradient boosting tree parameter optimization
图 10. 梯度提升树参数优选

Table 1. Parameter selection values of different methods
表 1. 不同方法的参数选择值

方法模型	参数名称	参数值
决策树	criterion	entropy(熵)
	max depth	7
	min samples split	3
	min samples leaf	3

Continued

随机森林	criterion	entropy(熵)
	estimator	25
	max depth	6
	min samples split	3
	min samples leaf	3
梯度提升树(GBDT)	learning rate	0.2
	estimator	20
	max depth	3
	min samples split	2
	min samples leaf	2
	subsample	0.6

5) 模型评价

a) 表 2 和表 3 是各类方法在训练和测试集上的查准率，是所有预测为该岩性的结果中真正是该岩性的比例。如在训练集上，随机森林方法在安山岩上查准率最高，表明预测为安山岩的结果中，真正是安山岩的概率最高 0.956，并且在测试集上随机森林准确率最高。

Table 2. Precision and recall rate of the test set
表 2. 测试集的查准率和召回率

岩性	样本数	随机森林		决策树		GBDT		贝叶斯	
		查准率	召回率	查准率	召回率	查准率	召回率	查准率	召回率
安山岩	232	0.956	0.862	0.921	0.806	0.933	0.853	0.917	0.767
火山角砾岩	269	0.925	0.973	0.868	0.959	0.909	0.970	0.918	0.966
流纹岩	330	0.955	0.984	0.946	0.957	0.951	0.957	0.804	0.824
凝灰岩	161	1	1	0.981	0.975	1	1	1	1
熔结角砾岩	133	0.977	0.984	0.962	0.969	0.963	0.984	0.992	0.947
玄武岩	101	1	0.980	0.990	0.980	1	0.960	0.970	0.970
英安岩	198	0.969	0.969	0.959	0.949	0.934	0.939	0.640	0.712
总	1424	0.962	0.962	0.936	0.936	0.948	0.948	0.867	0.867

Table 3. Precision and recall rate of training set
表 3. 训练集的查准率和召回率

岩性	样本数	随机森林		决策树		GBDT		贝叶斯	
		查准率	召回率	查准率	召回率	查准率	召回率	查准率	召回率
安山岩	526	0.991	0.901	0.903	0.855	0.953	0.853	0.936	0.750
火山角砾岩	647	0.941	0.993	0.903	0.955	0.920	0.981	0.908	0.964
流纹岩	743	0.976	0.993	0.956	0.966	0.954	0.977	0.808	0.816
凝灰岩	403	1	1	0.992	0.977	0.997	1	1	0.995

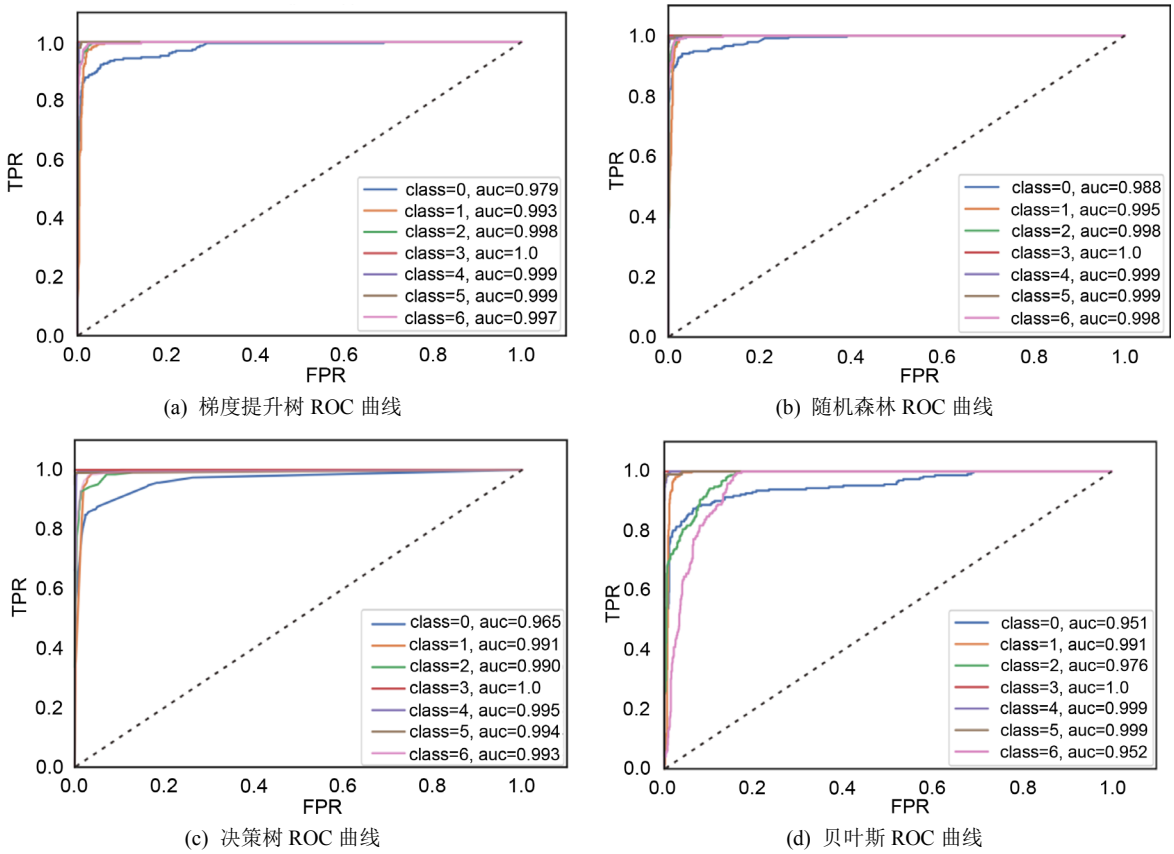
Continued

熔结角砾岩	300	0.986	1	0.957	0.980	0.961	1	0.967	0.980
玄武岩	229	1	1	0.969	0.978	1	0.982	0.982	0.965
英安岩	472	0.985	0.974	0.957	0.917	0.949	0.919	0.647	0.726
总	3320	0.978	0.978	0.943	0.943	0.955	0.955	0.868	0.868

b) ROC 曲线。

TPR 代表将正类样本分对的概率，FPR 代表将负类样本错分为正类的概率。ROC 空间中，每个点的横坐标是 FPR，纵坐标是 TPR。ROC 曲线中，(0,0)代表所有样本全部被判定为负类，(1,1)代表所有样本被判定为正类，(0,1)代表最完美分类。通过二维 ROC 曲线图形，能很直观的比较不同分类器模型的分类性能。从图 11 中看出，随机森林和梯度提升树的性能较好，能很好的向(0,1)点靠近。其次是决策树，而贝叶斯方法较差。

c) AUC 值。图 11 中“0”代表安山岩，“1”代表火山角砾岩，“2”代表流纹岩，“3”代表凝灰岩，“4”代表熔结角砾岩，“5”代表玄武岩，“6”代表英安岩。根据 AUC 值都接近 1 可知，随机森林和梯度提升树方法效果较好，每种岩性的 AUC 值都很接近于 1。



(a) 梯度提升树 ROC 曲线　(b) 随机森林 ROC 曲线

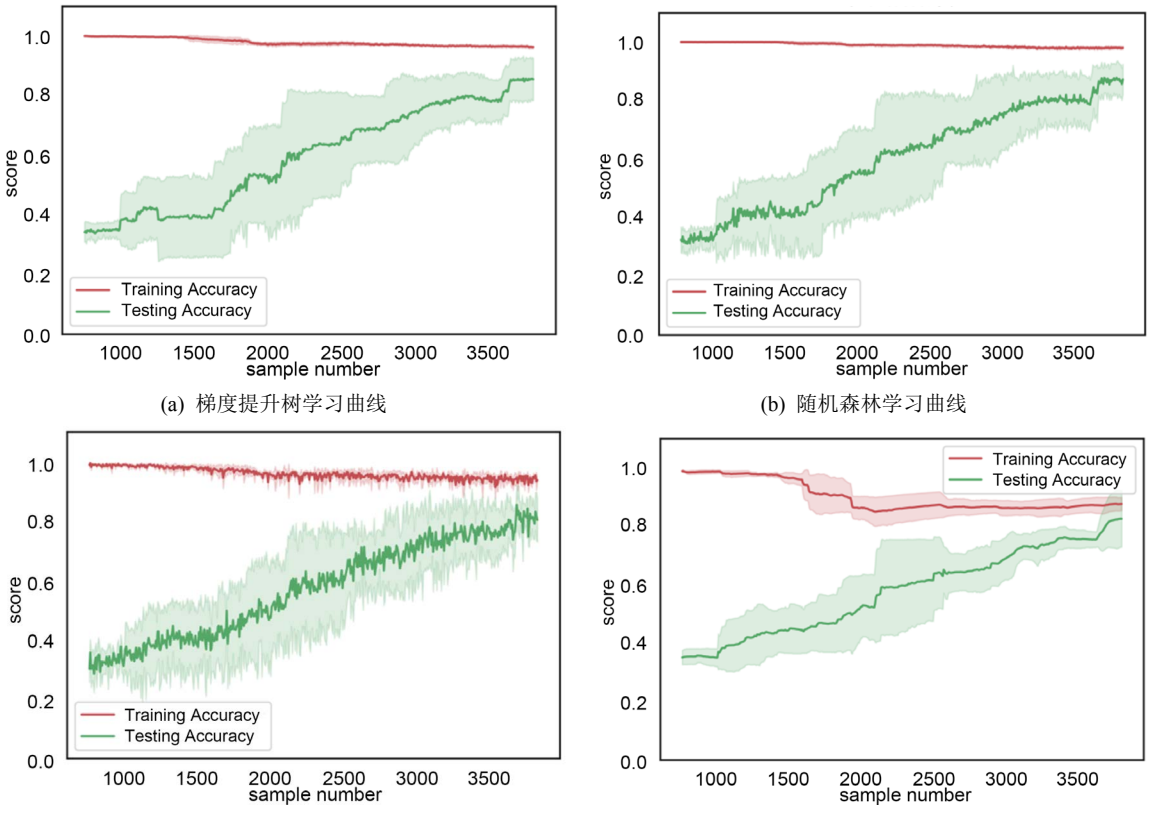
(c) 决策树 ROC 曲线　(d) 贝叶斯 ROC 曲线

Figure 11. ROC curves of different models
图 11. 不同模型 ROC 曲线

d) 模型稳定性

利用机器学习中的学习曲线判定训练模型的好坏。学习曲线中，横坐标为岩性样本数据量，纵坐标

为准确率。主要评估样本数据量的大小的变化对模型的影响。本次岩性识别岩性数据样本为 4745 个数据，每次从这些数据中按照不同比例随机抽取样本数据，本论文主要是在[10%, 100%]之间切分出 500 个不同比例的样本数据集，并分训练集和测试集。为确保模型的准确，采用交叉验证的方式在抽取的样本数据集内训练和测试，本次采用 5 折交叉验证。在梯度提升树和随机森林学习方法中，开始抽 500~1000 的岩性样本数据中，训练集上准确率高，测试集上准确率低，属于低偏差，高方差，模型处于欠拟合状态，随着数据点的增加，训练集上准确率下降缓慢，测试集上准确率明显升高，且方差变小，当数据点达到 4745 时，模型达到了低偏差低方差的稳定状态。在决策树模型中，根据学习曲线，上下波动，模型不稳定。在贝叶斯模型中，训练集准确率较低，且在随数据增大过程中，方差发生由大变小再变大的过程，模型不稳定。所以综上所述，梯度提升树和随机森林模型较稳定，见图 12。



(a) 梯度提升树学习曲线　(b) 随机森林学习曲线

(c) 决策树学习曲线　(d) 贝叶斯学习曲线

Figure 12. The learning curve of different models
图 12. 不同模型学习曲线

5. 预测效果

根据前面建立的模型对未知井段进行预测，为验证预测的准确性，采用前面未参与建模的取芯井进行预测验证。本文选取 JIN204 井和 KE301 井。JIN204 有三段取芯，第一段深度为 4193.3~4198.1 m，厚度为 4.8 m，取芯描述为安山岩。根据统计在该段随机森林方法预测为安山岩厚度为 4 m，预测准确率为 83%。决策树方法预测厚度为 3.45 m，预测准确率为 72%。贝叶斯方法预测厚度为 4 m，预测准确率为 83%。梯度提升树(GBDT)方法预测安山岩厚度为 2.88 m，预测准确率为 60%。第二段取芯深度为 4289.4~4294.2 m，厚度为 4.8 m，从图上可知，随机森林、决策树、贝叶斯和梯度提升树(GBDT)都预测准确为流纹岩，准确率为 100%。第三段取芯深度在 4329.3~4329.9 m，取芯描述该段岩性为流纹岩和凝灰岩，在深度为

4329.9~4330.5 m 的凝灰岩，随机森林和梯度提升树准确预测，而贝叶斯和决策树则预测错误。由于该段的测井曲线变化较大，流纹岩部分均未预测出来。在 KE301 井中(见图 13)，在图片道中，从上往下，深度为 3852.83 m、3854.28 m 和 3855.03 m 处的薄片定名均为安山岩，与四种方法预测均吻合。在 3849.0~3852.75 m 的深度段内，自然伽马，电阻率测井数据均有较明显的变化，在该段随机森林方法较好预测出来，见图 14。

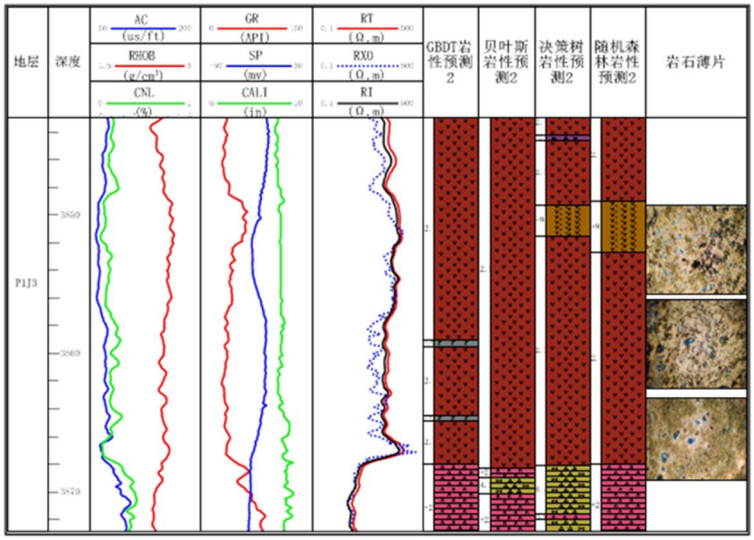


Figure 13. Lithology prediction of Well KE301
图 13. KE301 井岩性预测

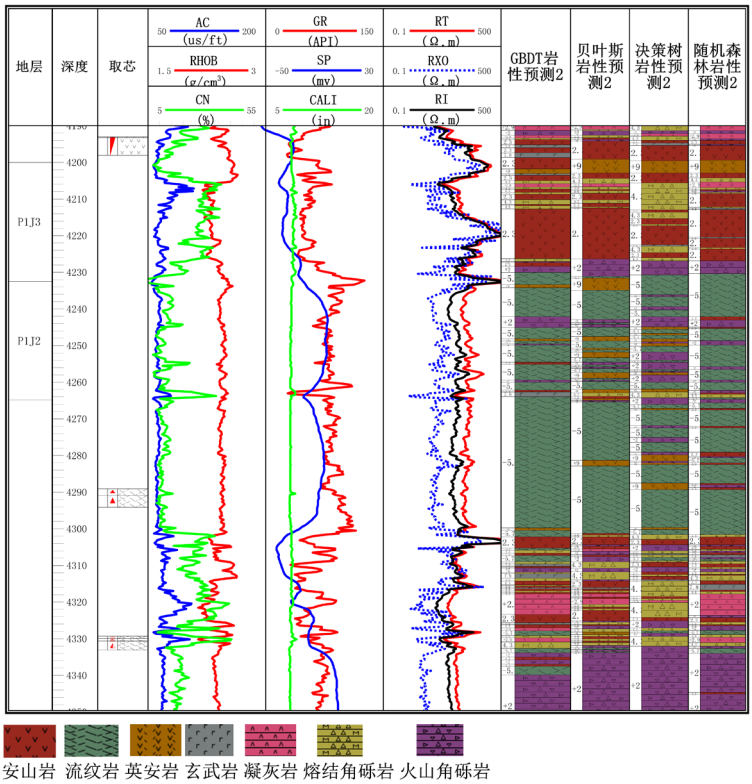


Figure 14. Lithology prediction of Well JIN204
图 14. JIN204 井岩性预测

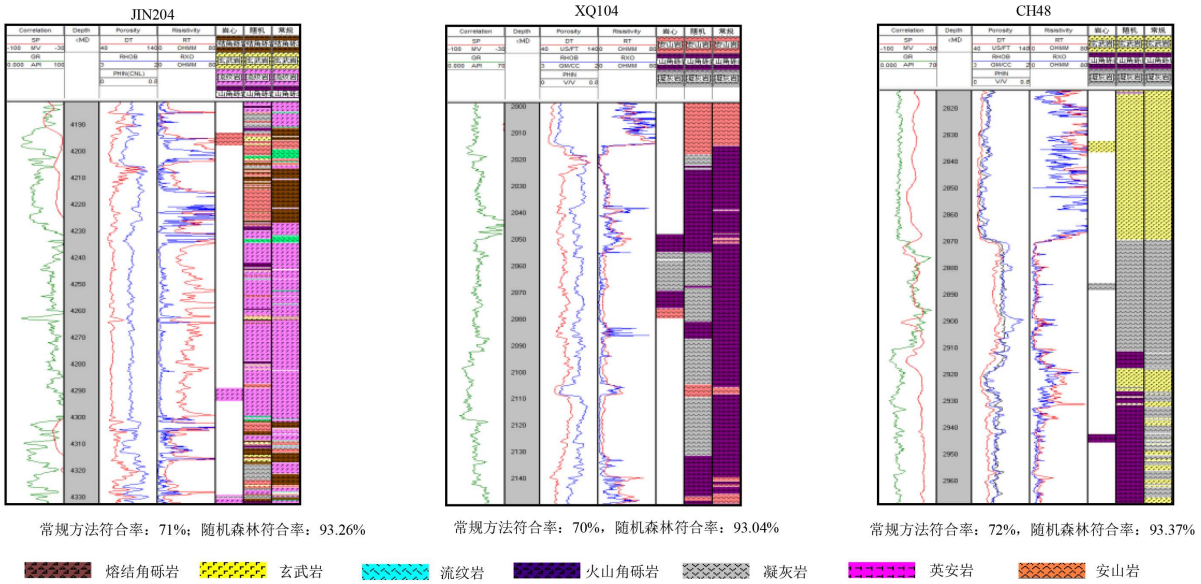


Figure 15. The results of intelligent recognition of lithology in multi-well volcanic reservoirs
图 15. 多井火山岩油藏岩性智能识别成果图

通过对四种方法的模型评价与优选表明，较好的模型有随机森林和梯度提升树两个模型。随机森林方法的准确率为 0.96，梯度提升树方法准率为 0.94，相差不大，根据盲井预测的统计结。如图 15 所示，利用常规方法进行复杂火山岩油藏岩性识别，符合率为 70%左右，而通过机器学习，特别是随机森林方法识别，符合率提高到 93%以上，效果显著。

6. 结论

1) 结合取芯，薄片，成像，实验分析化验数据确定岩性，并跟据测井响应特征分析标定测井曲线对应的岩性，形成测井曲线和岩性对应样本库，为后面机器学习训练预测打下基础。

2) 根据建立的样本库数据，对数据进行归一化处理，利用机器学习中的决策树，随机森林，梯度提升树和贝叶斯方法，采用交叉验证和网格搜索优选每个模型的最优参数，建立四种机器学习模型，并对这四种不同模型评价优选，优选出最优的随机森林方法模型。

3) 利用前面优选出学习最优的随机森林模型，对该研究区 45 口井岩性智能解释，通过对取芯井段的统计验证，与井段取芯的符合率在 90%以上。

参考文献

[1] 王乔. 火成岩裂缝地质——测井综合评价与地震预测[D]: [博士学位论文]. 长春: 吉林大学, 2016.

[2] Sanyal, S.K., Juprasert, S. and Jubasehe, J. (1980) An Evaluation of Rhyolite-Basalt-Voleanic Ash Sequence from Well Logs. *The Log Analyst*, **21**, 3-9.

[3] Benoit, W.R. and Darshan, K.S. (1980) Geothermal Well Log Analysis at Desert Peak, Nevada. *SPWLA* 21*st Annual Logging Symposium*, Lafayette, 8-11 July 1980.

[4] 范宜仁, 黄隆基, 代诗华. 交会图技术在火山岩岩性与裂缝识别中的应用[J]. 测井技术, 1999, 23(1): 53-56.

[5] 王泽华, 朱筱敏, 孙中春, 等. 测井资料用于盆地中火成岩岩性识别及岩相划分: 以准噶尔盆地为例[J]. 地学前缘, 2015, 22(3): 254-268.

[6] 谭伏霖, 王志章, 隆山, 等. 基于层次分解思想的火成岩岩性识别[J]. 测井技术, 2010, 34(2): 172-176.

[7] 谭伏霖, 王志章, 隆山, 等. 样品扩充法识别火成岩[J]. 中国石油大学学报(自然科学版), 2010(6): 45-49.

[8] 赵建, 高福红. 测井资料交会图法在火山岩岩性识别中的应用[J]. 世界地质, 2003, 22(2): 136-140.

[9] 王郑库, 欧成华, 李凤霞. 火山岩储层岩性识别方法研究[J]. 国外测井技术, 2007, 22(1): 8-11.

[10] 罗德江. 基于 Fisher 判别分析的弹性属性参数致密碎屑岩岩性识别[J]. 石油天然气学报, 2013(3): 85-89.

[11] 付光明, 严加永, 张昆, 等. 岩性识别技术现状与进展[J]. 地球物理学进展, 2017, 32(1): 26-40.

[12] 范存辉, 梁则亮, 秦启荣, 等. 基于测井参数的遗传 BP 神经网络识别火山岩岩性——以准噶尔盆地西北缘中拐凸起石炭系火山岩为例[J]. 石油天然气学报, 2012, 34(1): 68-71.

[13] 牟丹. 辽河盆地中基性火成岩测井岩性识别方法研究[D]: [博士学位论文]. 长春: 吉林大学, 2015.

[14] 牟丹, 王祝文, 黄玉龙, 等. 基于 SVM 测井数据的火山岩岩性识别——以辽河盆地东部坳陷为例[J]. 地球物理学报, 2015, 58(5): 1785-1793.

[15] Li, N., Qiao, D., Li, Q., *et al*. (2009) Theory on Logging Interpretation of Igneous Rocks and Its Application. *Petroleum Exploration and Development*, **36**, 683-692.

[16] 季桂树, 陈沛玲, 宋航. 决策树分类算法研究综述[J]. 科技广场, 2007(1): 9-12.

[17] 马骊. 随机森林算法的优化改进研究[D]: [硕士学位论文]. 广州: 暨南大学, 2016.

[18] 曹正凤. 随机森林算法优化研究[D]: [博士学位论文]. 北京: 首都经济贸易大学, 2016.

[19] 刘宇, 乔木. 基于聚类和 XGboost 算法的心脏病预测[J]. 计算机系统应用, 2019, 28(1): 228-232.

[20] 何世建. 基于梯度提升决策树与深度信念网络融合的推荐算法研究[D]: [硕士学位论文]. 桂林: 广西师范大学, 2017.

[21] 陈旋, 刘健, 冯新淇, 赵雪美. 基于朴素贝叶斯的差分隐私合成数据集发布算法[J]. 计算机科学, 2015, 42(1): 236-238.

[22] 何刚, 王志章, 谭伏霖, 等. 准噶尔盆地腹部火成岩分类及特征[J]. 新疆石油地质, 2010, 31(2): 125-127.

[23] 赵武生, 谭伏霖, 王志章, 等. 准噶尔盆地腹部火成岩岩性识别[J]. 天然气工业, 2010, 30(2): 21-25.

[24] 胡治华, 杨申谷, 夏锦芬, 等. 在火山岩相中的测井曲线特征及应用[J]. 天然气勘探与开发, 2007(3): 22-26.

[25] 尚玲, 谢亮, 姚卫江, 等. 准噶尔盆地中拐凸起石炭系火山岩岩性测井识别及应用[J]. 岩性油气藏, 2013, 25(2): 65-69.

[26] 许风光. 火成岩储层岩性识别及裂缝评价研究[D]: [硕士学位论文]. 青岛: 中国石油大学, 2004.

[27] 杨申谷, 刘笑翠, 胡志华, 等. 储层分析中火山岩岩性的测井识别[J]. 石油天然气学报, 2007, 39(6): 33-37.

[28] 邵阳. 含火山岩地层测井响应分析及岩性识别技术研究[D]: [硕士学位论文]. 大庆: 大庆石油学院, 2010.

[29] 陈建文, 魏斌, 李长山, 等. 火山岩岩性的测井识别[J]. 地学前缘, 2000, 7(4): 458.

[30] 覃豪, 李洪娟. 应用测井资料进行火山岩岩性识别[J]. 石油天然气学报, 2007, 29(3): 234-236.

[31] 王满. 基于 FMI 的火成岩组构分析[D]: [硕士学位论文]. 长春: 吉林大学, 2007.

[32] 张莹, 潘保芝, 印长海, 等. 成像测井图像在火山岩岩性识别中的应用[J]. 石油物探, 2007, 46(3): 288-293.

[33] 胡刚. 火山岩岩性识别方法研究[D]: [硕士学位论文]. 荆州: 长江大学, 2012.

[34] 高旭明, 张兵强, 庄玮. 利用测井资料识别火山岩岩性方法探讨[J]. 内蒙古石油化工, 2012(14): 42-43.

[35] 刘磊, 胡雪冰. 车排子地区东北部火山岩岩性–测井相特征及识别[J]. 河南科学, 2016, 34(6): 936-942.

[36] 刘传平, 郑建东, 杨景强. 徐深气田深层火山岩测井岩性识别方法[J]. 石油学报, 2006, 27(S1): 62-65.