Computer Science and Application 计算机科学与应用, 2021, 11(1), 28-35
Published Online January 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.111004

Hans汉斯

自适应特征选择的煤矿突水预测模型研究

陈梦圆，谢天保，童　柔，齐德伟

西安理工大学经济与管理学院，陕西　西安
Email: cmy4399@163.com

收稿日期：2020年12月13日；录用日期：2021年1月6日；发布日期：2021年1月13日

摘　要

针对煤矿突水影响因素复杂难以预测的问题，通过理论分析，构建影响煤矿突水因素的指标体系，并针对收集的相关数据，提出基于遗传算法的自适应预测模型的特征选择算法。实验结果表明：相比于稳定性特征选择算法和递归特征消除算法，自适应特征选择算法可以更好地提高模型预测准确率，适应性更强。

关键词

煤矿突水，影响因素，自适应特征选择，预测模型

Research on Coal Mine Water Inrush Forecasting Model Based on Adaptive Feature Selection

Mengyuan Chen, Tianbao Xie, Rou Tong, Dewei Qi

School of Economics and Management, Xi'an University of Technology, Xi'an Shaanxi
Email: cmy4399@163.com

Received: Dec. 13th, 2020; accepted: Jan. 6th, 2021; published: Jan. 13th, 2021

Abstract

In view of the complex and unpredictable influencing factors of coal mine water inrush, the index system of influencing factors of coal mine water inrush is constructed through theoretical analysis. And according to the relevant data collected, the feature selection algorithm of adaptive prediction model based on genetic algorithm is proposed. The results show that compared with stability feature selection algorithm and recursive feature elimination algorithm, adaptive feature selection algorithm can improve the accuracy of model prediction better and has stronger adaptability.

文章引用: 陈梦圆, 谢天保, 童柔, 齐德伟. 自适应特征选择的煤矿突水预测模型研究[J]. 计算机科学与应用, 2021, 11(1): 28-35. DOI: 10.12677/csa.2021.111004

Keywords

Coal Mine Water Inrush, Influencing Factor, Adaptive Feature Selection, Forecasting Model

1. 引言

煤矿突水是指掘进或采矿过程中当巷道揭穿导水断裂、富水溶洞、积水老窿，大量地下水突然涌入矿井的现象。在富水的岩溶水充水的矿区及顶底板有较厚高压含水层分布的矿山区，或者构造破碎的地段，常易发生矿井突水。据统计，“十一五”期间，全国煤矿发生特别重大水害事故 26 起，平均每年发生 5 起，共死亡 506 人。种种事实数据表明，目前，我国煤炭企业安全生产形势较为严峻，煤矿突水问题已成为制约煤炭工业发展的突出问题之一。

煤矿突水是一种十分复杂的工程地质现象，影响因素很多，且相互影响，这使得煤矿突水和影响因素之间的关系非常复杂，具有非线性，增加了预测难度。传统的煤矿突水预测方法有突水系数法[1] [2]、五图双系数法[3]、脆弱性指数法等[4]，其中突水系数法概念明确、公式简单实用，易于偏于保守，给深部开采带来束缚和限制[5]。脆弱性指数法采用 AHP 法确定主控因素权重，是通过人为判断给出权重值，主观性较强，使得评价结果与实际情况存在一定偏差。

近年来伴随着人工智能、数据挖掘、机器学习技术的不断发展，许多模型和算法例如回归分析(Logistic Regression, LR) [6]、支持向量机(Support Vector Machine, SVM) [7]、Cart 决策树[8]、深度神经网络[9]、PSO-WELM 模型[10]应用于煤矿突水预测，不同程度地提高了煤矿突水预测准确率，但就特定模型和数据，特征分析和选择研究较少。文献[11]建立了 LSTM 神经网络煤矿突水预测模型，并利用类似递归特征消除法的思想进行特征分析和选择，提高了模型预测精度。递归特征消除法并没有搜索特征组合的全局，针对特定模型不能保证特征选择结果最优。

为此本文提出自适应特征选择的煤矿突水预测模型，其主要思想为针对特定模型，利用遗传算法搜索特征组合全局，以五折交叉平均预测准确率为测量标准，自适应选择特征组合以保障该模型的平均预测准确率最高，现实中的煤矿突水样本、以及特征数都不会太大，因此算法不存在计算量大的问题。

2. 煤矿突水影响因素分析

煤矿突水涉及到的因素较多，各因素之间关系复杂，而影响因素之间复杂的相互关系使得煤矿突水与影响因素之间的关系变成复杂的非线性关系。煤矿中不同的地质环境、岩石性质、地应力的大小与分布和不同的断裂构造以及破碎带的发育情况等都有可能诱发不同机理的矿井突水。为此本文通过查阅资料以及煤矿专家的帮助，综合考虑构造条件、含水层条件、开采条件、岩性组合条件四个方面的相关影响因素，初步确定 21 个与煤矿突水有关的影响因素(如表 1 所示)。

3. 基于自适应特征选择的煤矿突水预测模型

表 1 从理论上分析煤矿突水的影响因素。要深入探究这些因素是否真的影响煤矿突水，各自的影响度有多大，就需要收集相关数据，通过构建分析预测模型进行验证。因为煤矿突水的危害性特大，要求模型具有非常高的准确率，特征选择是提高学习模型性能的重要手段，为此构建模型时需要进行特征选择。

Table 1. Related factors and data types of water inrush in coal mines
表 1. 煤矿突水相关因素及数据类型

构造条件	含水层条件	开采条件	岩性组合条件	其他
构造 x1	矿井充水含水层 x9	煤层倾角 x13	砂性岩段 x17	突水征兆 x21
陷落柱 x2	含水层与工作面距离 x10	采面面积 x14	泥性岩段 x18	
陷落柱充水 x3	含水层厚度 x11	走向长度 x15	灰岩段 x19	
断层 x4	含水层水压 x12	采高 x16	煤层厚度 x20	
断层充水 x5				
断层落差 x6				
裂隙带 x7				
裂隙带充水 x8				

3.1. 特征选择

传统的特征选取就是找到原始数据维度中的一个有用子集的过程，再运用一些有效的算法，实现数据的聚类、分类以及检索等任务。依据是否独立于后续的算法模型，特征选择可分为过滤式(Filter)和封装式(Wrapper)两种：Filter，一般直接利用所有训练数据的统计性能评估特征，选择 k 个预测能力较强的特征组成数据集，由于与后续学习算法无关，因此并不保证后续学习算法性能最佳；Wrapper 利用后续学习算法的训练准确率评估特征子集，偏差小，计算量大，不适合大数据集。由于煤矿突水预测模型要求非常高的预测准确度，数据样本量不大，为此，这里的特征选择是指针对收集的煤矿突水样本数据，依据不同的分类算法(例如支持向量机(SVM)、随机森林(RT)、神经网路(Net))从表 1 中的 21 个维度找到一个子集，能够提高分类算法预测准确度。

3.2. 自适应模型特征选择法

针对同一子集算法模型不同，模型预测性能不同。为此结合 Filter 和 Wrapper 各自的特点，本文提出自适应模型特征选择法，其基本思想为首先根据 Filter 法中的互信息理论，计算各特征与样本标签(突水/不突水)的相关度，确定各个特征被选择的概率，然后采用遗传算法进行特征空间搜索，以模型预测准确度作为空间搜索评价准则，最后选择出能使该模型预测性能最好的特征子集。具体算法如下：

1) 计算互信息及数据归一化。由于模型属于分类问题，本文没有采用最大信息系数，而采用互信息，经典的互信息评价定性自变量对定性因变量的相关性的。互信息计算公式：

$$I(x_i;y)=\sum_{x_i\in X}\sum_{y\in Y}p(x_i,y)\log\frac{p(x_i,y)}{p(x_i)p(y)} \tag{1}$$

其中：$x_i\,(i=1\cdots 20)$ 为表 1 中的影响因素，y 为突水标签，然后将 $I(x_i, y)$归一化。

2) 染色体编码。编码采用二进制，每个基因代表一个特征，染色体长度为 21 (对应表 1 中 21 个影响因素)。染色体基因位为 1 表示该特征被选中，否则未被选中。

3) 初始染色体产生。为了提高算法效率，尽快找到最优解。这里采用 Filter 法思想，优先选择互信息大的特征。对于任一染色体基因 Rst(i)，产生随机概率 P，if P < $I(x_i, y)$ then Rst(i) = 1 else Rst(i) = 0。这样在产生的初始染色体中，互信息大的特征被选取的概率较大。染色体种群数目为 N。

4) 自适应交叉算子。为了产生众多的样本数，这里采用三种交叉算子：单点交叉、对半交叉及区间交叉。交叉方式如图 1 所示。

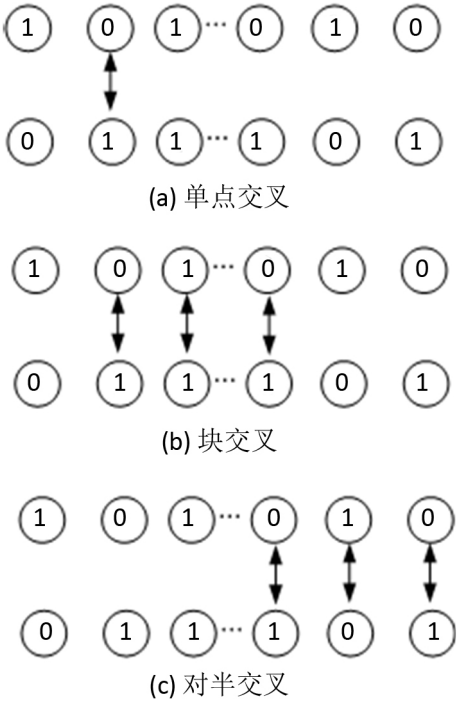

Figure 1. Three kinds of crossover operator
图 1. 三种交叉算子

交叉概率 pc 对遗传算法性能有很大的影响，虽然 pc 较大的时候种群更容易产生新个体，但是当其变大时，优良个体在种群中保留率也降低。这里采用 Srinvivas 提出的自适应遗传算法(Adaptive GA, AGA)方法：

$$pc = \begin{cases} pc_1 - \dfrac{(pc_1 - pc_2)(f' - f_{avg})}{f_{\max} - f_{avg}}, & f' \geq f_{avg} \\ pc_1, & f' < f_{avg} \end{cases} \tag{2}$$

f_{max}：群体中最大的适应度值；

f_{avg}：每代群体的平均适应度值；

f'：要交叉的两个个体中较大的适应度值。

通常情况下，$pc_1 = 0.9$，$pc_2 = 0.6$。

5) 自适应变异算子。变异概率 pm 过小时，不能保证样本多样性；过大时等同于随机算法，失去遗传算法意义，为此这里依然采用 AGA 方法：

$$pm = \begin{cases} pm_1 - \dfrac{(pm_1 - pm_2)(f - f_{avg})}{f_{\max} - f_{avg}}, & f \geq f_{avg} \\ pm_1, & f < f_{avg} \end{cases} \tag{3}$$

f：需要变异的染色体的适应度。

$$pm_1 = 0.1,\ pm_2 = 0.01$$

6) 适应度(S)计算。为了提高算法泛函性，稳定性，这里的适应度考虑模型的 3 个因素：五折交叉后预测准确率平均值 P、所选特征平均互信息 I 以及特征个数 k。

$$S = k1 * P + k2 * I + k3 * k/SN \tag{4}$$

其中： $k1 \gg k2 \gg k3$，SN 为特征总数，即适应度重点考虑模型的预测准确率。

7) 染色体进化。根据种群染色体适应度进行排序，采用轮盘法选择前 N 个染色体为新种群，转(4)。

8) 如连续几代最好染色体种群适应度不再进化，算法结束。

4. 实验

根据表 1 初始确定的特征因素，选取华北地区数据样本 368 例，样本比例基本平衡。为了验证本文提出的特征选择的自适应模型的能力，这里采用支持向量机(Support Vector Machine, SVM)、随机森林(Random Forest, RT)、逻辑回归(Logistic Regression, LR)、梯度提升树(Gradient Boosting Classifier, GBDT)、多层感知机也可称作多层神经网络(Multi-Layer Perceptron, MLP)五种算法进行测试。实验环境为华为笔记本 MateBook X Pro，Win10 + python3.5，五种分类器采用 python 中 sklearn 包，自适应特征选择遗传算法采用 python 自主编程实现。

4.1. 基于互信息的自适应遗传算法

样本特征与样本标签的互信息代表了它们之间的相关性。理论上讲，基于互信息产生的初始染色体相对与随机产生初始染色体相比，其平均适应度大，经过多代进化后，更容易快速获取最优解。

下面选取支持向量机 SVM 作为分类器进行实验，LinearSVC (penalty = ‘l2’, loss = ’squared_hinge’，tol = 0.0001, C = 1.0)，penalty 指定惩罚中使用的规范，loss 指定损失函数，tol 为公差停止标准，C 为错误项的惩罚参数。

图 2 为实验结果，与传统遗传算法相比，基于互信息生成的染色体在初始迭代阶段适应度较高，收敛较快。

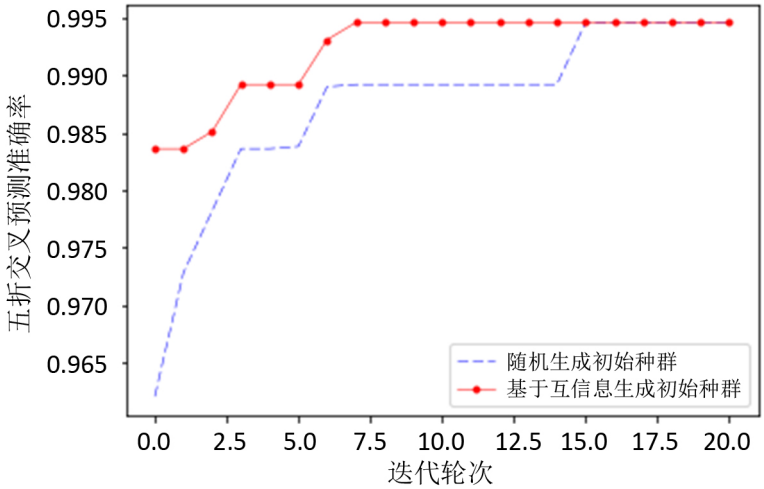

Figure 2. Adaptive genetic algorithm based on mutual information
图 2. 基于互信息的自适应遗传算法

4.2. 自适应特征选择预测模型

正如前文所述，不同的分类器算法有各自的特点，针对不同模型，如何选择样本特征，以提高模型的平均预测准确率(这里采用五折交叉平均值)，本文提出基于互信息生成初始染色体，然后通过自适应交叉(变异)算法，生成新的种群，以模型的平均预测准确率为染色体适应度选择新的种群进行进化，经过多

次迭代最终选择出能够提高模型预测准确率的特征集。为了验证本文算法思想，这里针对目前流行的五种分类器 SVM、RT、LR、GBDT 和 MLP 进行实验，结果如表 2。

Table 2. Test results for five algorithms
表 2. 五种算法测试结果

模型	全特征准确率	自适应选择特征		递归特征消除选择	
		准确率	选择结果	准确率	选择结果
SVM	0.9892	0.9945	x5, x8, x21	0.9945	x5, x21
RT	0.9836	1	x5, x8, x21	0.9945	x5, x12, x21
LR	0.9945	0.9945	x5, x21	0.9945	x5, x21
GBDT	0.9890	0.9945	x5, x21	0.9945	x5, x8, x17, x21
MLP	0.9782	1	x5, x8, x13, x21 x5, x8, x16, x21	不适宜	不适宜

表 2 的实验结果表明，自适应特征选择能够不同程度提高模型的预测精度，对于多层神经网络和随机森林模型，五折交叉均预测准确率为 100%，自适应模型选择的特征数目较少，这是因为初选的特征中，有很多关联性很强，例如 x1 和 x2，x3 相关，x2 和 x3 紧密相关，x4，x5 和 x6 紧密相关，x7 和 x8 紧密相关，就本文实验数据分析结果，煤矿构造条件对华北地区煤矿是否突水影响最大，其次煤矿开采条件(x13, x16)对煤矿突水也有影响。

4.3. 算法对比

特征选择的目的就是最大限度地从原始数据中提取特征以供模型使用，以提高模型的预测精确度，本节就两种顶层特征选择算法(稳定性选择、递归特征消除)与本文算法进行对比分析。

稳定性选择特征是一种基于二次抽样和选择算法相结合的选取方法，可以支持向量机 SVM 或者回归等算法，通过抽取数据子集以及特征子集来运行选择，不断重复，最终可以计算各特征作为重要特征出现的概率，作为特征筛选的依据。本文采用 python 中的 sklearn 包提供的 linear_model 模型进行回归，特征选择计算概率如下表 3。

Table 3. Results of feature selection for stability analysis
表 3. 稳定性分析特征选择结果

特征编号	X5	X21	X8	X2	X3	X7
选择概率	1	0.46	0.33	0.205	0.205	0.05

表 3 的实验结果中，共选择 6 个重要特征。其中 x2 和 x3 紧密相关(如果出现陷落柱充水 x3，那么必然出现陷落柱状况 x2)，类似地，x7 和 x8 紧密相关，这些特征都属于煤矿地质构造条件，再次表明目前的数据的研究结果，煤矿是否突水主要取决于煤矿的构造条件。这与学者杨淑敏观点一致(断裂构造是煤矿突水的主控因素)，稳定性特征选择只是求出各特征的重要性次序，并不确定特征选择的结果。根据图 2 不难分析，这种方法也不能保证模型最终预测结果最优。

递归特征消除的主要思想是反复的构建模型(如 SVM 或者回归模型)然后选出最差的特征进行剔除(可以根据回归系数选择)，然后在剩余的特征集重复这个过程，直到所有特征都遍历了。这个过程中特征被消除的次序就是特征的排序。因此，这是一种寻找最优特征子集的贪心算法。稳定性很大程度上取决于在迭代的时候底层用哪种模型，本文采用 python 中的 sklearn 包对文中四种模型(由于多层神经网络并

不计算特征的相关系数，不适宜该算法)进行了特征选择及模型预测，实验结果如表 2 所示，选择过程如图 3 所示。

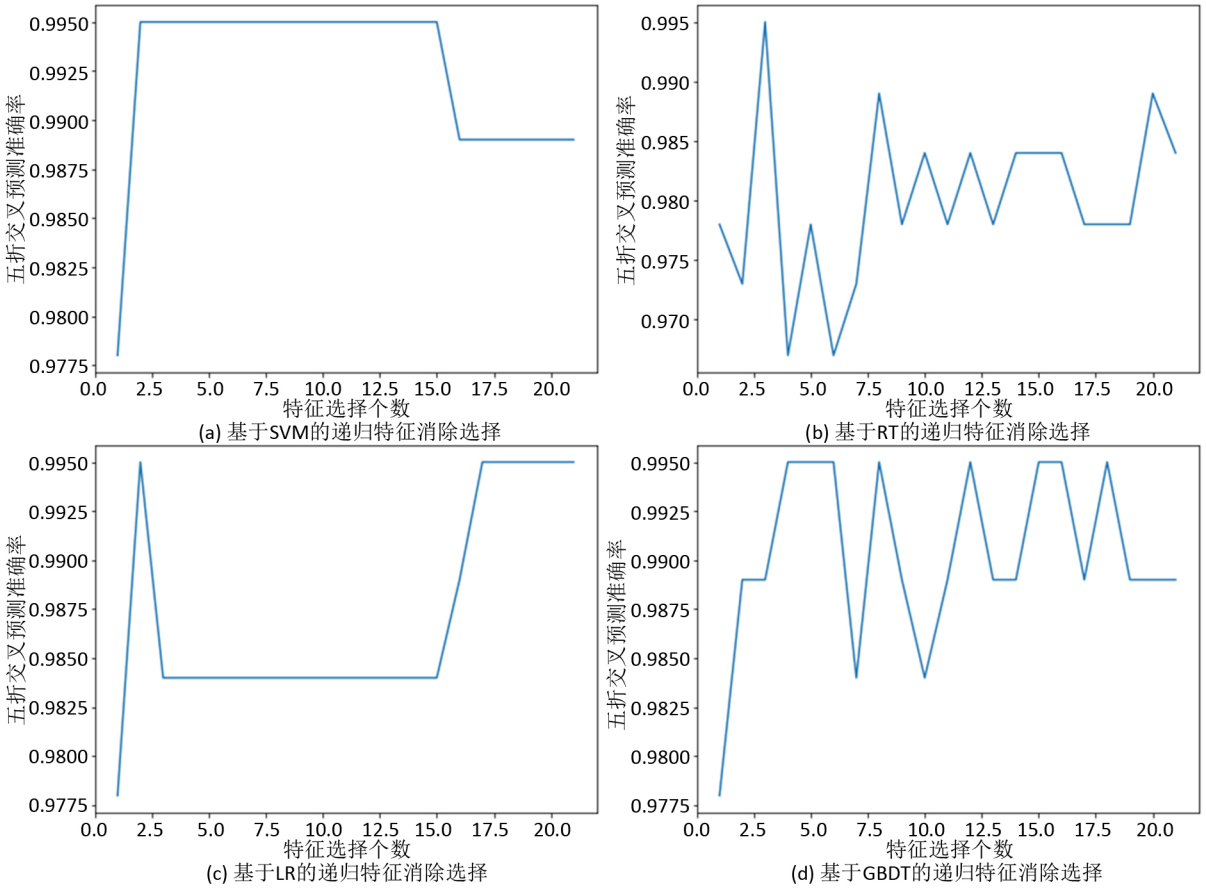

Figure 3. Selection results of recursive feature elimination for four classifiers
图 3. 四种分类器递归特征消除法的选择结果

图 3 和表 2 不难看出，相对于全部特征，递归特征消除法能够不同程度地提高模型的预测准确率(0.994)，但均不如本文算法，这是因为本文算法搜索特征组合的全部空间，而特征递归消除法每次迭代消除一个重要度较低的特征，并没有搜索特征组合全部空间。特征消除时需要考虑特征与样本标签的相关系数，应用范围受限。

5. 结论

尽管本文提出的基于自适应特征选择煤矿突水预测模型准确率较高，但这些样本针对于华北地区的煤矿，并且样本数量受限，目前的研究表明：煤矿突水的影响因素主要来源于构造条件。煤矿突水是由煤矿生产开采过程中各种复杂的因素综合作用的结果，要真正弄清煤矿突水发生的机理，还需要收集更多的样本，以及可视化程度更高的模型算法。

致　谢

感谢西安市科技计划项目对本论文的支持与帮助，感谢本论文所引用各位学者的专著，感谢在撰写本论文时周围的老师、同学的倾心教导，因为有了西安市科技计划项目的支持、各个学者所做研究成果的启发以及周边老师同学的帮助，本篇论文的最终写作才得以顺利完成，再次对他们表示衷心的感谢。

基金项目

本文受 2018 年西安市科技计划项目支持，项目编号：201805037YD15CG21(5)。

参考文献

[1] 段水云. 煤层底板突水系数计算公式的探讨[J]. 水文地质工程地质, 2003, 30(1): 96-99.

[2] 刘德旺. 突水系数法在回坡底煤矿奥灰突水危险性评价中的应用与研究[J]. 中国煤炭, 2016, 42(5): 118-120, 125.

[3] 王宗明, 来永伟, 段俭君. 五图双系数法在北辛窑煤矿底板突水评价中的应用[J]. 煤炭与化工, 2016, 39(5): 7-12.

[4] 陈建平, 李金柱, 王雪冬. 改进脆弱性指数法在煤矿底板突水评价中的应用[J]. 中国地质灾害与防治学报, 2019, 30(3): 67-74.

[5] 毛红川, 刘志, 邓春涛, 王档良. 煤矿突水预测方法探讨[J]. 河北工业科技, 2008, 25(4): 196-199.

[6] 刘伟韬, 廖尚辉, 刘士亮, 等. 主成分 logistic 回归分析在底板突水预测中的应用[J]. 辽宁工程技术大学学报(自然科学版), 2015, 34(8): 905-909.

[7] 闫志刚, 白海波, 张海荣. 一种新型的矿井突水分析与预测的支持向量机模型[J]. 中国安全科学学报, 2008, 18(7): 166-170.

[8] 杜春蕾, 张雪英, 李凤莲. 改进的 CART 算法在煤层底板突水预测中的应用[J]. 工矿自动化, 2014, 40(12): 52-56.

[9] 何风琴. 基于 PSO-WELM 模型的煤矿突水预测研究[J]. 煤炭技术, 2017, 36(10): 124-126.

[10] 曹超凡. 基于深度神经网络的煤矿突水水源判别分析[J]. 无线互联科技, 2017, 3(6): 137-140.

[11] 董丽丽, 费城, 张翔, 曹超凡. 基于 LSTM 神经网络的煤矿突水预测[J]. 煤田地质与勘探, 2019, 47(2): 137-142.