Advances in Geosciences 地球科学前沿, 2021, 11(5), 701-709
Published Online May 2021 in Hans. http://www.hanspub.org/journal/ag
https://doi.org/10.12677/AG.2021.115065

Hans汉斯

随机森林储层预测及关键参数探讨

——以SC某工区储层预测为例

范　晓，邹　文，刘　璞，李　乐，张　茜

东方地球物理公司物探技术研究中心成都分中心，四川　成都
Email: fanxiao_wt@cnpc.com.cn

收稿日期：2021年4月13日；录用日期：2021年5月24日；发布日期：2021年5月31日

摘　要

随机森林是一种高度灵活的机器学习算法，可以解决回归和分类的问题。本文建立随机森林地震储层预测流程，探讨了随机森林储层预测过程中的关键参数，并利用SC某工区储层伽马值预测为例，分别对比分析了不同参数对预测结果的影响，并给出最优参数设置。同时对参与建模的地震属性重要性进行分析，优化属性样本集。最后利用最佳随机森林模型，以优化属性样本集为输入开展储层预测，取得良好效果。实践证明，随机森林算法能够有效进行储层预测，随机森林模型的参数调优很重要，影响到算法的效率和最终预测精度。

关键词

随机森林，决策树，关键参数，储层预测，重要性分析

Random Forest Reservoir Prediction and Key Parameters Discussion

—Taking Reservoir Prediction of a Work Area in SC as an Example

Xiao Fan, Wen Zou, Pu Liu, Le Li, Xi Zhang

Chengdu Branch, Research and Development Center, BGP Inc., CNPC, Chengdu Sichuan
Email: fanxiao_wt@cnpc.com.cn

Received: Apr. 13th, 2021; accepted: May 24th, 2021; published: May 31st, 2021

Abstract

Random forest is a highly flexible machine learning algorithm, which can solve the problems of

文章引用: 范晓, 邹文, 刘璞, 李乐, 张茜. 随机森林储层预测及关键参数探讨[J]. 地球科学前沿, 2021, 11(5): 701-709.
DOI: 10.12677/AG.2021.115065

regression and classification. According to the process of random forest seismic reservoir prediction, this paper discusses the key parameters in the process of random forest reservoir prediction. Taking the reservoir gamma prediction of a work area in SC as an example, the influence of different parameters on the prediction results is compared and analyzed, and the optimal parameter settings are given. At the same time, the importance of seismic attributes is analyzed, and the optimal attribute sample set is obtained. The optimal random forest model is used to carry out reservoir prediction with the optimal attribute as the input, and good results are achieved. It is proved that random forest algorithm can effectively predict reservoir. Parameter optimization of random forest model is very important, which affects the efficiency and final prediction accuracy of the algorithm.

Keywords

Random Forest, Decision Tree, Key Parameters, Reservoir Prediction, Importance Analysis

1. 引言

储层预测已经成为重要的油气勘探手段之一，储层预测技术也在不断发展中。近几年，人工智能和机器学习算法发展迅速，快速渗透到各行各业，包括地球物理勘探领域，尤其是遗传算法、随机森林算法、蚁群算法、支持向量机等机器学习算法的发展和应用，将地震储层预测技术推向了一个全新的历程。

随机森林算法是 2001 年 Breiman 提出来的[1]，随机森林模型是由多颗决策树组合而成。作为高度灵活的一种机器学习算法，随机森林对于回归问题和分类问题有很好的效果。随机森林拥有广泛的应用前景，已广泛应用在市场营销、医疗保健、图像识别、疾病预测等领域[2] [3] [4] [5] [6]。在地球物理勘探领域，随机森林方法也有一定的应用，将该方法应用在分类中，可进行岩性识别和地震相划分；将该方法应用于解决回归问题中，可预测储层物性参数。许多学者开展了随机森林算法在地震储层预测领域的应用。王志宏等通过测井数据对储层岩性进行识别，选取 7 种测井参数作为判别指标。对相关性较高的指标进行因子分析，提取公共因子作为随机森林模型的输入，建立基于因子分析和随机森林的储层岩性判别模型，证明模型判别准确率高，判别岩性与实际岩性具有较好的一致性，为测井资料岩性识别提供了一种新方法[7]。宋建国等建立地震属性与储层特征参数之间的非线性关系，以预测值与实际值之间的均方根误差值为评价标准，分析随机森林回归算法在地震储层预测中的特点。并将方法应用于某陆地工区的自然电位预测和某海上工区的自然伽马预测，取得较好效果[8]。柴明锐等分别利用支持向量机、神经网络、随机森林等多个机器学习方法对西北缘 X723 井百口泉组开展砂砾岩岩屑成分预测、对比和分析，结果表明随机森林的预测效果最好，误差绝对值最低[9]。周雪晴等针对复杂岩性碳酸盐岩储层原有岩性识别方法精度较低、泛化能力不足、结果不稳定等问题，提出基于粗糙集-随机森林算法的复杂岩性识别方法。通过对某区块 502 块岩心数据处理，证明该方法且实现简单，有较强泛化能力，可作为复杂岩性储层岩性识别方法[10]。

本文采用随机森林方法实现测井储层参数与地震数据之间的非线性关系拟合。形成了一套基于随机森林的储层预测技术和流程，分析了随机森林建模过程关键参数。以 SC 某工区储层伽马预测为例，对比了不同关键参数设置预测结果的影响，并给出最优参数设置。同时对参与建模的地震属性重要性进行

分析，优化属性样本集。最后利用最佳随机森林模型，以优化属性样本集为输入开展储层预测，取得良好效果。研究表明，随机森林模型建立关键参数的合理设置对预测结果影响较大，随机森林可以作为储层参数预测的有效工具。

2. 方法原理

2.1. 随机森林算法

2.1.1. 随机森林原理及特征

随机森林的基本单元是决策树，它是通过集成学习的思想将多棵树集成的一种算法。多棵决策树就组成了森林。由于决策树容易过度拟合，而随机森林可以过集成学习来减少过度拟合现象。随机森林可以解决分类和回归的问题。假设当前随机森林解决的是分类问题，随机森林中的每棵决策树都是一个分类器，那么对于一个输入样本，N 棵树会有 N 个分类结果。随机森林集成了所有的决策树分类结果，将分类次数最多的类别指定为最终的输出，这就是随机森林解决分类问题的基本原理。

随机森林的两个重要特征是随机性和集成特征。随机森林中的随机具有两层含义，一是指随机选取训练样本，另一个指随机选取特征子集。这两个随机性的引入对随机森林的分类性能非常重要，由于这种随机性，使得随机森林不容易陷入过拟合，并且具有很好的抗噪能力。随机森林的本质属于机器学习中的集成学习方法。集成学习是通过建立几个模型组合来解决单一预测问题。它的工作原理是生成多个分类器/模型，各自独立地学习和作出预测。最后将这些预测结合成单预测。集成学习通过将多个学习器进行结合，常常可以获得比单一学习器更为显著优越的泛化性能(图 1)。

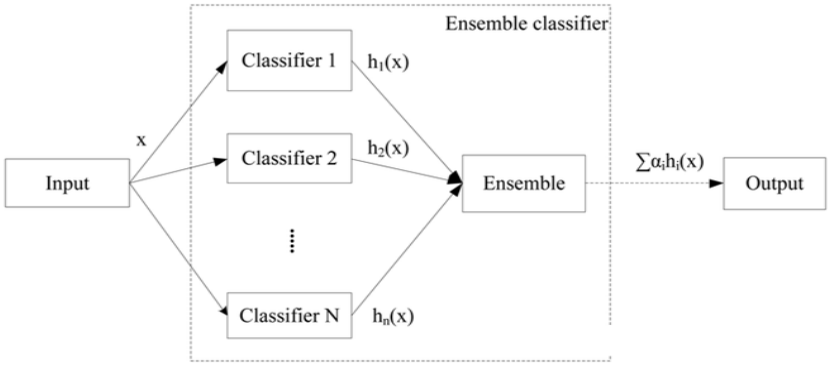


Figure 1. Schematic diagram of ensemble learning
图 1. 集成学习示意图

2.1.2. 随机森林模型构建

随机森林构建思路如下。

1) 构建初始模型

假设随机森林中有 X 个样本，Y 个特征。随机森林的随机性就体现在对于每棵树都随机地抽取 2/3X 个样本作为训练样本集(另外为测试数据集)，同时有放回地随机选取 y ($y \le Y$)个特征作为该树的分支依据。每棵树都据此建立。多棵树构成了随机森林的初始模型。

2) 特征集选择

在 Y 个特征中，每个特征的重要性是不同的，我们需要找到构建树是最重要的特征。判断特征的重要程度，可以通过随机改变特征值，判断特征值改变前后测试数据集的误差来决定，这个误差又叫做该特征的袋外误差(*errOOB*)。每个特征在多颗树中的重要程度的均值作为该特征在随机森林中的重要程度。

$$\mathrm{M}\left(Y_i\right)=\frac{1}{N_i}\sum_{t=1}^{N_i}\left(errOOB_{t1}-errOOB_{t2}\right) \tag{1}$$

上式为特征 Y_i 在随机森林中的重要性计算公式，其中 N_i 为特征 Y_i 在随机森林中出现的次数，$errOOB_{t1}$ 为 t 棵树中特征 Y_i 值改变后的袋外误差，$errOOB_{t2}$ 为 t 棵树中特征 Y_i 的袋外误差。将所有特征按照重要程度排序，去除重要性低的特征，得到新的特征集合，完成随机森林的一次迭代。

3) 优化随机森林

按照步骤(2)不断迭代，逐步淘汰重要性低的特征，利用新的特征集重新生成新的随机森林模型。每生成一次随机森林后，利用随机森林预测的样本结果与真实值进行比较，得到该随机森林的袋外误差。当袋外误差达到最小时的随机森林模型为最优模型[11] [12]。

2.2. 随机森林储层预测

储层预测主要的目的就是通过估算储层参数来达到研究储层空间分布状况的目的，储层参数包括储层厚度、孔隙度、渗透率、饱和度等，这些参数通常可以由测井数据得到，然而测井数据仅能提供井点位置上的储层参数信息，若要预测整个工区三维尺度上的储层参数还需将测井数据与三维地震数据结合起来。地震数据中包含着丰富的岩性、物性、流体等地下储层信息，这些信息可以通过提取不同的地震属性来实现。不同的地震属性突出的储层信息不一样，有的主要反映岩性变化，有些能够反映流体性质[13]。然而单纯使用地震属性进行储层预测存在不确定性和多解性，因此我们可以将地震属性与测井数据结合起来，建立二者之间的联系，以地震属性为基础，将井上储层参数在三维空间内进行有效的扩展，然后依据物理意义明确的特征参数对储层进行解释，从而能够更加全面准确地研究储层的岩性物性等特征在三维空间的分布规律[14] [15]。

由上述分析可知，建立测井数据与地震数据间的准确关系是进行储层预测的关键，而两者间的关系往往是非线性的，因此建立一种能够准确拟合两种数据的非线性表达式是难点。传统的非线性表达方式有非线性回归，多项式，指数公式等。随着人工智能的发展，一些更加精确的计算方法，如支持向量机、神经网络等成功应用于关系表达。随机森林方法具有训练速度快，容易做成并行化方法、抗过拟合能力强、能够处理高维数据等特点，是一种灵活度和运算效率高的机器学习方法。利用随机森林进行储层预测首先利用井点位置的测井参数和地震属性建立随机森林模型，然后利用建立的模型及三维地震属性数据预测整个工区的储层参数[16]。具体实现步骤如下。

1) 层位标定。

层位标定是连接测井数据和地震数据的关键步骤，该步骤是为了得到准确的井震时深关系，因而建立测井数据与地震响应间的对应关系。

2) 数据重采样。

由于测井数据采样率高，地震数据采样率低，若要建立两者间的关系，需在层位标定的基础上对测井数据进行重采样，得到与地震数据时间采样率一致的测井数据。

3) 样本集确定。

样本集是测井数据样本集和地震数据样本集的组合。地震数据样本的选择需从井旁道获取。由于单一地震属性只能突出某一方面信息，采用单一属性数据可能无法达到良好的拟合效果，因此可以提取多个地震属性，通过属性相关性分析剔除相关性较高的属性，得到最终的地震数据样本集。

4) 随机森林模型建立。

将样本集输入训练随机森林模型，训练得到最优模型。模型建立过程中需要设置一些关键参数，这些参数的设置不仅影响最终模型预测的精度，同时对计算效率也有较大影响。下面介绍最重要的几个参

数。①决策树的数量，决策树数量越大，效果越好，但是计算时间也会随之增加。此外当树的数量超过一个临界值之后，算法的效果并不会很显著地变好。②随机森林的最大深度，这个参数可根据样本数量多少，特征多少来确定，当样本数量多，特征也多的情况下，需要限制树的最大深度。③叶子节点最少样本数。叶子是决策树的末端节点。较小的叶子使模型更容易捕捉训练数据中的噪声。实际应用中，可以尝试多种叶子大小种类，以找到最优的那个值。④内部节点再划分所需最小样本数。这个值限制了子树继续划分的条件，如果某节点的样本数少于该值，则不会继续再尝试选择最优特征来进行划分。如果样本量不大，可以不管这个值。如果样本数量非常大，则应设置较大值。

5) 储层参数预测。

根据建立的随机森林模型开展属性重要性分析，剔除重要性较低的属性，得到新的属性集合。利用新的属性集合输入到随机森林模型中开展储层参数预测，能够在不影响预测精度的情况下提高运算效率(图 2)。

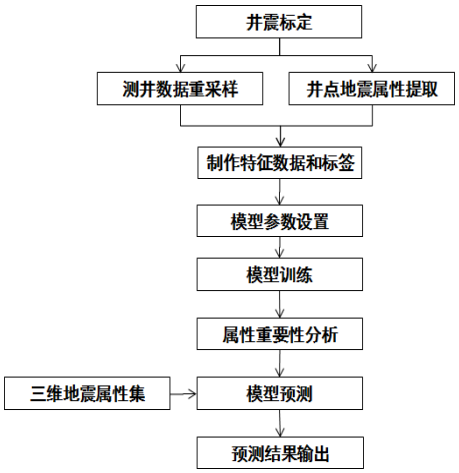


Figure 2. Flow chart of reservoir prediction using Random Forest
图 2. 随机森林储层预测流程图

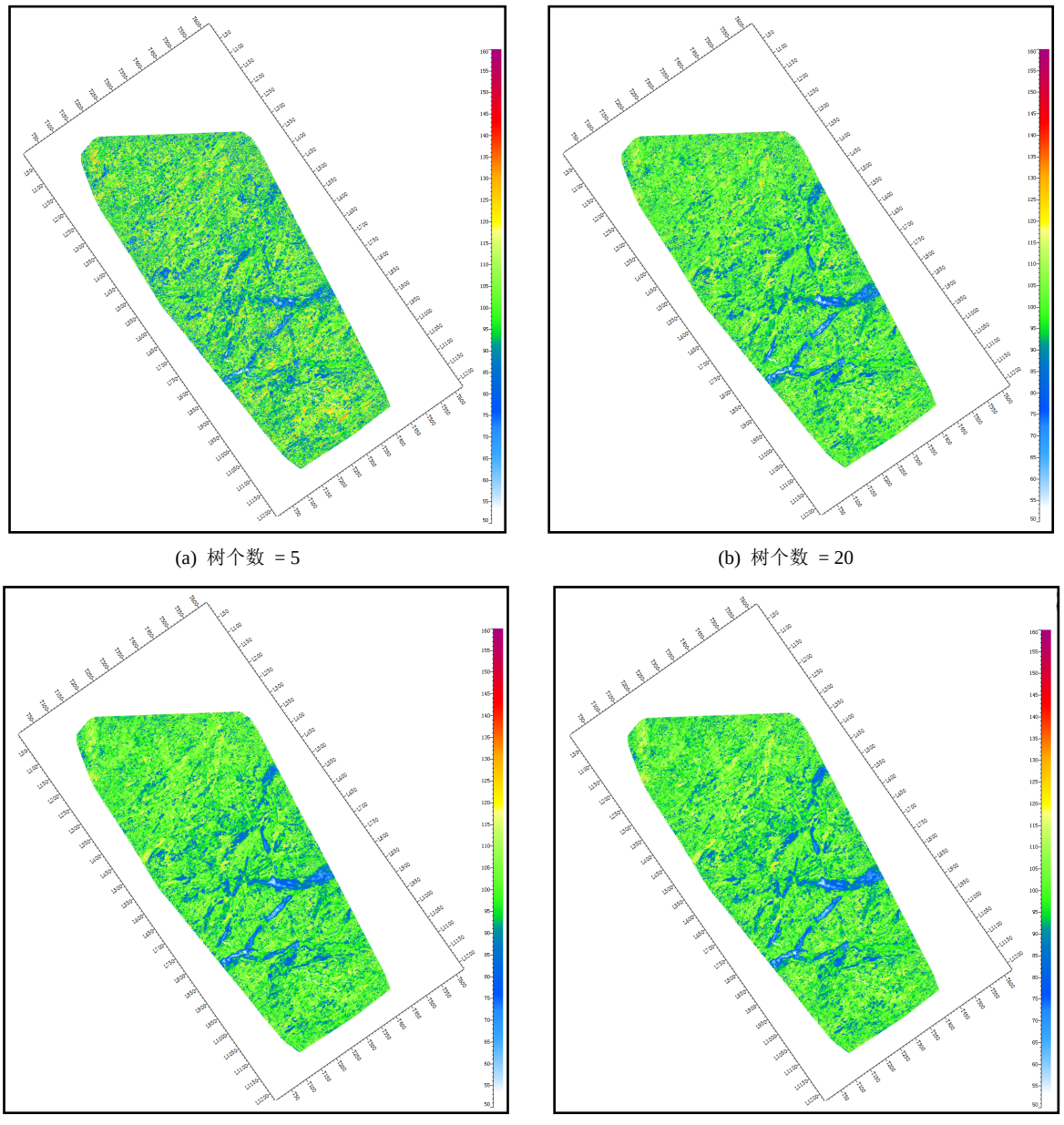
3. 应用实例

以 SC 某工区 SXM 组储层段伽马值预测为例，分别讨论不同参数设置对模型预测结果的影响。在获取样本集时，根据经验提取三瞬属性、反射强度属性、品质因子属性、振幅均值属性、能量半衰时、弧长属等十一种属性，详见表 1。建立井点处测井信息与地震属性信息相对应的样本集。

Table 1. Training sample data set (The first seven lines)
表 1. 数据样本集(前 7 行)

伽马值	弧长属性	反射强度	品质因子	振幅均值	能量半衰时	单频属性 10 HZ	单频属性 40 HZ	单频属性 80 HZ	瞬时振幅	瞬时相位	瞬时频率
61.662	10.444	1289.39	1.171	1862.13	0.58	23,495.9	8972.44	22,000.9	0.0818	−50.803	63.153
65.901	10.697	2704.56	0.871	1099.79	0.58	14,189.3	3972.55	19,097.7	0.2702	−5.333	43.73
52.912	10.824	3894.19	1.247	356.502	0.54	8000.18	13,621.5	12,909.1	0.3507	26.152	39.064
53.621	10.854	4730.82	2.555	−203.152	0.52	7189.98	18,451.2	7832.77	0.2771	54.278	36.708
88.201	10.877	5164.91	2.753	−544.464	0.48	8430.18	19,422.6	7058.02	0.0837	80.707	34.904
85.233	10.906	5180.65	1.168	−762.064	0.56	10,479.1	19,158.4	7778.58	−0.1419	105.838	32.983
89.392	10.981	4794.53	0.605	−1011.3	0.54	13,957.7	18,082.6	9998.2	−0.3066	129.585	29.987

将上述样本集分为训练集和测试集两部分，利用训练样本集建立随机森林模型。为了提高建模精度，需选择最优的参数组合，建立最优模型。下面对随机森林最大深度和决策树个数进行分析。图 3 为不同决策树个数下随机森林预测结果。图 3(a)随机森林中有 5 棵树，图 3(b)中随机森林模型中有 20 棵树，图 3(c)随机森林模型中有 50 颗树，图 3(d)随机森林模型中有 80 颗树。对比可见，低伽马值的砂体预测结果差别较大。随着模型中决策树个数的增多，预测结果越准确，砂岩刻画的越完整。随着决策树数量的进一步增大(图 3(d))，对结果的影响差别不大，即图 3(c)和图 3(d)结果差别不大，此时可判断最佳决策树个数为 50。在实际应用中，随着决策树个数的增加，运算效率也会降低，因此需要综合考虑计算精度和计算效率，找到两者之间的平衡。

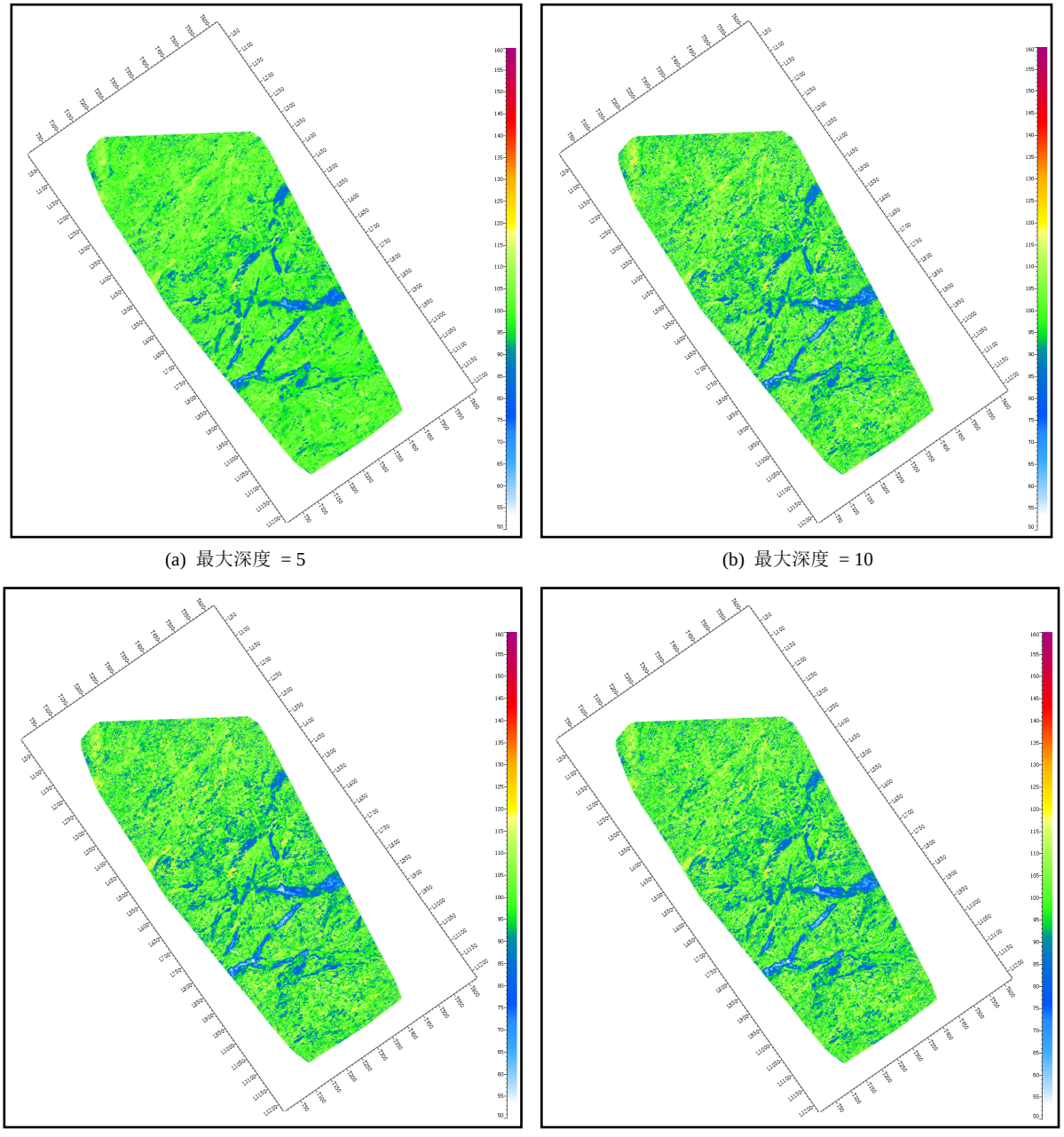


(a) 树个数 = 5　(b) 树个数 = 20

(c) 树个数 = 50　(d) 树个数 = 80

Figure 3. Comparison of different numbers of decision trees
图 3. 不同决策树个数模型预测结果对比

图 4 为随机森林的最大深度对比。当最大深度较浅时，预测结果的精度不高，图 4(a)中最大深度设置过浅，无法得出准确预测结果。当最大深度过深时，随机森林中的每个节点中的样本数量过少，单棵树容易过拟合。对比图 4，本次最大深度最佳设置值为 20。



(a) 最大深度 = 5　(b) 最大深度 = 10

(c) 最大深度 = 20　(d) 最大深度 = 不限制

Figure 4. Comparison of different maximum depth
图 4. 随机森林最大深度预测结果对比

对参与的十一个属性重要性开展分析，分析哪些属性对预测的储层物性参数更加敏感。图 5 为属性重要性对比图，最上面的属性是重要性最高的属性，也是对伽马值最敏感的属性。最下面的属性是最不敏感的属性。所有属性重要性加和为 1。一般来说保留重要性前 0.8 左右的属性数据集，可将重要性最小

的属性忽略。这样既节约时间，对预测结果精度影响不大。因此我们将重要性小于 0.1 的最后三个属性去除，保留重要性较高的前八个属性开展后面的预测。图 6 为属性剔除前后预测结果对比，可以看出，剔除敏感度低的属性对预测结果几乎无影响。

Figure 5. Comparison chart of importance of attributes
图 5. 随机森林最大深度预测结果对比

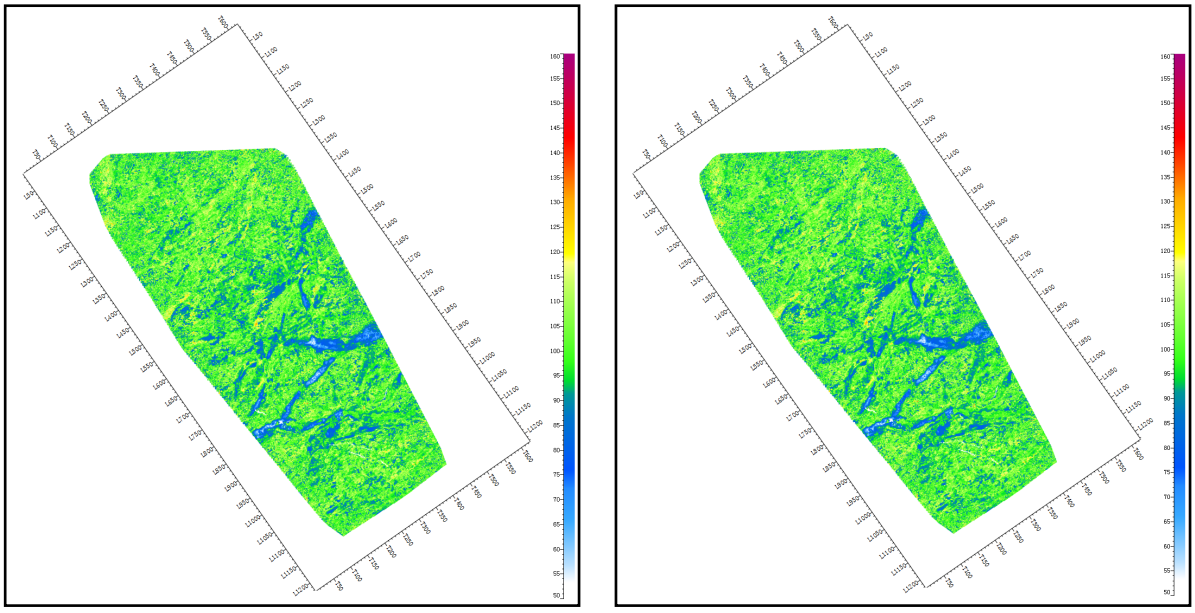


Figure 6. Comparison of results before (left) and after (right) attribute optimization
图 6. 属性优化前(左)，后(右)预测效果对比

4. 结论

本文形成了随机森林储层预测方法及流程，并利用该流程以某工区储层段伽马值预测为例，讨论了随机森林模型构建过程中的关键参数，及最优参数设置。通过分析得到如下结论。

1) 随机森林可以解决回归问题(储层物性参数预测)和分类问题(地震相划分、岩性识别)，实践证明随机森林模型预测结果良好。

2) 随机森林算法可以对特征进行重要性分析，即对参数计算的不同类型的地震属性进行敏感性分析。敏感性高的属性组合与测井参数具有更准确的非线性映射关系。

3) 随机森林模型的参数调优很重要，影响到算法的效率和预测结果精度，是建模过程中的必须步骤。

基金项目

本项技术成果受东方公司基金项目“基于结构张量的属性提取方法研究与应用”(合同号：11-06-2020)资助。

参考文献

[1] Breiman, I. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324

[2] 方匡南, 吴见彬, 朱建平, 等. 随机森林方法研究综述[J]. 统计与信息论坛, 2011, 26(3): 32-38.

[3] Reif, D.M., Motsiniger, A.A., McKinne, B.A., *et al.* (2006) Feature Selection Using a Random Forests Classifier for the Integrated Analysis of Multiple Data Types. CIBCB, 1-8. https://doi.org/10.1109/CIBCB.2006.330987

[4] Diaz-Uriarte, R. and Andres, S.A.D. (2005) Variable Selection from Random Forests: Application to Gene Expression Data. Spanish Bioinformatics Conference.

[5] 张华伟, 王明文, 甘丽新. 基于随机森林的文本分类模型研究[J]. 山东大学学报(理学版), 2006, 41(3): 5-9.

[6] 方匡南. 随机森林组合预测理论及其在金融中的应用[M]. 厦门: 厦门大学出版社, 2012.

[7] 王志宏, 韩璐, 戚磊. 随机森立分类方法在储层岩性识别中的应用[J]. 辽宁工程技术大学学报(自然科学版), 2015, 34(9): 1083-1088.

[8] 宋建国, 高强山, 李哲. 随机森林回归在地震储层预测中的应用[J]. 石油地球物理勘探, 2016, 51(6): 1202-1211.

[9] 柴明锐, 程丹, 张昌民, 朱锐, 唐勇, 瞿建华. 机器学习方法对砂砾岩岩屑成分的预测——以西北缘 X723 井百口泉组为例[J]. 西安石油大学学报(自然科学版), 2017, 32(5): 22-28.

[10] 周雪晴, 张占松, 张超谟, 聂昕, 朱林奇, 张宏悦. 基于粗糙集——随机森林算法的复杂岩性识别[J]. 大庆石油地质与开发, 2017, 36(6): 127-132.

[11] 高强山. 基于随机回归森林的储层预测方法研究[D]: [硕士学位论文]. 青岛: 中国石油大学(华东), 2017.

[12] 何健. 基于随机森林算法的储层预测[D]: [硕士学位论文]. 成都: 成都理工大学, 2020.

[13] Brown, A.R. (2001) Understanding Seismic Attributes. *Geophysics*, **66**, 47-48. https://doi.org/10.1190/1.1444919

[14] Chen, Q. and Sidney, S. (1997) Seismic Attribute Technology for Reservoir Forecasting and Monitoring. *The Leading Edge*, **16**, 445-448. https://doi.org/10.1190/1.1437657

[15] Taner, M. (2001) Seismic Attributes. *CSEG Recorder*, **26**, 8-56.

[16] 魏佳明, 韩家新. 随机森林在储层孔隙度预测中的应用[J]. 智能计算机与应用, 2018, 8(5): 79-82.