doi: 10.11720/wtyht.2021.1508
郭建宏,张占松,张超谟,等.用地球物理测井资料预测煤层气含量——基于斜率关联度—随机森林方法的工作案例[J].物探与化探,2021,45(1):18-28. http://doi.org/10.11720/wtyht.2021.1508
Guo J H, Zhang Z S, Zhang C M, et al. The exploration of predicting CBM content by geophysical logging data: A case study based on slope correlation random forest method[J]. Geophysical and Geochemical Exploration, 2021, 45(1): 18-28. http://doi.org/10.11720/wtyht.2021.1508

用地球物理测井资料预测煤层气含量
——基于斜率关联度—随机森林方法的工作案例

郭建宏[1,2],张占松[1,2],张超谟[1,2],周雪晴[1,2],肖航[1,2],秦瑞宝[3],余杰[3]
(1.长江大学 地球物理与石油资源学院,湖北 武汉 430100; 2.长江大学 油气资源与勘探技术教育部重点实验室,湖北 武汉 430100; 3.中海油研究总院,北京 100027)

摘 要: 煤层气含量是煤层勘探开发研究的重点参数之一,由于煤层气含量受多因素影响,能有效预测其含量至关重要。本文将斜率关联度法与随机森林算法相结合,以地球物理测井资料为基础进行煤层气含量预测。首先利用改进的斜率关联度法,计算得到对煤层气含量敏感的测井曲线,再利用交叉验证法探究合适的随机森林决策树个数,并结合选出的超参数利用随机森林算法预测煤层气含量。以沁水煤田柿庄北区 3 号层为例,对该区块进行评价预测,并将预测结果与多元回归模型拟合结果进行对比,同时对本文方法模型的泛化性进行研究分析。结果表明,应用斜率关联度法对测井曲线与煤层气含量进行分析计算能准确有效地找到可用于煤层气含量预测的测井曲线;用随机森林算法训练得到的模型预测非夹矸段煤岩的煤层气含量准确,计算结果可信度高,在夹矸段预测能力较弱,总体对煤层气勘探开发有指导意义,具有实际应用价值。

关键词: 煤层气含量;斜率关联度法;测井曲线;随机森林;地球物理测井资料
中图分类号: P631　　**文献标识码:** A　　**文章编号:** 1000-8918(2021)01-0018-11

0 引言

煤层气勘探是近年来非常规油气资源开发的重点研究方向,准确评价煤层气含量对煤层气井单井产量预测与煤层气产能评估及勘探开发尤为关键[1-3]。煤层气资源作为非常规油气资源,储集与渗流机理与常规天然气差异较大[4],且煤层气含量受控于多因素,机理复杂,例如与其埋深、层厚,煤体结构及变质程度,以及储层压力、温度等地质因素均有一定关系[5-7]。评价煤储层气含量一直是煤层研究的重、难点,煤层气含量评价方法最为直接的是对煤层取心样本直接进行解吸测量,这一方法最为准确,但由于煤层大多较薄且机械强度差易破碎,导致煤层取心率低,对应煤心解吸实验资料较少[8]。国内外学者针对这一问题,结合煤层气储集机理与实验等,提出了一系列方法:从 KIM 法将储层因素与水分等工业组分相关联,后基于这一方法将工业组分引入并对其分析得到 KIM 改进方程[9-10];后有学者通过实验建立等温吸附模型,利用等温吸附线对煤层气含量进行预测,并基于这一理论提出兰氏煤阶方程进行评价预测[11-12]。

上述实验方法虽能评价煤层气含量,但多为对样本点进行评价,难以应用到整口井或整个区块,因此通过地球物理测井方法评价煤层气含量等煤层参数逐渐成为研究热点。相较于成本较高的取心方法,测井手段连续性强,性价比与可靠性均较高,将两者结合评价煤层气含量成为了接受度更高,使用更广泛的方法。利用地球物理测井资料预测煤层气含量的方法主要可概括为:原理法、数学地质法及数

收稿日期: 2020-11-08; **修回日期:** 2020-11-19
基金项目: 国家科技重大专项(2016ZX05060001-012)
作者简介: 郭建宏(1997-),男,山东招远人,主要研究方向为测井方法与解释、煤层气测井智能评价。Email:87942024@qq.com
通讯作者: 张占松(1965-),男,河南登封人,教授,主要从事测井方法与解释、油藏描述等工作。Email:Zhangzhs@yangtzeu.edu.cn

学统计法。原理法多为直接基于煤层测井资料,通过理论方法形成煤层气含量预测模型,例如将测井体积模型用于评价煤层气含量[13],或利用背景值法[14]计算煤层气含量,但两种方法中参数的选择对结果影响较大,且该类方法泛化性差,只能用于单井或单层评价。也有部分数学地质方法被用于煤层气含量预测,田敏等[15]将灰色系统理论结合实验数据对煤层气含量建立灰色多变量静态模型,随后郭建宏等[16]基于此将灰色多变量静态模型与测井曲线相结合将这一方法泛化性增强,能连续且准确地评价出整段煤层的气含量曲线,这类方法更多从数据上出发,得到的结果不一定能与理论完全相符。相比之下,数学统计法在煤层气含量预测中应用的更为广泛。由于煤层的复杂性,测井响应与煤层气含量间的关系也复杂多样,可能为线性亦或非线性关系,因而统计法多以回归分析及机器学习算法为主。回归分析法即是通过研究测井曲线与目标气含量的相关关系找到与煤层气含量敏感的测井曲线,利用最小二乘法计算出煤层气含量回归评价模型,这一方法简单且效果稳定,被广泛应用于煤层气含量评价。梁亚林等[17]利用测井曲线建立多元回归方程预测气含量并以此为基础对相应区块进行气含量预测,结果与地质情况相吻合;黄兆辉等[18]与金泽亮等[19]针对沁水盆地将多元线性回归法与兰氏方程相结合,建立煤层气含量评价模型,结果准确度较高,具有有效性。当线性关系难以表征煤层气含量与测井曲线间的关系时,可利用机器学习等方法进行预测,这类方法非线性逼近能力强,以神经网络方法为主,已有许多学者对此进行研究,将特征参数与目标参数通过神经网络进行训练形成网格模型,对测试集进行泛化性测试,以此评价模型的实用性。上述方法对存在潜在联系但无法直接用表达式展示的问题有明显优势,例如将煤层气含量与测井曲线资料通过 BP 神经网络进行训练,后对区块其他井进行验证发现这一方法预测煤层气含量精度高[20-21];随后支持向量机[22]等更多算法被引入到煤层气含量预测中。

在实际应用中,各类方法均受到不同程度的限制,体积模型法等原理传统方法受参数选择影响大且泛化性差而无法被推广使用;多元回归法由于各测井曲线对气含量响应的灵敏度不同使得结果会出现偏差,且这类方法对数据量要求大,与煤层取心率低样本少的特点相冲突;BP 神经网络训练的复杂性大,参数选择对模型影响大且对样本量有一定要求,使用局限性明显;支持向量机回归对小样本适用性强但容易过拟合;随机森林算法可利用袋外数据直接检测泛化性,且可利用有放回抽样解决样本数据少的问题[23],因此也被应用于复杂储层参数预测中[24],相比其他传统机器学习方法,随机森林算法更适合解决煤层小样本参数预测问题。基于此,笔者将斜率关联度法与随机森林相结合,基于测井曲线对煤层气含量进行斜率关联度分析,剔除冗余数据,即通过斜率关联法筛选出与煤层气含量敏感的测井曲线作为特征向量,并基于分析结果结合随机森林算法进行决策树个数优选,建立模型对煤层气含量进行预测,并用实际数据来验证本文方法的有效性与实用性。

1　基本原理

1.1　斜率关联度计算

一般关联度最早由邓聚龙教授提出,该分析法对样本数量小且分布无明显规律的数据有较强的实用性,计算结果与定性分析符合。一般关联度基本思想为将各序列与目标序列曲线形态进行对比,其几何形状接近,序列间关联度大,反之则小[25]。实际使用时,普通的关联度法存在缺陷,许多学者提出了改进,例如为了克服在规范性与保序性上的不足提出普通斜率关联度法[26],即在不同序列上对比各序列段斜率的接近程度来计算各序列间关联度大小,斜率越接近则关联度越大,反之则越小。后在此基础上进行了改进,对斜率的正负进行了计算[27],使其既能反映正关联也能找到负关联,极大提高了评价的精确性。规定一参考序列 x_0 与一对比序列 x_i,其形式分别为:

$$x_0 = \{x_0(k) \mid k = 1,2,3,\cdots,n\}, \tag{1}$$

$$x_i = \{x_i(k) \mid k = 1,2,3,\cdots,n\}, \tag{2}$$

则改进的斜率关联法公式为[28]:

$$\gamma_{(x_0,x_i)} = \frac{1}{n-1}\sum_{k=1}^{n-1}\delta(k)\cdot \frac{1}{1+\left|\frac{|x_0(k+1)-x_0(k)|}{\bar{\Delta}_0}-\frac{|x_i(k+1)-x_i(k)|}{\bar{\Delta}_i}\right|}, \tag{3}$$

式中:$\bar{\Delta}_0 = \frac{1}{n-1}\sum_{k=1}^{n}|x_0(k+1)-x_0(k)|$;$\bar{\Delta}_i = \frac{1}{n-1}\sum_{k=1}^{n}|x_i(k+1)-x_i(k)|$;$\delta(k)=\pm1$,当$[x_0(k+1)-x_0(k)][x_i(k+1)-x_i(k)]\geqslant 0$ 时值为 1,当$[x_0(k+1)-x_0(k)][x_i(k+1)-x_i(k)]<0$ 时值为-1。

1.2　随机森林

1.2.1　随机森林原理

随机森林法于2001年被提出[29]，该算法是一种以决策树为基础的集成算法，将单个决策树视作其对目标建立的模型结果进行综合得到新的模型。其中一组决策树可写为：$\{h(X,\theta_k),k=1,2,\cdots,K\}$。式中 θ_k 为随机变量，服从独立同分布，X 与 K 分别表示自变量与决策树的个数。随机森林预测的结果基于各决策树的结果取均值而得[29]：

$$\bar{h}(X)=\frac{1}{K}\sum_{k=1}^{K}\{h(X,\theta_k)\}。\tag{4}$$

为了防止模型出现过拟合或精度低的问题，通过引入 Bagging[23] 和随机子空间思想[30]。Bagging 即套袋思想，对原始样本有放回的进行 n 次抽取以生成训练样本，n 为原始样本量，并基于每个训练样本生成回归决策树 K。若 M 为原始样本，N 为原始样本中的样本，由于是有放回的进行抽取，则 S 中每个样本没被抽中的概率为 $\left(1-\frac{1}{N}\right)^N$，当 N 趋近于无穷大时则有：

$$\lim_{n\to\infty}\left(1-\frac{1}{N}\right)^N\approx\frac{1}{e}\approx 0.368,\tag{5}$$

即每棵树约有36.8%的样本未被抽取参与建模，将此类数据称为袋外数据(OOB，out of bag)。Bagging 思想在随机化建立更多的决策树时还保证其相互独立性。与 Bagging 思想类似，随机子空间思想可以保证不同树节点与其节点间的特征子集的差异性，以及树的独立性与多样性，即在构建决策树的过程中，每个分裂节点的特征数选取一般为从总特征空间 F 中随机抽取 f(推荐为 $f=\log_2 F$)个特征，并依照 Gini 指标选取最优特征进行分支生长。因而在随机森林回归中，决策树 K 与特征数 f 对模型预测性能存在显著影响。

1.2.2　随机森林泛化误差

以遵循独立同分布的随机向量 (X,Y) 为例，结合式(5)，则 $h(X)$ 对应均方泛化误差为：

$$E_{X,Y}(Y-h(X))^2,\tag{6}$$

在随机森林回归中，若决策树的个数趋于无穷时，存在：

$$E_{X,Y}(Y-\bar{h}(X,\theta_k))^2\to E_{X,Y}(Y-E_\theta h(X,\theta))^2=PE^*_{\text{tree}},\tag{7}$$

式中：θ_k 为第 k 个决策树的随机变量；E_θ 对应数学期望；PE^*_{tree} 为随机森林回归的泛化误差。若对于随机变量 θ，回归决策树无偏，有 $EY=E_X h(X,\theta)$，则：

$$PE^*_{\text{forest}}=E_\theta E_{X,Y}(Y-h(X,\theta))^2\leqslant\bar{\rho}PE^*_{\text{tree}},\tag{8}$$

式中：$\bar{\rho}$ 为剩余 $Y-h(X,\theta)$ 及 $Y-h(X,\theta')$ 的相关系数，θ 与 θ' 相互独立。综上，随机森林随着决策树数目不断增加最终会收敛且泛化误差会趋于一定值。

1.2.3　随机森林流程

随机森林回归算法流程为：

1）应用 boostrasp 采样随机生成训练数据集，未被抽中的为袋外数据，再随机抽取 m 个特征进行节点分裂，结合数据集中建模数据构建决策树；

2）按照上述方法构建 K 棵回归决策树，令其充分生长，不进行剪枝，形成随机森林；

3）利用袋外数据误差(OOB error)评价对效果进行评价，公式为：

$$\text{MSE}_{\text{OOB}}=\frac{1}{M}\sum_{1}^{M}(y_i-y_i^{\text{OOB}})^2,\tag{9}$$

式中：y_i 与 y_i^{OOB} 分别为目标实际值与模型对袋外误差数据的预测值；

4）利用上述步骤确定的模型对目标数据样本进行预测，随机森林各决策树预测结果的平均为最终预测输出结果。

1.3　煤层气含量评价步骤

结合本文实际内容，实行步骤为：

1）利用斜率关联度计算各测井曲线与煤层气含量的关联性，并根据实际计算结果筛选出有利于煤层气含量建模的数据；

2）利用选取出的测井曲线结合随机森林算法进行建模，并探究出合适的回归决策树的数目；

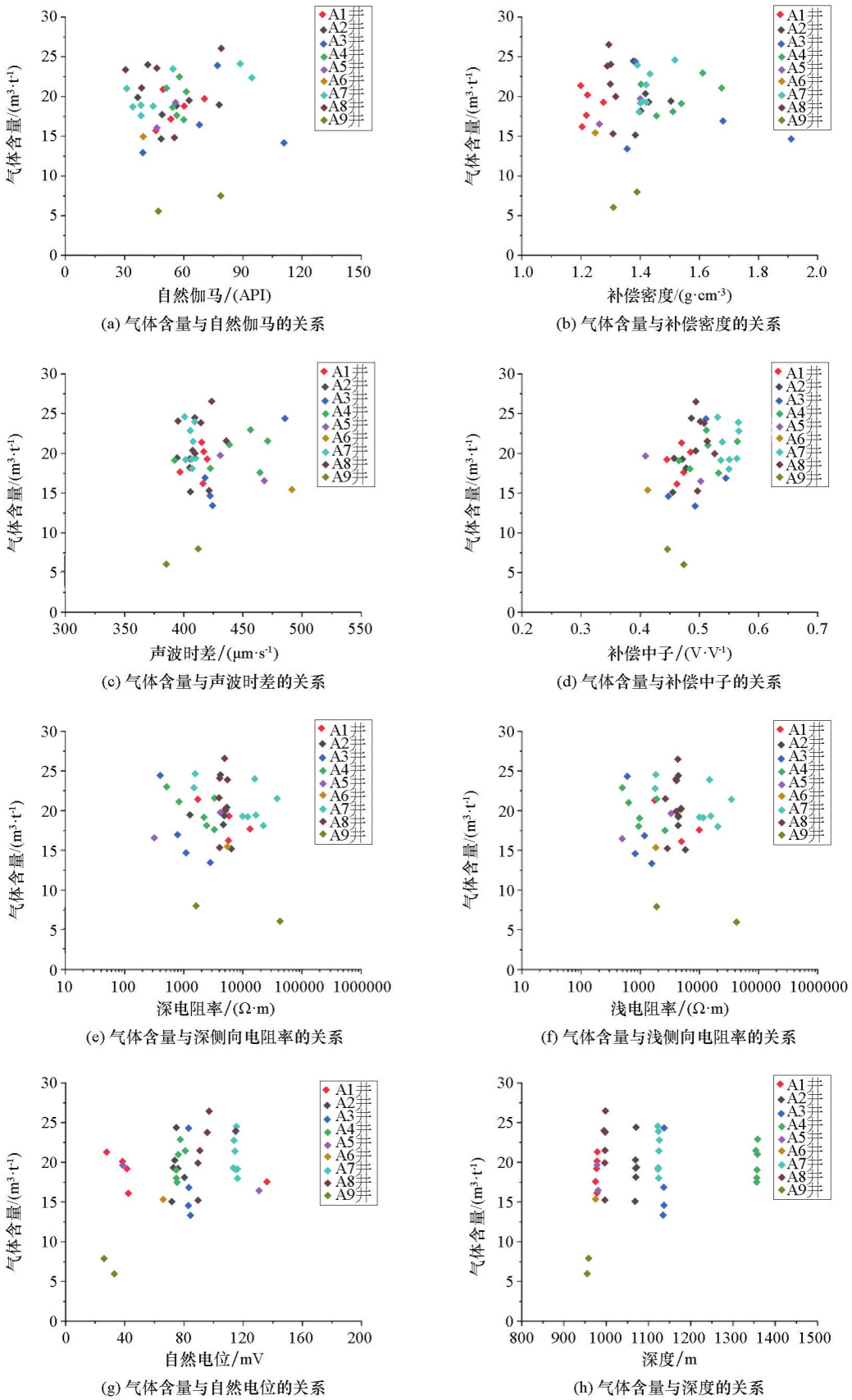
3）根据探究得到的特征个数与回归子树个数进行建模，并用未参与建模的数据进行预测验证。

2　煤层气含量预测模型

2.1　应用工区概况

使用沁水煤田柿庄北地区部分井3号煤层数据，结合本文所述方法对该区块3号层气含量进行评价预测。沁水煤田为石炭—二叠纪煤田，资源储量丰富，储层条件稳定，具有巨大开发潜力[31]。柿庄北区位于该区块，共取得该区块9口井共40组煤心数据，将煤心样本取得后，通过对样本进行多次采样实验测试对应样品气含量，最后对实验结果求取平均值。同时对煤心样本对应的深度段取平均深度值对应的各测井曲线响应值，并进行制表。表1为3号煤层标准化后的测井响应范围，图1为各测井响应曲线与煤层气含量交会图。

理论上，煤层埋深一定程度上决定了煤岩产生的气体能否有效储存，在埋深较浅处，煤层气含量随



(a) 气体含量与自然伽马的关系　(b) 气体含量与补偿密度的关系

(c) 气体含量与声波时差的关系　(d) 气体含量与补偿中子的关系

(e) 气体含量与深侧向电阻率的关系　(f) 气体含量与浅侧向电阻率的关系

(g) 气体含量与自然电位的关系　(h) 气体含量与深度的关系

图 1　煤层气含量与测井参数间的关系

Fig. 1　Relationship between coalbed methane content and logging parameters

表 1　3 号煤层测井响应范围

Table 1　Logging response range of No. 3 Coal Seam

参数	测试气量/($m^3 \cdot t^{-1}$)	自然伽马/API	自然电位/mV	补偿密度/($g \cdot cm^{-3}$)	声波时差/($\mu s \cdot m^{-1}$)	补偿中子/($V \cdot V^{-1}$)	深电阻率/($\Omega \cdot m$)	浅电阻率/($\Omega \cdot m$)
范围	5.91~26.07	30.4~109.4	26~134	1.19~1.89	384~489	0.41~0.56	165~33766	470~18355
平均值	18.83	55.3	82	1.39	419	0.49	4845	4428

深度增加而增大。孔隙度测井系列包含补偿密度测井、声波时差测井及补偿中子测井。由于煤的基质密度较低,煤层密度值随其致密程度的增加而增大,相应的孔隙度及气含量会降低,因而随着煤层气含量的增加,对应煤的体积密度减小,在补偿密度测井资料上补偿密度测井响应值与煤层气含量理论上应呈负相关关系;煤岩分子结构相对松散,声波时差测井曲线的响应为时差值较高,且其对储层含气性敏感,遇气层会明显增大或出现周波跳跃现象[32],理论上在声波时差测井资料上两者呈正相关关系;煤储层由碳、氢、氧组成且煤层气中含有甲烷,导致含氢指数高,使补偿中子测井资料呈现出一种虚高假象,而实际孔隙度通常较低。岩性测井系列提供了自然伽马测井曲线和自然电位测井曲线。由于煤的自然放射性通常较弱,煤的天然放射性多取决于成煤过程中的外来矿物质,粘土矿物会通过影响煤的吸附性能进而影响煤层气储集,煤层中粘土矿物增多,对应自然伽马测井响应增大,但煤层气含量由于有效孔隙降低而使得气含量减少,即在自然伽马测井资料上呈现出两者为负相关关系;在自然电位测井上,煤层的岩性相对更纯且导电性差,煤岩与泥浆间的化学作用和动电学作用弱,对应自然电位响应较低。电阻率测井系列提供了深、浅侧向电阻率曲线:煤岩电阻率受多因素影响,从煤层气含量考虑,气含量越大,电阻率测井响应越大。

从理论上分析后结合实际交会图进行判断,3 号煤层深度范围为 953~1 350 m 间,每口井实验样本数大多在 4~7 组,从交会图 1h 中可发现不同井 3 号层深度相近,与气含量无明显关联,总体上随深度增加煤层气含量增大。分析煤层气含量与孔隙度测井系列曲线的交会图,结合图 1b 及补偿密度测井资料得到的响应范围,3 号煤层补偿密度测井资料反映煤层的响应区间为 1.19~1.89 g/cm^3,但纯煤密度较低,若煤层中含泥岩夹矸则会使得补偿密度侧向响应值增大,将 A4 井中补偿密度过高值与 A9 井中气含量过低值剔除,则可发现煤层补偿密度测井值与煤层气含量呈负相关关系。图 1c 与图 1d 能看出声波时差测井曲线资料中的响应值与煤层气含量趋势上为正相关,但关系较差,补偿中子测井曲线资料上其响应值与煤层气含量呈正相关且关系相对明显,即 3 号煤层由于煤层气的存在将使得补偿中子测井资料的“虚高假象”更为突出。对应图 1a 与图 1g 分析,不同井自然伽马基线存在差异,每口井中存在自然伽马测井响应高值,这一原因多为煤层中泥岩夹矸所致,由于煤层中含泥岩夹矸段会导致自伽马测井响应异常增高进而直接影响了两者相关性;自然电位测井响应与煤层气含量总体上为正相关,但每口井中自然电位测井响应与煤层气含量无明显关系。煤岩电阻率受多方面因素影响,其变质程度、煤体结构、矿物质含量及分布等均会对电阻率测井响应值产生影响,通过图 1e 与 1f 分析,煤层气含量与深侧向电阻率总体上无相关关系,仅单井部分样品存在相关性,且煤层气含量与浅侧向电阻率相对深侧向电阻率存在差异,单井来看趋势也并不明显,多因煤层受泥浆侵入影响或扩径导致其表征的并非为原状地层。

综上分析可以看出,煤层气含量与地球物理测井曲线响应间的关系极为复杂,测井响应受多方面因素影响,煤岩本身以及夹矸存在等均会使得煤层段测井曲线响应出现变化。煤层取心率低,样本少,简单数据清洗会使得样本数据减少,且趋势也不一定能准确找到,而传统交会图分析对样本数据量有一定要求且容易受异常值的影响,因而靠交会图难以准确得到适合随机森林算法的特征参数。基于此,本文通过斜率关联度进行相关性分析,这一方法对实验数据具有更好的隐性挖掘能力,且受异常值影响相对小,能对样本数据总体与目标数据进行综合分析,不会由于单个异常点对结果产生较大影响。

2.2　斜率关联度计算

通过改进的斜率关联度法,对煤层测井曲线参数进行计算分析,表 2 为参与斜率关联度计算的数据,表 3 为斜率关联度计算结果。

通过表 3 可以得到 6 条与煤层气含量正关联的测井曲线,自然电位与浅侧向电阻率为负关联,正关联曲线中,均能找到理论支撑。在正关联曲线中,自然伽马曲线关联度相对其他测井曲线较低,为了验证这一曲线是否适合用于煤层气含量预测,利用随机森林中袋外误差曲线进行求证。如图2所示,

表 2　3 号煤层斜率关联度计算样本

Table 2　Calculation sample of slope correlation degree of No. 3 Coal Seam

样号	测试气量/($m^3 \cdot t^{-1}$)	深度曲线/m	自然伽马/API	自然电位/mV	补偿密度/($g \cdot cm^{-3}$)	声波时差/($\mu s \cdot m^{-1}$)	补偿中子/($V \cdot V^{-1}$)	深电阻率/($\Omega \cdot m$)	浅电阻率/($\Omega \cdot m$)
A1-1	17.32	972.94	52.96	134.34	1.21	395.9	0.47	11946	9024
A1-2	18.93	975.28	59.63	41.22	1.27	418.5	0.44	5121	3892
A1-3	15.90	976.00	45.51	42.10	1.20	415.1	0.46	5224	4563
⋮				⋮					
A9-2	7.81	956.55	77.90	25.93	1.38	411.0	0.44	1514	88

表 3　3 号煤层斜率关联度计算结果

Table 3　Calculation results of slope correlation degree of No. 3 coal seam

	$\gamma(x_0,x_1)$ 深度曲线	$\gamma(x_0,x_2)$ 自然伽马	$\gamma(x_0,x_3)$ 自然电位	$\gamma(x_0,x_4)$ 声波时差	$\gamma(x_0,x_5)$ 补偿密度	$\gamma(x_0,x_6)$ 补偿中子	$\gamma(x_0,x_7)$ 深电阻率	$\gamma(x_0,x_8)$ 浅电阻率
关联度	0.134	0.056	−0.049	0.163	0.183	0.168	0.196	−0.076
关联序	5	6	7	4	2	3	1	8

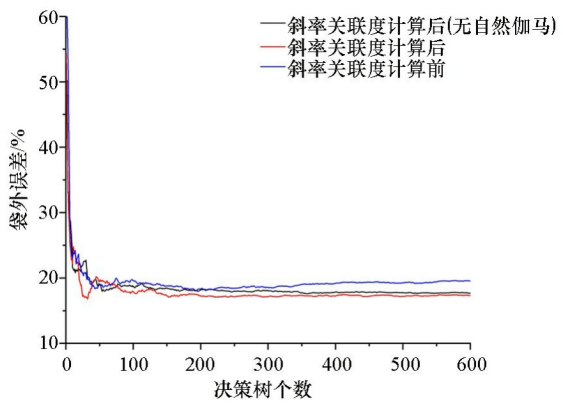


图 2　斜率关联度计算前后随机森林袋外误差结果

Fig. 2　Results of random forest out of bag error before and after slope correlation calculation

将随机森林决策树个数选定为 600 个,共作出 3 条曲线,曲线 1 为在斜率关联度计算后筛选出的曲线基础上去掉了自然伽马曲线得到的袋外误差数据,曲线 2 为斜率关联度计算筛选出的曲线得到的袋外误差数据,曲线 3 为未经斜率关联度计算的全曲线得到的袋外误差数据。经分析可发现,曲线 3 初始袋外误差大且收敛速度慢但相对稳定,经特征筛选后的袋外误差数据初始误差相对较小且收敛速度慢,曲线 1 与曲线 2 均在收敛过程中出现震荡,但很快趋于稳定,且最终曲线 2 袋外误差最低,即斜率关联度计算结果具有可靠性,包含自然伽马曲线的曲线特征组袋外误差相对低且收敛相对更快。因而证明斜率关联度能更深地发掘与煤层气含量相关的测井曲线,计算结果准确且与理论相符。

2.3　随机森林决策树优选

为了使随机森林建立的模型具有可靠性和对煤层气含量预测的有效性,需对随机森林的参数进行探究。在优选特征个数的基础上,还需确定决策树的个数。就随机森林这一算法而言,决策树个数的选择能直接影响模型的性能与精度,决策树过少,建立的模型精度低,数据利用不充分,模型效果发挥不充分,决策树过多会导致模型成型慢且增加过拟合发生的风险。由于煤层取心率低且数据稀少,为有效利用数据,将已有的 40 组数据随机分成 4 份,每组 10 个数据,其中 1 份为测试集,不参与随机森林建模,另外 3 份数据用于交叉验证以确定决策树的优选范围。具体做法为将 3 份原始数据中选取两组数据作为训练集对随机森林模型进行训练,再用另外一组数据进行验证,对验证集中的数据进行预测,以验证集中预测值与实验值的 MSE 作为判别指标。在 4 组分布中,为保证交叉验证的有效性,煤层气含量分布相对平均,除测试集外,另外 3 组数据中利用其中两组数据进行训练得到模型,预测另一组样本,通过观测预测结果随决策树个数变化来判断每组合适的决策树个数,结合 3 组结果进行判断。如图 3 所示,通过交叉验证,随着决策树个数不断增加,3 个组分别作为验证集时的预测值与实验值的均方误差逐渐稳定,在决策树为 500 个时,3 组验证集均方误差趋于稳定且达到低值,因而确定决策树个数为 500。如图 4 所示,以上述 3 组数据为训练集对随机森林进行训练得到模型,决策树个数设为 500,观察其袋外误差, 发现 500 个决策树时袋外误差已达到最低值且稳定,因而证明上述探究结果有效。

2.4　随机森林预测煤层气含量

基于上述对测井曲线特征的优选和对决策树个数的选择,利用上述 3 组训练集训练得到的随机森林模型预测测试集煤层气含量,结果如图5及表4

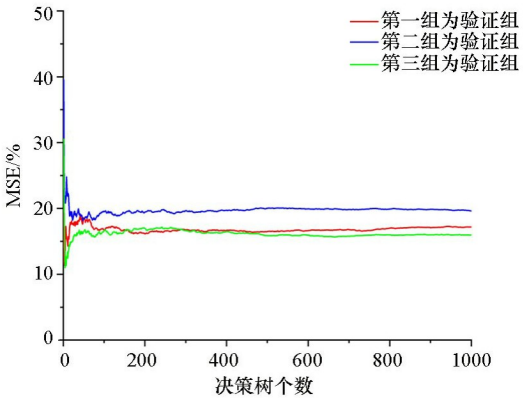


图 3　交叉验证探究决策树范围结果

Fig. 3　Cross validation to explore decision tree range results

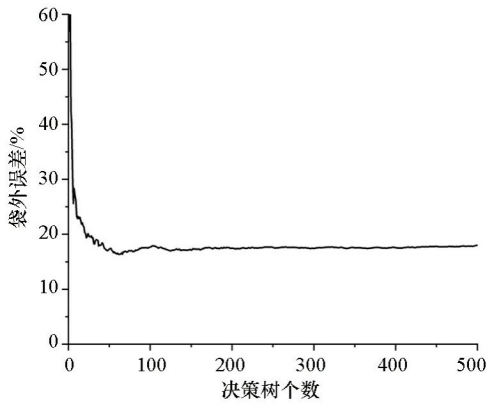


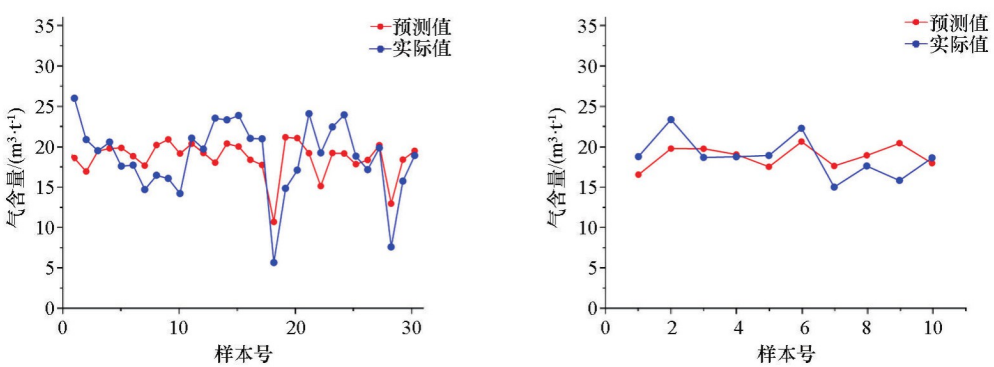
图 4　决策树个数为 500 时袋外误差

Fig. 4　Out of bag error when the number of decision trees is 500

所示。随机森林计算得到的模型在训练集回判相对误差为 19%，针对测试集预测，平均相对误差在 11.1%，并以此为基础对该区块单井 3 号煤层进行评价预测，以 A7 井为例，结果如图 6 所示。随机森林训练得到的模型在测试集上表现稳定，能有效预测煤层气含量，并能以此为基础对区块各井 3 号煤层进行煤层气含量曲线预测，且预测结果与实验结果相符合，说明该算法对训练集有效且泛化性强，能有效抗过拟合。此外，为了进行对比还对数据进行多元回归拟合，用同样曲线回归拟合出的模型在训练集与测试集上的平均相对误差分别为 21% 和 19%，误差均大于本文算法预测的结果，也说明本文方法相对应用较为广泛的多元回归法能进一步提升预测精度。在预测结果中，发现当煤层气含量为低值时的预测结果都存在较大误差，即含气量低值预测结果相对偏高，针对这一问题，笔者进行了分析。

2.5　误差异常值分析

如图 7 所示，以 A3 井为例，对比该井 3 号煤层测井响应值，发现煤层中下段部分存在响应异常值，7 号样本自然伽马测井响应值与补偿密度测井响应值明显偏高，深侧向电阻率测井响应值相对较浅部分减小且补偿密度测井响应值超出煤岩最大密度范围，结合柿庄北区综合柱状图发现，该区 3 号层存在泥岩或炭质泥岩岩性的夹矸，理论上自然伽马测井响应值增加，密度测井响应值增加与深侧向电阻率测井响应值减小理论上表征的应为煤层气含量减小，而 A3 井 7 号样本实验结果表明取心处气含量仅略低于其他处且与 3 号样本持平，这一现象会导致针对该样本的预测结果远低于实际实验情况，即夹矸的存在对煤层气含量预测结果造成了影响。综合分析，夹矸的存在对煤层测井响应会产生较大影响，自然伽马值与补偿密度值异常增高且泥岩电阻



(a) 训练集回判结果　　(b) 测试集预测结果

图 5　斜率关联度—随机森林预测煤层气含量结果

Fig. 5　Slope correlation degree-prediction of coalbed methane content by random forest

表 4　3 号煤层测试集预测结果

Table 4　Prediction results of No. 3 coal seam test set

	测试气量/($m^3 \cdot t^{-1}$)	预测气量/($m^3 \cdot t^{-1}$)	绝对误差/($m^3 \cdot t^{-1}$)	相对误差/%
A1-2	18.93	16.67	2.26	0.12
A7-5	23.53	18.93	4.60	0.20
A7-3	18.83	19.90	1.07	0.06
A7-2	18.91	19.19	0.28	0.01
A2-6	19.07	17.66	1.41	0.07
A8-5	22.44	20.79	1.65	0.07
A6-2	15.14	17.77	2.63	0.17
A4-2	17.76	19.08	1.32	0.07
A3-1	15.96	20.59	4.63	0.29
A4-3	18.77	18.10	0.67	0.04
平均值			2.05	11.10

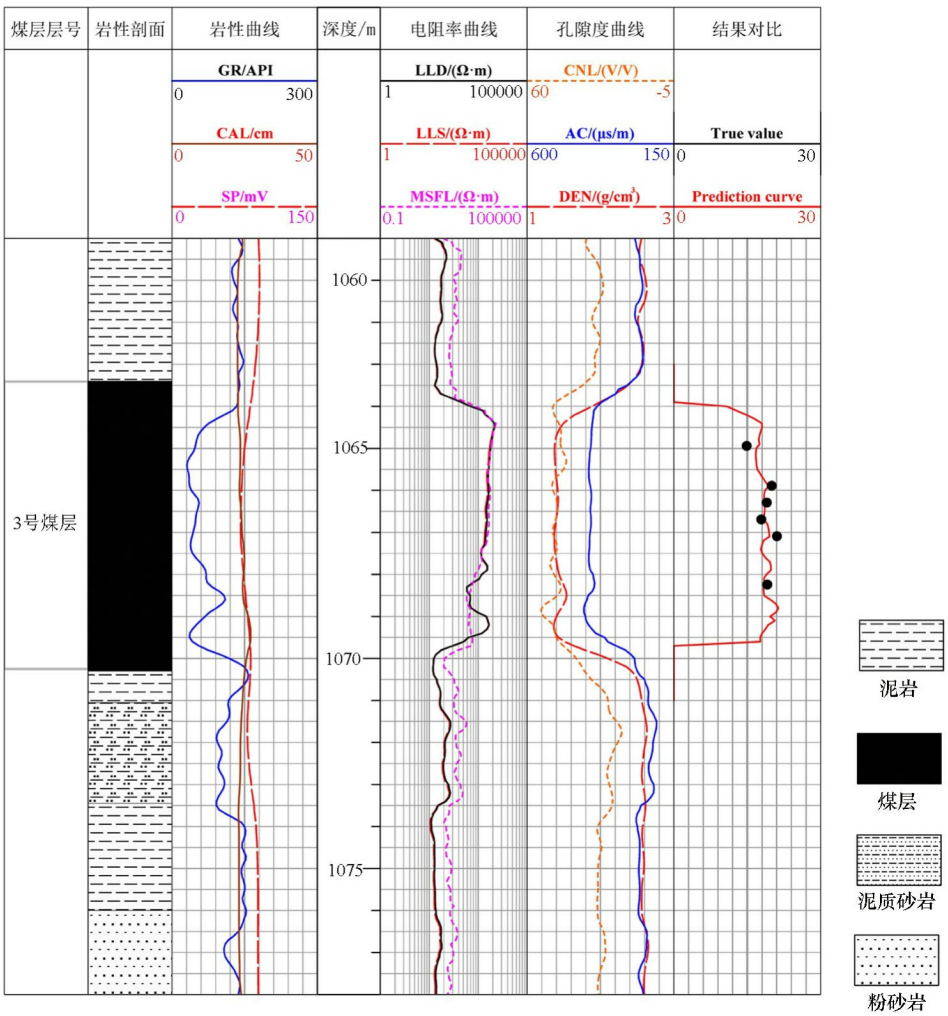


图 6　A7 井 3 号煤层气含量预测成果

Fig. 6　Prediction results of No. 3 coalbed methane content in A7 well

率低会使得电率测井资料响应值出现减小波动,所以对应夹矸深度段用于预测煤层气含量的测井资料响应会受到干扰,使得夹矸段气含量评价结果相对异常,而煤层取样难度大,样本量小,受夹矸影响的实验样本少,多元回归法或机器学习法都难以单独对这类情况进行建模评价,随机森林法对该类样本预测误差相对该算法对其他层段预测误差较大,为 38.4%,多元回归法对该井夹矸处气含量预测的相对误差为 54.8%,相比之下虽然随机森林算法预测误差相对略低,但预测效果均较差,两种方法都无法

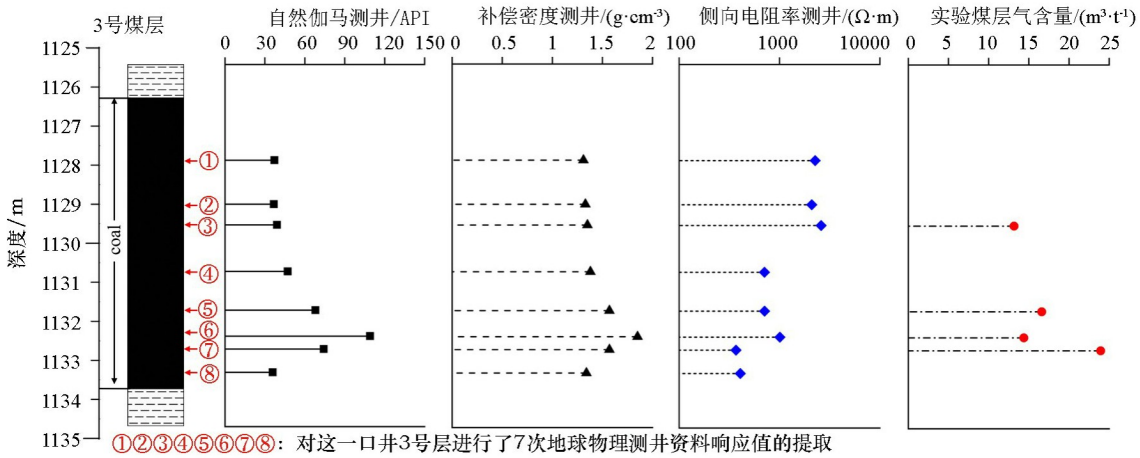


图 7　A3 井 3 号煤层响应与实验值分析

Fig. 7　Response and experimental value analysis of No. 3 coal seam in well A3

准确预测。因而随机森林算法能有效预测煤层非夹矸段气含量值,夹矸段难以准确预测,总体对生产上能进行准确指导,对煤层气含量预测评价提供了一种新的思路方法。

3　结论

1）斜率关联度算法能更好发掘测井资料与煤层气含量间的关系,通过对各条测井曲线与煤层气含量值进行斜率关联度计算分析,对于煤层气含量预测问题,自然伽马、补偿密度、声波时差、补偿中子、深侧向电阻率及深度与煤层气含量为正关联,利用上述测井曲线相对其他曲线组合能降低随机森林算法的袋外误差,提升该算法在煤层气含量预测能力上的泛化性。

2）针对随机森林算法的超参数中的决策树个数选择中,利用交叉验证计算得到决策树个数为 500 时,该算法学习效率达到稳定且能充分发挥算法性能,训练出的模型准确且强健。

3）通过实际计算分析,利用斜率关联度—随机森林法能有效预测煤层气含量,计算精度相对多元回归法更高,但在煤岩夹矸段煤层气含量预测效果欠佳,总体上能有效评价区块煤层气含量。

综上,利用斜率关联度—随机森林法能有效预测煤层气含量,构建出的模型强健且泛化性强,实际应用价值突出,对煤层气勘探开发可提供帮助。

参考文献(References):

[1] 赵庆波. 中国煤层气地质特征及勘探新领域[J]. 天然气工业, 2004, 24(5):4－7.
Zhao Q B. Geological features of the coalbed methane in China and its new exploration domains[J]. Natural Gas Industry, 2004, 24(5):4－7.

[2] 孟召平, 田永东, 雷旸. 煤层含气量预测的 BP 神经网络模型与应用[J]. 中国矿业大学学报, 2008(4):28－33.
Meng Z P, Tian Y D, Lei Y. Prediction models of coal bed gas content based on BP neural networks and its applications[J]. Journal of China University of Mining & Technology, 2008(4):28－33.

[3] 连承波, 赵永军, 李汉林, 等. 煤层含气量的主控因素及定量预测[J]. 煤炭学报, 2005, 30(6):726－729.
Lian C B, Zhao Y J, Li H L, ei al. Main controlling factors analysis and prediction of coal bed gas content[J]. Journal of China Coal Society, 2005, 30(6):726－729.

[4] 娄剑青. 影响煤层气井产量的因素分析[J]. 天然气工业, 2004, 24(4):62－64.
Lou J Q. Factors of influencing production of coal bed gas wells [J]. Natrual Gas Industry, 2004, 24(4):62－64.

[5] 高波, 马玉贞, 陶明信, 等. 煤层气富集高产的主控因素[J]. 沉积学报, 2003, 21(2):345－349.
Gao B, Ma Y Z, Tao M X, et al. Main controlling factors analysis of enrichment condition of coalbed methane[J]. Acta Sedimentologica Sinica, 2003, 21(2):345－349.

[6] 吴永平, 李仲东, 王允诚. 煤层气储层异常压力的成因机理及受控因素[J]. 煤炭学报, 2006, 31(4):475－479.
Wu Y P, Li Z D, Wang Y C. The formation mechanisms of abnormal pressure and factor in control of the coal bed gas in Qinshui Basin[J]. Journal of China Coal Society, 2006, 31(4):475－479.

[7] 叶建平, 武强, 王子和. 水文地质条件对煤层气赋存的控制作用[J]. 煤炭学报, 2001, 26(5):63－67.
Ye J P, Wu Q, Wang Z H. Controlled characteristics of hydrogeological conditions on the coalbed methane migration and accumulation[J]. Journal of China Coal Society, 2001, 26(5):63－67.

[8] 陈跃, 汤达祯, 许浩, 等. 基于测井信息的韩城地区煤体结构的分布规律[J]. 煤炭学报, 2013, 38(8):1435－1442.
Chen Y, Tang D Z, Xu H, et al. The distribution of coal structure in Hancheng based on well logging data[J]. Journal of China Coal Society, 2013, 38(8):1435－1442.

[9] 李贵红, 张鸿, 崔永君, 等. 基于多元逐步回归分析的煤储层含气量预测模型——以沁水盆地为例[J]. 煤田地质与勘探, 2005, 33(2):22-25.
Li G H, Zhang H, Cui Y J, et al. A predictive model of gas content in coal reservoirs based on multiple stepwise regression analysis: A case study from Qinshui Basin[J]. Coal Geology & Exploration, 2005, 33(2):22-25.

[10] Kim A G. Estimating methane content of bituminous coal beds from adsorption data[R]. United States Department of the Interior, Report of Investigations-Bureau of Mines 8245,1977:1-11.

[11] Ahmed U, Johnston D, Colson L. An advanced and integrated approach to coal formation evaluation[C]//SPE22736, 1991,755-770.

[12] Hawkins J M, Schraufnagel R A, Olszewsk A J. Estimating coal bed gas content and sorption isotherm using well log data[C]//SPE24905, 1992:491-501.

[13] 邵先杰, 孙玉波, 孙景民, 等. 煤岩参数测井解释方法——以韩城矿区为例[J]. 石油勘探与开发, 2013, 40(5):559-565.
Shao X J, Sun Y B, Sun J M, et al. Logging interpretation of coal petrologic parameters: A case study of Hancheng mining area[J]. Petroleum Exploration and Development, 2013, 40(5):559-565.

[14] 董红, 侯俊胜, 李能根, 等. 煤层煤质和含气量的测井评价方法及其应用[J]. 物探与化探, 2001, 25(2):138-143.
Dong H, Hou J S, Li N G, et al. The logging evaluation method for coal quality and methane[J]. Geophysical and Geochemical Exploration, 2001, 25(2):138-143.

[15] 田敏, 赵永军, 颛孙鹏程. 灰色系统理论在煤层气含量预测中的应用[J]. 煤田地质与勘探, 2008, 36(2):24-27.
Tian M, Zhao Y J, Zhuansun P C. Application of grey system theroy in prediction of coalbed methane content[J]. Coal Geology & Exploration, 2008, 36(2):24-27.

[16] 郭建宏, 张占松, 张超谟, 等. 基于灰色系统与测井方法的煤层气含量预测及应用[J]. 物探与化探, 2020, 44(5): 1190-1200.
Guo J H, Zhang Z S, Zhang C M, et al. Prediction and application of coalbed methane content based on grey system and logging method[J]. Geophysical & Geochemical Exploration, 2020, 44(5):1190-1200.

[17] 梁亚林, 原文涛. 测井预测煤层气含量及分布规律——以山西省沁水煤田为例[J]. 物探与化探, 2018, 42(6):1144-1149.
Liang Y L, Yuan W T. The prediction of the content and distribution of coalbed gas: A case study in the Qinshui coalfield based on logging[J]. Geophysical and Geochemical Exploration, 2018, 42(6): 1144-1149.

[18] 黄兆辉, 邹长春, 杨玉卿, 等. 沁水盆地南部 TS 地区煤层气储层测井评价方法[J]. 现代地质, 2012, 26(6):1275-1282.
Huang Z H, Zou C C, Yang Y Q, et al. Coal bed methane reservoir evaluation from wireline logs in TS District, southern Qinshui Basin[J]. Geoscience, 2012, 26(6):1275-1282.

[19] 金泽亮, 薛海飞, 高海滨, 等. 煤层气储层测井评价技术及应用[J]. 煤田地质与勘探, 2013, 41(2):42-45.
Jin Z L, Xue H F, Gao H B, et al. Technology for evaluation of CBM reservoir logging and its application[J]. Coal Geology & Exploration, 2013, 41(2):42-45.

[20] 潘和平, 黄智辉. 煤层含气量测井解释方法探讨[J]. 煤田地质与勘探, 1998, 26(2):58-60.
Pan H P, Huang Z H. Discussion on the interpretation method of coalbed methane content[J]. Coal Geology & Exploration, 1998, 26(2):58-60.

[21] 吴东平, 吴春萍, 岳晓燕. 煤层气测井评价的神经网络技术[J]. 天然气勘探与开发, 2001, 24(1):31-34.
Wu D P, Wu C P, Yue X Y. Neural network of coal bed gas logging evaluation[J]. Natural Gas Exploration & Development, 2001, 24(1):31-34.

[22] 连承波, 赵永军, 李汉林, 等. 基于支持向量机回归的煤层含气量预测[J]. 西安科技大学学报, 2008, 28(4):707-709.
Lian C B, Zhao Y J, Li H L, et al. Prediction of coal bed gas content based on support vector machine regression[J]. Journal Center of Xi'an University of Science and Technology, 2008, 28(4):707-709.

[23] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123-140.

[24] 冯明刚, 严伟, 葛新民, 等. 利用随机森林回归算法预测总有机碳含量[J]. 矿物岩石地球化学通报, 2018, 37(3):475-481.
Feng M G, Yan W, Ge X M, et al. Predicting total organic carbon content by random forest regression algorithm[J]. Bulletin of Mineralogy, Petrology and Geochemistry, 2018, 37(3):475-481.

[25] 肖新平, 谢录臣, 黄定荣. 灰色关联度计算的改进及其应用[J]. 数理统计与管理, 1995, 14(5):27-30.
Xiao X P, Xie L C, Huang D R. A modified computation method of grey correlation degree and its application[J]. Journal of Applied Statistics and Management, 1995, 14(5):27-30.

[26] 马保国, 成国庆. 一种相似性关联度公式[J]. 系统工程理论与实践, 2000(7):69-71.
Ma B G, Cheng G Q. A formula of similarity correlation degree[J]. Systems Engineering-Theory & Practice, 2000(7):69-71.

[27] 李明凉. 灰色关联度新判别准则及其计算公式[J]. 系统工程, 1998, 16(1):68-70.
Li M L. A new descriminant byelaw for grey interconnet degree and its calculation formulas[J]. Systems Engineering, 1998, 16(1):68-70.

[28] 张绍良, 张国良. 灰色关联度计算方法比较及存在问题分析[J]. 系统工程, 1996, 14(3): 45-49.
Zhang S L, Zhang G L. Comparison between computation modles of grey interconnet degree and analysis on their shortages[J]. Systems Engineering, 1996, 14(3):45-49.

[29] Breiman L, Cutler A. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[30] Ho T K. The random subspace method for constructing decision forests[J]. IEEE Transactions on Pattern Analysisand Machine

Intelligence, 1998, 20(8): 832-844.

[31] 贾承造, 郑民, 张永峰. 中国非常规油气资源与勘探开发前景[J]. 石油勘探与开发, 2012, 39(2):129-136.
Jia C Z, Zheng M, Zhang Y F. Unconventional hydrocarbon resources in China and the prospect of exploration and development [J]. Petroleum Exploration and Development, 2012, 39(2):129-136.

[32] 雍世和, 张超谟. 测井数据处理与综合解释[M]. 东营:中国石油大学出版社, 2007:134-139.
Yong S H, Zhang C M. Logging data processing and comprehensive interpretation[M]. Dongying: China University of Petroleum Press, 2007:134-139.

The exploration of predicting CBM content by geophysical logging data: A case study based on slope correlation random forest method

GUO Jian-Hong[1,2], ZHANG Zhan-Song[1,2], ZHANG Chao-Mo[1,2], ZHOU Xue-Qing[1,2], XIAO Hang[1,2], QIN Rui-Bao[3], YU Jie[3]

(1. *College of Physics and Petroleum Resources, Yangtze University, Wuhan* 430100, *China*; 2. *Key Laboratory of Exploration Technologies for Oil and Gas Resources, Ministry of Education, Yangtze University, Wuhan* 430100, *China*; 3. *CNOOC Research Institute, Beijing* 100027, *China*)

Abstract: Coalbed methane content is one of the key parameters in coal seam exploration and development research. Due to the influence of many factors on coalbed methane content, it is very important to predict coalbed methane content effectively. In this paper, slope correlation degree method and random forest algorithm are combined to predict coalbed methane content based on geophysical logging data. Firstly, the improved slope correlation degree method is used to obtain the favorable geophysical logging curves for CBM content prediction, and then the cross validation method is used to explore the appropriate number of random forest decision trees, and the random forest algorithm is used to predict the coalbed methane content for the logging curve sequence with positive correlation. With the No. 3 seam in Shizhuang north area of Qinshui coalfield as an example, the block was evaluated and predicted with the results compared with the results of multiple regression model, and the anti-interference ability of the model was studied and analyzed. The results show that the application of slope correlation method to analyzing and calculating the geophysical logging curve and coalbed methane content can accurately and effectively find the logging curve that can be used to predict the content of coalbed methane, the model trained by random forest algorithm is accurate in predicting the content of coalbed methane in the non-gangue section, and the calculation result has high reliability, but the prediction ability is weak in the gangue section. The results obtained by the authors are of guiding significance to the exploration and development of coalbed methane and have practical application value.

Key words: coalbed methane content; slope correlation method; logging curve; random forest; geophysical logging data

(本文编辑:王萌)