



基于改进随机森林算法的地质构造识别模型

王怀秀 冯思怡 刘最亮

引用本文:

王怀秀, 冯思怡, 刘最亮. 基于改进随机森林算法的地质构造识别模型[J]. 煤炭科学技术, 2023, 51(4): 149–156.

WANG Huaixiu, FENG Siyi, LIU Zuiliang. Geological structure recognition model based on improved random forest algorithm[J]. Coal Science and Technology, 2023, 51(4): 149–156.

在线阅读 View online: <https://doi.org/10.13199/j.cnki.cst.2021-0754>

您可能感兴趣的其他文章

Articles you may be interested in

基于地震多属性融合技术的煤层巷道探测识别方法

Coal seam roadway identification method based on seismic multi-attribute fusion technology

煤炭科学技术. 2021, 49(6): 235–241 <http://www.mtkxjs.com.cn/article/id/78b5d837-da58-43c0-b39f-51305af7fc16>

基于C-F方法三维地质模型精度评价研究

Study on accuracy evaluation of 3D geological model based on C-F Method

煤炭科学技术. 2017(11) <http://www.mtkxjs.com.cn/article/id/2c6f6a74-3639-4ef6-9e58-822afab597fe>

智能开采透明工作面地质模型梯级优化试验研究

Experimental study on cascade optimization of geological models in intelligent mining transparency working face

煤炭科学技术. 2020, 48(7) <http://www.mtkxjs.com.cn/article/id/623337c8-793e-4eeb-acf6-95ef8f50260e>

基于改进 OTSU算法的采煤沉陷耕地作物绝产边界识别

Recognition of out-of-production boundary of crops in mining subsidence arable land based on improved OTSU algorithm

煤炭科学技术. 2020, 48(4) <http://www.mtkxjs.com.cn/article/id/d03a7d75-e1f0-46d1-9b95-4412652e7408>

大数据下冻结井筒整体结构模糊随机可靠性模型建立

Fuzzy random reliability model establishment of freezing shaft lining structure under big data environment

煤炭科学技术. 2019(11) <http://www.mtkxjs.com.cn/article/id/1ebd5496-069f-4102-b1a8-338574b84bee>

基于瓦斯地质特征的突出预警模型研究

Study on outburst early warning model based on gas geological features

煤炭科学技术. 2018(12) <http://www.mtkxjs.com.cn/article/id/6ff983b7-e43b-4f9e-be1d-016bce340365>



关注微信公众号, 获得更多资讯信息



移动扫码阅读

王怀秀, 冯思怡, 刘最亮. 基于改进随机森林算法的地质构造识别模型[J]. 煤炭科学技术, 2023, 51(4): 149–156.
WANG Huaixiu, FENG Siyi, LIU Zuiliang. Geological structure recognition model based on improved random forest algorithm[J]. Coal Science and Technology, 2023, 51(4): 149–156.

基于改进随机森林算法的地质构造识别模型

王怀秀¹, 冯思怡¹, 刘最亮²

(1. 北京建筑大学 电气与信息工程学院, 北京 102616; 2. 华阳新材料科技集团有限公司, 山西 阳泉 045000)

摘要:地震属性常常用来进行构造解释以及预测。为克服单一地震属性预测带来的多解性和不确定性的问题, 采用地震多属性融合技术对地质构造进行解释以及预测。基于经典的机器学习随机森林算法模型, 提出了一种改进的随机森林算法对多种地震属性进行融合分类, 将地震多属性融合技术与改进的随机森林算法结合, 建立了基于改进随机森林算法的地质构造识别模型。以山西新元煤炭责任有限公司二条带二采区作为研究区域, 基于三维地震勘探成果提取到的 12 种地震属性, 通过对 12 种属性进行属性相关性分析以及特征重要性分析, 依据结果保留了全部 12 种属性进行后续的属性融合。利用揭露验证后的地质构造—断层和陷落柱作为样本标签, 提出一种改进网格搜索的优化算法, 将分类器数目与单棵决策树的最大特征数组成参数对进行网格搜索, 基于 Python 语言平台建立算法模型, 实验结果表明改进后的算法模型预测准确率达到 97%, 经过后续的模式验证, 证明了相比于逻辑回归、梯度提升与决策树等几种算法, 改进后的随机森林算法能够更加有效地识别地质构造中的断层与陷落柱等异常体, 且识别准确率更高, 算法适用性更加广泛。

关键词:地质构造识别; 地震属性融合; 随机森林算法; 地质模型

中图分类号: TD163

文献标志码: A

文章编号: 0253-2336(2023)04-0149-08

Geological structure recognition model based on improved random forest algorithm

WANG Huaixiu¹, FENG Siyi¹, LIU Zuiliang²

(1. School of electrical and information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 102616, China;

2. Huayang New Material Technology Group Co., Ltd., Yangquan 045000, China)

Abstract: Seismic attributes are often used for structural interpretation and prediction. In order to overcome the problems of multiple solutions and uncertainty caused by single seismic attribute prediction, seismic multi-attribute fusion technology is used to interpret and predict geological structures. Based on the classical machine learning random forest algorithm model, an improved random forest algorithm is proposed to fuse and classify multiple seismic attributes. Combining the seismic multi-attribute fusion technology with the improved random forest algorithm, a geological structure recognition model based on the improved random forest algorithm is established. Taking the second mining area of the second belt of Shanxi Xinyuan Coal Co., Ltd. as the research area, based on the twelve seismic attributes extracted from the three-dimensional seismic exploration results, through the attribute correlation analysis and feature importance analysis of the twelve attributes, according to the results, all twelve attributes are retained for subsequent attribute fusion. Using the exposed and verified geological structure faults and collapse columns as sample labels, an improved grid search optimization algorithm is proposed. The number of classifiers and the maximum feature number of a single decision tree are combined to search the grid. The algorithm model is established based on Python language platform. The experimental results show that the prediction accuracy of the improved algorithm model reaches 97%, After subsequent model verification, it is proved that compared with several algorithms such as logistic regression, gradient lifting and decision tree, the improved random forest algorithm can more effectively identify abnormal bodies such as faults and collapse

收稿日期: 2022-04-15

责任编辑: 周子博

DOI: 10.13199/j.cnki.cst.2021-0754

基金项目: 国家重点研发计划支撑资助项目 (2018YFC0807806); 北京建筑大学研究生创新资助项目

作者简介: 王怀秀(1966—), 女, 安徽歙县人, 副教授, 博士。E-mail: wanghxiu@126.com

通讯作者: 冯思怡(1997—), 女, 山西忻州人, 硕士。E-mail: 1960309474@qq.com

columns in geological structures, with higher recognition accuracy and wider applicability.

Key words: geological structure identification; seismic attribute fusion; random forest algorithm; geological model

0 引言

随着煤炭资源向深部开采,煤层开采难度增大,各种复杂的地质构造严重影响煤矿开采人员安全。地震属性就是经过数学变换而导出的有关地震波的几何形态、运动学特征、动力学特征和统计学特征,通过对地震属性进行分析,并做出标定,消除畸变,就有可能揭示有关储层信息。然而地下地质情况的复杂性和地震信息的影响因素太多,存在较大的不确定性或模糊性,应用任何单一的地震属性都不能准确地进行构造识别,展开地震多属性融合分析就显得十分必要。

地震属性融合的研究有很多种, BALCH^[1] 于 1971 年将地震资料用彩色进行显示,提高了对地下地质异常的识别能力。2002 年,我国乐友喜教授^[2] 优先将聚类分析的方法应用于地震属性融合,多元线性回归法^[3] 也可以用于属性融合(季玉新和欧钦, 2003)。随着大数据时代的来临,目前发展较快的是基于地震属性数据的融合,即通过数学统计、人工智能等方式提取最优地震属性,如 2010 年,曹琳昱^[4] 首次将基于粒子群优化的 BP 网络技术应用于多属性融合中。神经网络融合属性法识别速度很快,并且自适应性以及容错能力强,该方法适用范围广。但这种方法不能自主优选属性,同时需要足够的样本数据来对网络进行训练;2012 年, Bruno 将 PCA^[5] 用于断层属性的融合,通过对地震属性进行 PCA 融合,得到了融合后的新属性,对于微小断层识别的准确度有了较大的提高。但是 PCA 是一种线性降维方法,当数据中存在非线性关系的时候,PCA 的效果会大打折扣;2017 年,孙振宇^[6] 将 SVM 算法用于地震小断层识别, SVM 模型融合各属性预测断层的优势,从不同的角度挖掘断层信息,降低了解释人员主观因素对解释结果的影响。但是在构建 SVM 模型时,模型本身的结构直接影响模型识别准确率,且地震属性的选择对模型准确率影响也很大。

近年来,地震属性融合技术发展迅速,已广泛应用于储层预测^[7-8]、砂体预测^[9-10]、构造识别等各个领域。在地震属性融合过程中,需要解决的关键问题是选择一种准确率高并且适用于多种样本数据集的算法,能够更有效地对地震属性数据进行解释,提高构造识别模型的准确率。

随机森林算法作为一种高度灵活的算法近年来

广受欢迎,拥有广泛的应用前景。在当前所有的算法中,作为一种集成算法的随机森林算法本身精度比大多数单个算法好,准确性高,且对数据集的要求不高,适用于多种数据集(线性与非线性、高维数据集等)。随机森林算法的随机性在于 2 个方面:①每棵树的训练样本是随机的,②树中每个节点的分裂属性集合也是随机选择确定的。正因为这两个随机性,随机森林对噪声数据不敏感,克服了过拟合的问题。但是目前为止,对随机森林中决策树的数量 k 、单棵决策树的最大特征数 m 等参数进行优化与选择的研究还比较少,一般情况下都是通过经验选择参数,往往可能不是最优参数。

针对上述问题,提出一种改进的网格搜索算法,基于模型得分对算法模型进行评估,对随机森林算法的分类器数量与单棵决策树的最大特征数这 2 个参数进行优化,克服以往依据经验选择参数的缺点,选取最优参数值,并且利用得到的算法模型进行地质构造识别与预测。

1 随机森林算法及优化

1.1 随机森林算法原理

2001 年, Breiman 等提出随机森林算法^[11] (Random Forest, 简称 RF)。该算法是一种基于传统的决策树理论的集成学习 (Ensemble Learning) 方法。随机森林在决策树的训练中引入随机属性选择。具体来说,传统决策树在选择划分属性时是在当前结点的属性集合(假定有 d 个属性)中选择一个最优属性;而在随机森林中,对基决策树的每个结点,先从该结点的属性集合中随机选择一个包含 d 个属性的子集,然后再从这个子集中选择一个最优属性用于划分。随机森林通过 Bagging(集成)方法,生成彼此之间互不相同的训练样本集,该算法主要用于分类和回归,对于分类问题,采用简单多数投票法的结果作为随机森林的输出;对于回归问题,根据单棵树输出结果的简单平均作为随机森林的输出^[12]。文中,我们选择随机森林分类算法进行分类预测。

随机森林的算法流程如图 1 所示^[13]。

1) 假设原始训练集有 N 个样本,应用 bootstrap 法随机有放回的抽样组成训练集;

2) 设样本有 D 个特征,在每一棵树的每个节点处随机抽取 $d(d < D)$ 个特征,融合在 d 个特征中选择具有分类能力的变量,变量分类的阈值通过

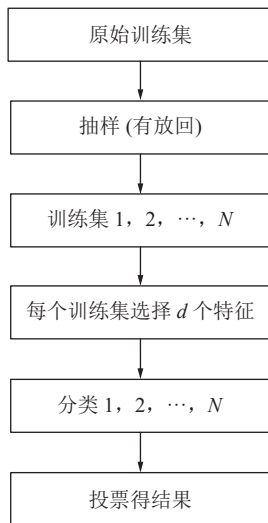


图1 随机森林算法流程
Fig.1 Random forest algorithm flow

检查每一个分类点确定。

3)用抽取的特征进行构造随机森林模型。

4)输入数据,随机森林分类器对新的数据及逆行判别和分类,分类结果按照树分类器的投票多少而定。

随机森林中的每一棵树都是按照自顶向下的递归分裂原则,即从根节点开始依次对训练集进行划分。随机森林在试验中性能较好,由于随机性,使得随机森林不容易过拟合,有很好的抗噪能力并且可以处理很高维度的数据。但是在处理非平衡性数据集的时候,存在缺陷。选择的数据集是某矿区地质勘探得到的地震属性数据,数据标签分布较为均衡,因此该数据集适用于随机森林算法。

1.2 网格搜索

网格搜索(Grid Search CV),又被称作穷举搜索,是目前机器学习中很常用的一种寻优调参的方法。其基本原理是将变量区域网格化,遍历所有网格点,求解满足约束函数的目标函数值,最终比较选择出最优点。在随机森林算法中,每棵树的分类能力越强,整个森林的错误率越低;减小特征选择的个数,树的相关性和分类能力也会相应的降低,增大特征选择的个数,两者也会随之增大。所以关键是如何选择最优的特征个数。目前为止仍然没有单棵决策树的分类正确率和树的多样性两者之间的关系对随机森林性能影响的研究^[14-15]。

基于经典的随机森林算法,提出一种改进的网格搜索优化算法。首先在较大范围内大步长划分网格,进行初步粗搜索选择出最优点;然后在最优点附近进行小步长划分网格,再次进行网格搜索出最优点。

随机森林中有2个重要的参数,分别是决策树数目“n_estimators”和单棵决策树的最大特征数“max_features”,这两个参数能够较好的提升模型对噪声的处理能力,克服过拟合问题^[16-18]。本研究选择这两个参数作为需要搜索的超参数,将二者组成参数对,进行搜索寻优(图2),最终利用网格搜索模型得分作为评估标准得到最优模型以及最佳参数。

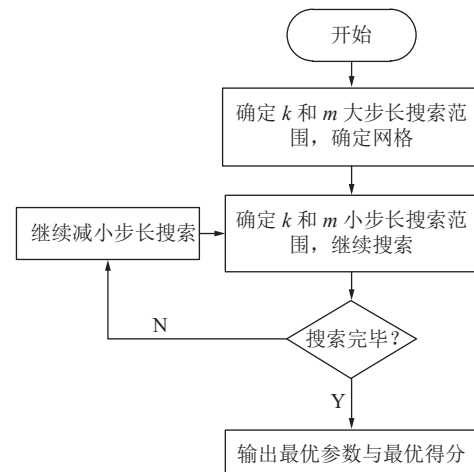


图2 改进网格搜索算法的寻优流程
Fig.2 Optimization process of improved grid search algorithm

具体步骤如下:

1)确定决策树的数量 k 和最大特征数 m 的范围,设定大步长搜索范围。

2)大步长搜索结果确定小步长搜索范围,若输出的最优参数满足要求,则进行下一步,否则,缩小步长,重复上述步骤,继续搜索。

3)对网格节点上的每一组参数构建随机森林,选择得分最优的参数 k, m 。

随机森林通过网格搜索得到最优参数值,最优参数得到的模型可以使用网格搜索参数best_score,即模型的平均交叉验证得分来评估分类效果的好坏,得分越高表明该分类模型的分类效果更好。

2 基于优化随机森林算法的地质构造识别模型

以山西新元煤矿二条带二采区三煤层作为研究区域,该区域的地质异常体主要是断层,还含有较少的陷落柱。断层改变了煤岩层的埋藏条件,使煤层错断并发生显著位移,一方面破坏了煤层的连续性和完整性,为煤层开采带来阻力;另一方面,断层处容易发生瓦斯突水、透水等事故,严重影响矿区的安全开采^[19-20];而陷落柱会破坏煤层的稳定性及连续性,减少煤炭资源储量,同时陷落柱的存在还影响了

工作面的常规布置,给煤矿安全生产带来了重大的不利影响^[21]。前期工程中利用地震勘探等地球物理探测方法得到的地震属性数据信息量非常大。由于不同的属性对相同目标体敏感度不同,针对特定目标体合理选择敏感度较高的属性数据进行分析有利于提高识别精度和结果的准确性^[22]。

研究中利用三维地震勘探成果,按照 5×5 网格提取出研究区域 3 号煤层所对应的 x 、 y 坐标及相关属性数据,基于已有研究成果及专家推荐选取对构造敏感的十二种地震属性,根据矿方提供的实际揭露构造 CAD 图,对该区域内的属性数据进行分类标记,将断层区域标记为 2,陷落柱区域标记为 1,无构造区域标记为 0。经过属性敏感性测试,最后得到 1 397 组包含 x 、 y 坐标以及标记的数据集,而且该数据集包含有 12 种地震属性分别是:方差体切片、相干体切片、分频、均方根振幅、平均能量、倾角、曲

率、瞬时相位、瞬时振幅、瞬时频率、最小振幅和最大振幅)。在算法改进以及模型构建过程中,首先对这 12 种属性进行特征分析,之后进行优化网格搜索的随机森林模型的构建以及模型预测效果的验证。

2.1 特征分析

利用特征相关性分析与特征在随机森林算法分类效果中的影响二者结合对特征进行综合性分析。首先对 12 种属性进行特征相关性分析,得到的属性间相关性见表 1,相关系数越大,两个特征间的相关性越强,特征存在冗余,相关性越小则两个特征间的相关性越弱,当相关系数为 0 时表明两个特征之间是独立的。之后在随机森林算法中,对属性进行特征重要性分析,确定这 12 种属性对于分类器构建以及算法预测的重要性,见表 2 与图 3。依据特征间的相关性分析与特征重要性分析来选择特征。

表 1 特征相关性分析
Table 1 Feature correlation analysis

特征	相关系数										最大振幅	最小振幅	瞬时频率
	特征	方差体	分频	均方根振幅	平均能量	倾角	曲率	瞬时相位	瞬时振幅				
方差体	1												
分频	−0.17	1											
均方根振幅	−0.26	−0.49	1										
平均能量	−0.2	−0.46	0.93	1									
倾角	0.49	−0.065	−0.19	−0.140	1								
曲率	−0.22	−0.15	0.7	0.730	−0.16	1							
瞬时相位	0.083	0.004 4	0.039	0.091	−0.11	−0.15	1						
瞬时振幅	−0.29	−0.44	0.97	0.940	−0.25	0.73	0.048	1					
相干体	−0.46	0.09	0.43	0.400	−0.27	0.48	−0.012	0.46	1				
最大振幅	−0.18	−0.43	0.89	0.980	−0.14	0.7	0.140	0.92	0.380	1			
最小振幅	−0.19	−0.3	0.51	0.470	−0.25	−0.052	0.480	0.51	0.190	0.46	1		
瞬时频率	−0.073	0.14	−0.11	−0.310	0.044	−0.042	−0.27	−0.22	0.075	−0.45	−0.16	1	

表 2 特征重要性
Table 2 Feature importance

属性	特征重要性	属性	特征重要性
分频	0.40	倾角	0.04
方差体	0.12	曲率	0.04
均方根振幅	0.10	最小振幅	0.04
平均能量	0.06	瞬时频率	0.04
瞬时振幅	0.06	瞬时相位	0.03
最大振幅	0.05	相干体	0.03

本数据集中特征相关性分析见表 1,可以看到最大振幅与平均能量、瞬时振幅与均方根振幅之

间相关性较大,而从表 2 和图 3 中可以看到,不同的特征对于分类效果的影响不同,且这 4 种属性间的特征重要性差距不大。经过进一步的算法测试对比试验,发现 4 个特征对本数据集分类效果影响都比较大,而且删掉其中某一个特征后,算法预测的准确率会下降(约下降 3%),且本数据集特征较少,因此选择保留原本 12 个特征进行后续的算法优化。

2.2 算法流程

首先对分类器数量进行大步长搜索,设定随机森林分类树的数量 $n_estimators$ 初始搜索范围为 [50, 1 000],步长设置为 50,设定 $max_features$ 的范围为

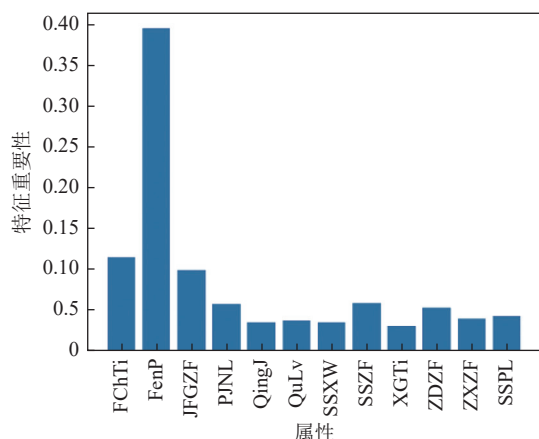


图3 地震属性特征重要性

Fig.3 Importance of seismic attribute feature

[1, 12], 步长为 1。利用 Python 中的模型得分参数对模型进行评估, 大步长搜索过程中模型得分受 2 个参数影响的曲线如图 4 所示。小步长搜索过程中模型得分受 2 个参数影响的曲线如图 5 所示。

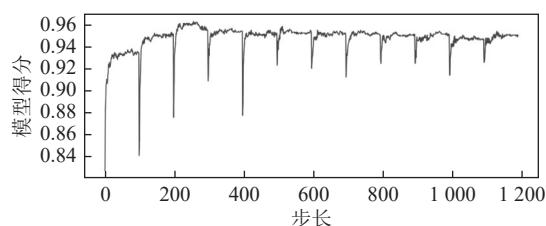


图4 大步长搜索模型得分

Fig.4 Large step search model score

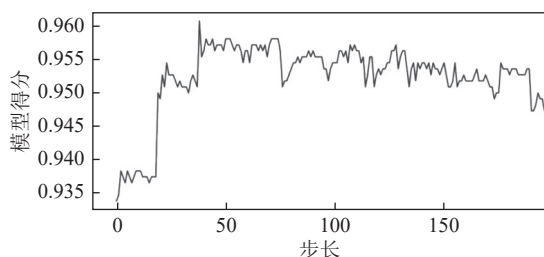


图5 小步长搜索得分

Fig.5 Small step search scores

由图 4 中曲线峰值为 0.960 5, 输出参数值 ('max_features': 3, 'n_estimators': 50)。而且从图 4 中可以看到, 当基分类器的数目超过一定值时, 模型的得分基本收敛, 再增加基分类器的数目, 效果基本不会提升, 而且代码运行速度会变慢。

下一步进行小范围搜索, 初步设定 n_estimators 的范围为 [1, 100]; max_features 的范围为 [1, 12]。得到最优参数对 ('max_features': 3, 'n_estimators': 58)。由于参数数据较多, 因此在本论文中选择最优参数

点附近的 20 组数据作为参考, 见表 3。当分类器数目为 58, 最大特征数为 3 时, 模型得分最高为 0.963 3。

表3 参数对及得分

Table 3 Parameter pairs and scores

max_features	n_estimators	得分
1	56	0.932 95
	57	0.932 06
	58	0.932 94
	59	0.933 84
	56	0.949 04
2	57	0.950 82
	58	0.949 93
	59	0.950 82
	56	0.961 55
	57	0.960 64
3	58	0.963 34
	59	0.960 64
	56	0.955 29
	57	0.955 29
	58	0.954 39
4	59	0.955 29
	56	0.950 82
	57	0.949 92
	58	0.950 82
	59	0.951 72

2.3 算法验证

为进一步验证模型的可靠性, 利用实地采集并且经过处理后的其他几个开采区域的地震属性数据集进行验证。在进行验证时, 为了节省工作, 直接进行大步长(步长设定为 50)网格搜索, 初步得到较优分类器数目, 进行第二步小范围搜索(步长设定为 10), 得到更加精确的分类器数目取值, 最后进行步长为 1 的细化分, 得到最终模型参数以及参数。各数据集在改进网格搜索的随机森林模型中的得分见表 4。可以看到, 经过改进后的随机森林算法模型精确度均有不同程度的提高。

将改进后的随机森林算法与 GBDT(Gradient Boosting Decision Tree, 梯度提升树)、逻辑回归、决策树等 3 种算法在本数据集上的预测结果进行比较, 在比较预测分类效果时, 为了对分类器的评估更全面, 利用正确率、准确率和 f_1 score 等评估指标来评估模型的分类效果。

在训练样本中, 真阳性(True Positive, TP): 指被分类器正确分类的正例数据; 真阴性(True Negative, TN): 指被分类器正确分类的负例数据; 假阳性(False Positive, FP): 被错误地标记为正例数据的负例数据; 假阴性(False Negative, FN): 被错误地标记为负例数据的正例数据。

表 4 随机森林参数优化算法验证

Table 4 Validation of optimization algorithm for random forest parameters

测试区	随机森林算法	改进随机森林算法
1	0.956 81	0.971 88
2	0.926 44	0.935 74
3	0.942 07	0.953 76
4	0.980 85	0.982 55
5	0.803 33	0.813 75
6	0.813 47	0.822 88
7	0.885 11	0.897 56
8	0.760 43	0.782 97
9	0.915 45	0.931 59

针对全体训练样本,正确率(Accuracy, A)计算方式为

$$A = \frac{TP + TN}{TP + FP + TN + FN}$$

针对正例:准确率 P 计算公式为

$$P = \frac{TP}{TP + FP}$$

召回率 R 的计算公式为

$$R = \frac{TP}{TP + FN}$$

f_1 score 被定义为准确率和召回率的调和平均数,用它来综合评估模型性能调和平均数:

$$f_1 = \frac{2PR}{P + R}$$

结果见表 5。可以看到相比于其他算法,改进后的随机森林算法模型预测准确率更高,而且从算法正确率 A 、准确率 P 、 f_1 score 等得分来看,模型的分类效果得到了相应的提高。

表 5 算法对比

Table 5 Algorithm comparison

算法	预测准确率/%	正确率	准确率	f_1 score
逻辑回归	70.71	0.71	0.24	0.28
决策树	93.21	0.95	0.91	0.91
GBDT	96.43	0.96	0.96	0.92
改进算法	97.50	0.97	0.97	0.93

3 模型验证

利用改进后的随机森林算法模型进行构造识别预测。依据矿方提供的新元煤矿二条带二采区实际揭露后的勘探成果,按照 1 : 5 000 比例尺绘制得到

该区域的实际揭露构造 CAD 图,如图 6 所示。构造图中,断层为线(图 6 中紫色线),陷落柱为面(图 6 中红色面区域),蓝色线部分为开采划定的巷道。从图 6 可以看到,该区域中主要构造为断层,陷落柱比较少。

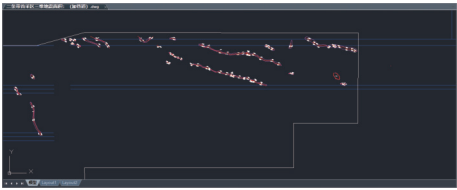


图 6 二条带实际揭露构造 CAD 图

Fig.6 CAD drawing of actual exposed structure of the second belt

研究中将地震属性数据按照 7 : 3 的比例划分为训练集和测试集对模型进行训练以及预测。预测得到地震属性数据坐标点以及标记类型文件,利用课题组开发的软件进行识别结果显示,模型预测呈现出来的图中断层为线,陷落柱为面,得到识别结果如图 7 所示。

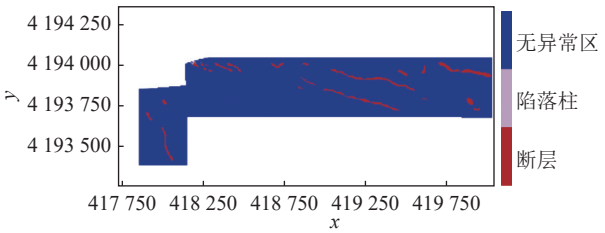


图 7 预测构造

Fig.7 Prediction structure

与该区域勘探得到的三维地震构造 CAD 图(图 6)进行对比,可以看到预测得到的结果中,构造数量预测较为准确,而且预测生成的文本文件中构造对应的坐标点比较精确,可以有效地预测该区域构造所处地理位置,且进行有效识别。

为了进一步验证模型,利用新元北采区西部地震属性数据进行验证,该矿区的三维地震构造如图 8 所示,该验证矿区中陷落柱构造较多,断层较少。得到的构造预测如图 9 所示,改进后的算法模型预测得到的结果与该矿区构造类型基本一致而且通过对生成文本文件中的数据对比观察到坐标基本符合。二条带采区断层构造较多,陷落柱构造较少;而北采区西部矿区陷落柱构造较多,断层构造相对较少。但是通过试验结果,可以看到该算法模型对于断层和陷落柱都有较准确地识别效果。

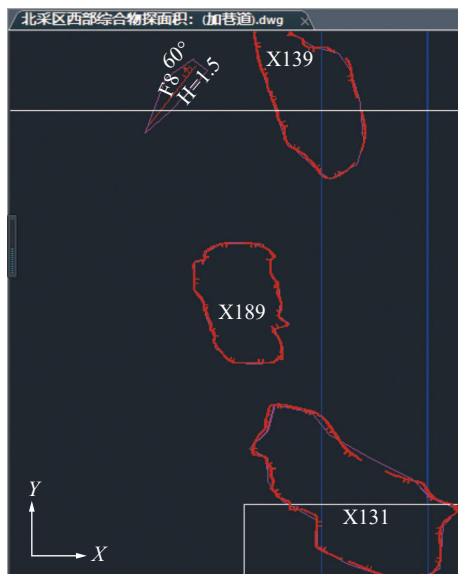


图8 北采区西部实际揭露构造 CAD 图

Fig.8 CAD drawing of actual exposed structure in the west of North Mining Area

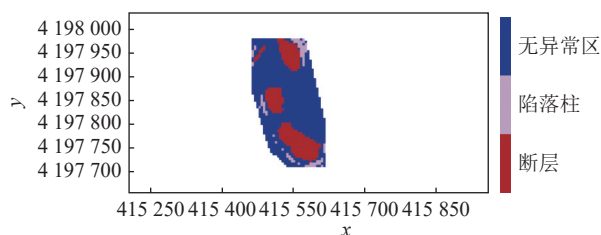


图9 验证矿区构造预测结果

Fig.9 verification of structural prediction results of the mining area

4 结 语

利用山西新元煤矿二条带二采区三维地震勘探成果提取到的地震属性数据集(感谢矿方与勘探方提供的数据),针对现有的研究中对地震属性数据集要求较高且数据处理较复杂的问题,基于经典的随机森林算法模型,提出了一种改进网格搜索优化随机森林算法模型的方法。通过对网格搜索进行分步长搜索,对参数对进行调整,利用改进后的随机森林算法对地震属性数据进行融合分类预测,进而建立地质构造识别模型。将该算法模型与逻辑回归、决策树、GBDT 等几种算法模型进行比较,在模型预测准确率、算法正确率、准确率、 f_1 score 等评估标准进行比较证明该算法优于其他机器学习算法。而且经过在新元北采区西部地震属性数据集上的测试,验证了该算法模型在断层和陷落柱等构造识别中分类结果的准确性与适用性。

参考文献(References):

- [1] A. H. Balch. COLOR SONAGRAMS: A New Dimension In Seismic Data Interpretation[J]. A. H. Balch, 2012, 36(6): 1043-1251.
- [2] 乐友喜,王永刚. 非参数回归法在孔隙度参数预测中的应用[J]. 地质科学, 2002(1): 118-126.
LE Youxi, WANG Yonggang. Application of nonparametric regression method in porosity parameter prediction[J]. *Geological Sciences*, 2002(1): 118-126.
- [3] 季玉新,欧 钦. 优选地震属性预测储层参数方法及应用研究[J]. 石油地球物理勘探, 2003(S1): 57-62,138.
JI Yuxin, OU Qin. Method and application of optimizing seismic attributes to predict reservoir parameters[J]. *Petroleum Geophysical Exploration*, 2003(S1): 57-62,138.
- [4] 曹琳昱,朱仕军,周 强. 基于粒子群优化的BP网络在地震属性融合技术中的应用[J]. 石油与天然气地质, 2010, 31(05): 685-688.
CAO Linyu, ZHU Shijun, ZHOU Qiang. Application of BP network based on Particle Swarm Optimization in seismic attribute fusion technology[J]. *Petroleum and Natural Gas Geology*, 2010, 31(05): 685-688.
- [5] NISHINO KO, NAYAR SHREE K, JEBARA TONY. Clustered blockwise PCA for representing visual data. [J]. *IEEE transactions on pattern analysis and machine intelligence*, 2005, 27(10).
- [6] 孙振宇,彭苏萍,邹冠贵. 基于SVM算法的地震小断层自动识别[J]. 煤炭学报, 2017, 42(11): 2945-2952.
SUN Zhenyu, PENG Suping, ZOU Guangui. Automatic recognition of small seismic faults based on SVM algorithm[J]. *Acta coalae Sinica*, 2017, 42(11): 2945-2952.
- [7] 施尚明,王 杰,段彦清. 基于RGB多地震属性融合的储层预测[J]. 黑龙江科技大学学报, 2016, 26(5): 502-505.
SHI Shangming, WANG Jie, DUAN Yanqing. Reservoir prediction based on RGB multi seismic attribute fusion[J]. *Journal of Heilongjiang University of Science and Technology*, 2016, 26(5): 502-505.
- [8] 朱可丹,王雅春,衣启樊,等. 地震属性融合技术在海拉尔盆地乌东斜坡带南屯组储层预测中的应用[J]. 物探计算技术, 2017, 39(1): 109-115.
ZHU Kedan, WANG Yachun, YI Qifan, et al. Application of seismic attribute fusion technology in reservoir prediction of Nantun Formation in Wudong slope belt of Hailar Basin[J]. *Geophysical and Geochemical Exploration Calculation Technology*, 2017, 39(1): 109-115.
- [9] 杨宏伟,吴海燕,魏国华,等. 基于概率核地震属性融合的砂体预测方法研究[J]. 地球物理学进展, 2020, 35(1): 216-221.
YANG Hongwei, WU Haiyan, WEI Guohua, et al. Study on sand body prediction method based on probabilistic nuclear seismic attribute fusion[J]. *Progress in Geophysics*, 2020, 35(1): 216-221.
- [10] 邱 晗,丛向元,王 虹,等. 地震属性融合技术在松辽盆地扶余油层砂岩预测中的应用[A]//中国石油学会石油物探专业委员会、中国地球物理学会勘探地球物理委员会. 中国石油学会2019年物探技术研讨会论文集[C]. 中国石油学会石油物探专业委员会、中国地球物理学会勘探地球物理委员会: 石油地球

- 物理勘探编辑部, 2019: 4.
- QIU Han, CONG Xiangyuan, WANG Hong, *et al.* Application of seismic attribute fusion technology in sandstone prediction of Fuyu reservoir in Songliao Basin [A]//Petroleum Geophysical Committee of Chinese Petroleum Society, Exploration Geophysical Committee of Chinese Geophysical Society. Proceedings of Geophysical Technology Symposium of Chinese Petroleum Society in 2019 [C]. Exploration geophysical Committee of Chinese Geophysical Society: Editorial Department of Petroleum Geophysical Exploration, 2019: 4.
- [11] BREIMAN L. Random Forests[J]. Machine Learning, 2001, 45: 5–32.
- [12] 李贞贵. 随机森林改进的若干研究[D]. 厦门: 厦门大学, 2013.
- LI Zhengui. Some studies on random forest improvement [D]. Xiamen: Xiamen University, 2013
- [13] 赵金超, 李 仪, 王 冬, 等. 基于优化的随机森林心脏病预测算法[J]. 青岛科技大学学报(自然科学版), 2021, 42(2): 112–118.
- Zhao Jinchao, Li Yi, Wang Dong, Zhang Junhu. Stochastic Forest heart disease prediction algorithm based on optimization[J]. Journal of Qingdao University of Science and Technology (NATURAL SCIENCE EDITION), 2021, 42(2): 112–118.
- [14] MD NASIM ADNAN, MD ZAHIDUL ISLAM. Optimizing the number of trees in a decision forest to discover a subforest with high ensemble accuracy using a genetic algorithm[J]. Knowledge-Based Systems, 2016, 110.
- [15] 温博文, 董文瀚, 解武杰, 等. 基于改进网格搜索算法的随机森林参数优化[J]. 计算机工程与应用, 2018, 54(10): 154–157.
- WEN Bowen, DONG Wenhan, XIE WuJie, *et al.* optimization of random forest parameters based on improved grid search algorithm[J]. Computer Engineering and Application, 2018, 54(10): 154–157.
- [16] LIU Xiao, SONG Mingli, TAO Dacheng, *et al.* Random forest construction with robust semisupervised node splitting. [J]. IEEE Transactions on image Processing: A Publication of the IEEE Signal Processing Society, 2015, 24(1).
- [17] 赵 东, 臧雪柏, 赵宏伟. 基于果蝇优化的随机森林预测方法[J]. 吉林大学学报(工学版), 2017, 47(2): 609–614.
- ZHAO Dong, ZANG Xuebai, ZHAO Hongwei. Random forest prediction method based on Drosophila optimization[J]. Journal of Jilin University (Engineering Edition), 2017, 47(2): 609–614.
- [18] 石文兵, 苏树智. 基于优化随机森林算法的高校餐饮企业营业额预测模型[J]. 通化师范学院学报, 2021, 42(2): 88–94.
- SHI Wenbing, SU Shuzhi. Turnover prediction model of University catering enterprises based on optimized random forest algorithm[J]. Journal of Tonghua Normal University, 2021, 42(2): 88–94.
- [19] 邹冠贵, 任 珂, 吉 寅, 等. 基于主成分分析和最近邻算法的断层识别研究[J]. 煤田地质与勘探, 2021, 49(4): 15–23.
- ZOU guangui, REN Ke, JI Yin, *et al.* Fault recognition based on principal component analysis and nearest neighbor algorithm[J]. Coal geology and Exploration, 2021, 49(4): 15–23.
- [20] 董守华, 石亚丁, 汪 洋. 地震多参数BP神经网络自动识别小断层[J]. 中国矿业大学学报, 1997, 4(3): 16–20.
- DONG Shouhua, SHI Yading, WANG Yang. Automatic identification of small faults by BP artificial neural network with multi parameters[J]. Journal of China University of Mining and Technology, 1997, 4(3): 16–20.
- [21] 刘冬花. 陷落柱对煤矿安全的影响[J]. 内蒙古煤炭经济, 2014, 4(11): 148,165.
- LIU Donghua. Influence of collapse column on coal mine safety[J]. Inner Mongolia Coal Economy, 2014, 4(11): 148,165.
- [22] 庄益明. 煤层小断层地震多属性精细解释方法研究[D]. 徐州: 中国矿业大学, 2018.
- ZHUANG Yiming. Research on seismic multi-attribute fine interpretation method of small fault in coal seam [D]. Xuzhou: China University of Mining and Technology, 2018