煤炭科学技术
COAL SCIENCE AND TECHNOLOGY

基于PSO-XGBoost的矿井突水水源快速判识模型

董东林 张陇强 张恩雨 傅培祺 陈宇祺 林新栋 李慧哲

引用本文:
董东林，张陇强，张恩雨，等. 基于PSO-XGBoost的矿井突水水源快速判识模型[J]. 煤炭科学技术, 2023, 51(7): 72-82.
DONG Donglin, ZHANG Longqiang, ZHANG Enyu. A rapid identification model of mine water inrush based on PSO-XGBoost[J]. Coal Science and Technology, 2023, 51(7): 72-82.

在线阅读 View online: https://doi.org/10.13199/j.cnki.cst.2023-0446

董东林, 张陇强, 张恩雨, 等. 基于 PSO-XGBoost 的矿井突水水源快速判识模型[J]. 煤炭科学技术, 2023, 51(7): 72−82.
DONG Donglin, ZHANG Longqiang, ZHANG Enyu, *et al*. A rapid identification model of mine water inrush based on PSO-XGBoost[J]. Coal Science and Technology, 2023, 51(7): 72−82.

移动扫码阅读

基于 PSO-XGBoost 的矿井突水水源快速判识模型

董东林，张陇强，张恩雨，傅培祺，陈宇祺，林新栋，李慧哲
（中国矿业大学 (北京) 地球科学与测绘工程学院, 北京 100083）

摘 要: 矿井突水是煤矿安全生产面临的主要威胁之一，快速分析突水成因和准确判别突水水源是矿井突水灾害治理的关键步骤。为有效防治矿井突水灾害，准确快速地判识矿井突水水源，提出一种基于粒子群优化算法(PSO)结合极限梯度提升回归树(XGBoost)的矿井突水水源识别模型(PSO-XGBoost)，通过高效的参数全局搜索模式进一步提高突水水源识别效率与精度，并将该模型成功应用于辽宁抚顺煤田老虎台矿区以验证模型的实用性。基于老虎台矿 40 组水样光谱数据，首先利用多元散射校正、平滑去噪、标准化及主成分分析对原始光谱数据预处理，依据分层随机抽样按照 7∶3 比例进行训练集和测试集划分。其次，初始化粒子个体最优值和全局最优值，利用 PSO 对 XGBoost 算法的 learning_rate、n_estimatiors、max_depth 等 7 项参数进行迭代寻优，构建最优参数组合下的分类识别模型。为进一步研究该模型的优越性，选取平均判识准确率和对数损失值作为评价指标，对比 PSO-XGBoost 模型与 PSO-SVM、PSO-RF 模型的分类识别结果，同时通过 100 次重复交叉验证评价各模型的泛化能力。对比结果表明，XGBoost、PSO-SVM、PSO-RF 和 PSO-XGBoost 模型对测试集数据的平均判识准确率分别为 87.76%、87.56%、91.67% 和 91.67%。对于重复交叉验证，XGBoost、PSO-SVM、PSO-RF 和 PSO-XGBoost 模型的平均准确度分别为 87.76%、87.56%、90.63% 和 93.18%，相应的对数损失平均值分别为 0.545 3、0.546 0、0.562 3 和 0.453 4。综合分析评价指标结果得出，PSO-XGBoost 模型在矿井突水水源识别方面具有更高的判别精度和更好的泛化能力。

关键词: 矿井突水; 水源识别; 粒子群优化算法; XGBoost; 机器学习; 参数优化
中图分类号: TD745 **文献标志码:** A **文章编号:** 0253−2336(2023)07−0072−11

A rapid identification model of mine water inrush based on PSO-XGBoost

DONG Donglin, ZHANG Longqiang, ZHANG Enyu, FU Peiqi, CHEN Yuqi, LIN Xindong, LI Huizhe
(*School of Geosciences & Surveying Engineering, China University of Mining and Technology-Beijing, Beijing* 100083, *China*)

Abstract: Mine water inrush is one of the main threats to mine safety production. Rapid analysis of the cause of water inrush and accurate identification of water inrush source are the key steps of mine water inrush disaster control. In order to effectively prevent and control mine water inrush disaster and identify mine water inrush source accurately and quickly, a mine water inrush source identification model (PSO-XGBoost) based on particle swarm optimization algorithm (PSO) and limit gradient lifting regression tree (XGBoost) was proposed. The efficiency and accuracy of water inrush source identification were further improved by the efficient parameter global search model, and the model was successfully applied to the Laohutai mine in Fushun coal field, Liaoning Province to verify the practicability of the model. Based on the spectral data of 40 groups of water samples from Laohutai mine, the original spectral data were preprocessed by multiple scattering correction, smoothing denoising, standardization and principal component analysis, and the training set and test set were divided according to the ratio of 7∶3 according to stratified random sampling. Secondly, the individual optimal value and the global optimal value of particles are initialized, and PSO is used to iteratively optimize seven parameters of XGBoost algorithm, such as learning_rate, n_estim-

收稿日期: 2023−03−30 **责任编辑:** 黄小雨 **DOI:** 10.13199/j.cnki.cst.2023-0446
基金项目: 国家自然科学基金资助项目(41972255); 国家重点研发计划资助项目(2017YFC0804104); 国家自然科学基金联合基金资助项目(U1710258)
作者简介: 董东林(1969—), 男, 陕西乾县人, 教授, 博士生导师。E-mail: ddl9266@163.com

ators, max_depth, etc., to construct the classification and recognition model under the optimal parameter combination. To further investigate the superiority of the model, the average discrimination accuracy and log loss value were selected as evaluation indexes to compare the classification recognition results of PSO-XGBoost model with PSO-SVM and PSO-RF models, while the generalization ability of each model was evaluated by 100 repetitions of cross-validation. The comparison results showed that the average discrimination accuracies of XGBoost, PSO-SVM, PSO-RF and PSO-XGBoost models for the test set data were 87.76%, 87.56%, 91.67% and 91.67%, respectively. For repeated cross-validation, the average accuracy of XGBoost, PSO-SVM, PSO-RF, and PSO-XGBoost models were 87.76%, 87.56%, 90.63%, and 93.18%, respectively, with corresponding log-loss averages of 0.545 3, 0.546 0, 0.562 3, and 0.453 4, respectively. Comprehensive analysis of evaluation indexes shows that PSO-XGBoost model has higher discrimination accuracy and better generalization ability in mine water inrush source identification.

Key words: mine water inrush；water source identification；particle swarm optimization algorithm；XGBoost；machine Learning；parameter optimization

0 引　　言

煤炭作为我国短时间内不可替代的稳定主体能源，煤炭资源的安全、清洁、高效开采关乎我国能源安全和经济的健康可持续发展[1-3]。然而，我国煤矿水文地质条件异常复杂，煤层在开采过程中水害频发[4]，不仅长期制约着我国煤炭资源的安全开采，还严重威胁着井下矿工的生命安全。矿井发生突水后，快速、准确地判识突水水源是矿井水害治理的关键[5-6]。传统的矿井突水水源判识法主要有水位动态观测[7]、水温、水化学[8-9]、同位素示踪[8-10]及基于GIS[11-12]等方法。地下水化学能真实反应地下水中不同离子的本质特征[13]，众多学者基于水化学参数分析，以不同离子作为主要判别指标，建立相应的突水水源识别模型[14]，为基于水化学判识突水水源奠定了深厚基础。常见的突水水源判识模型大体分为基于数学方法（多元统计[15-16]、模糊数学法[17]、灰色系统法[18]）和结合计算机技术（神经网络[19]、SVM法[20]、可拓识别法[21]）两大类。这些方法在提高矿井突水水源判别准确性的同时，还推动了矿井防治水理论的发展，然而这些方法在应用过程中仍旧存在局限和不足。比如，基于水化学参数的判别模型虽然性能较为稳定，但识别时间长，效率低，较难适应矿井突水的实时监测[22]，不宜在短时间内快速开展矿井突水灾害防治；神经网络具有较强的非线性映射能力、自学习能力和容错性，但对训练样本要求较高，样本选择的合适与否将直接影响到最终的判识结果[23]；模糊数学和灰色数学法较难确定因子权重和隶属度[24]；可拓识别法在处理差异较小的样本数据时容易造成误判。近年来，紫外–可见分光光度法被广泛应用于医学、环境、化工、地质学等诸多领域，但鲜有文献报道利用该技术方法来识别矿井突水水源的应用研究。紫外可见分光技术凭借时间响应快、精度高和抗干扰等特点，为矿井突水水源判识提供新的思路和方法。

光谱数据承载着测试水源的重要特征信息[25]，因此，基于光谱数据的突水水源判别是高维数据的多分类问题。常见分类算法有决策树（DT）、随机森林（RF）、BP 神经网络、支持向量机（SVM）、XGBoost 等。决策树算法易于理解和解释，但如果树深度过大，会导致过拟合问题，此外，决策树算法对于连续性变量处理较为困难[26]。随机森林算法虽然改善了决策树容易过拟合的缺陷[27]，但该算法在解决不平衡数据集分类精度方面仍有待进一步提高。BP 神经网络基于反向误差传播原理，网络训练耗时长，且需多次训练才能获取较为稳定的模型。SVM 分类适合于处理高维模式下的小样本数据[28]，但受参数影响明显，对于噪声和异常值较敏感[29]。XGBoost 算法是一种提升树模型，基于对 GBDT 算法的高效改进，兼具对高维数据集的规模求解和精准分类，并且该算法引入正则项技术避免了过拟合情况，有效提高了模型分类识别的泛化能力。

研究基于辽宁抚顺煤田老虎台矿不同类型水样的 40 组光谱数据，采用 XGBoost 模型进行矿井突水水源识别研究，并利用粒子群优化算法（PSO）对 XGBoost 模型的学习率（learning rate）、随机采样率、最小叶子节点样本权重（min child weight）和树最大深度（max depth）等 7 项参数进行优化，建立 PSO—XGBoost 分类识别模型，以实现对矿井突水水源的精准高效预测，为矿井突水水源识别提供新的思路与方法。

1　PSO-XGBoost 模型基本原理及构建

1.1　XGBoost 算法基本原理

XGBoost 是基于 Gradient Boosting Decision Tree(GBDT) 算法高效改进而来，兼具良好的分类性能和运行速度[30]。相较于 Boosting 库，XGBoost 算法通过对损失函数二阶泰勒展开，并引入正则项以

实现整体最优解，以此来控制模型整体复杂度，从而有效提高了算法的泛化能力。此外，为防止出现过拟合问题，该算法采用同时对特征选择并行处理的方法，使得该算法运行速度更快，且结果更具有可解释性。

假设给定包含n个样本数量的数据集$D=\{(x_i,y_i)|x_i\in R^m,y_i\in R,i=1,2,\cdots,n\}$由$m$个特征组成，共$n$个样本，其中$R^m$和$R$分别为$m$维实数向量数据集和实数集合。

$$\widehat{y_i}=\sum_{k=1}^{K}f_k x_i,f_k\in F \tag{1}$$

式中：f_k为一棵回归树；K为回归树的总数目；F为回归树空间。

目标函数O_{bj}为

$$O_{bj}=\sum_{i=1}^{m}l(y_i,\widehat{y_i})+\sum_{k=1}^{K}\Omega(f_k) \tag{2}$$

式中：l为损失函数，用来衡量分类预测值和真实值之间的误差；$\widehat{y_i}$为分类预测值；y_i为真实值；$\Omega(f_k)$为正则项。

XGBoost 算法采用梯度提升迭代运算，每经过一次迭代过程，将添加新的回归树，则第 t 次迭代运算结果为：

$$\widehat{y_i}^{(t)}=\sum_{j=1}^{t}f_k(x_i)=\widehat{y_i}^{(t-1)}+f_t(x_i) \tag{3}$$

将式(3)代入式(2)，计算出第t次迭代的目标函数表达式为$O_{bj}{}^{(t)}$：

$$O_{bj}{}^{(t)}=\sum_{i=1}^{m}l[(y_i,\widehat{y_i}^{(t-1)}+f_t(x_i)]+\Omega(f_k)+\sigma \tag{4}$$

式中：σ为常数项。

将式(4)二阶泰勒展开，并加入正则项$\Omega(f_k)$防止出现过拟合现象。

$$\begin{cases}O_{bj}{}^{(t)}\cong\sum_{i=1}^{m}[\partial\widehat{y}^{(t-1)}l\left(y_i,\widehat{y_i}^{(t-1)}\right)f_t(x_i)+\\ \frac{1}{2}\partial^2\widehat{y_i}^{(t-1)}l(y_i,\widehat{y_i}^{(t-1)})f_t^2(x_i)]+\Omega(f_k)+\sigma\\ \Omega(f_k)=\gamma T+\frac{1}{2}\lambda\|\omega^2\|\end{cases} \tag{5}$$

式中：γ为叶子树惩罚系数；T为树叶子节点数目；ω为叶子权重；λ为权重惩罚系数；$\|\ \|$为对ω^2进行正则运算。

由于 XGBoost 算法中参数较多，调参过程随机性大，需要通过参数寻优来提高模型的分类预测精度。因此选用处理多参数优化问题效果显著的 PSO 算法优化模型参数，通过减少参数选取的随机性，以此来提高模型的分类预测性能。

1.2　粒子群优化算法(PSO)

粒子群优化算法(PSO)是一种基于模仿鸟群觅食行为构建起来的群智能算法[31-32]。相比于蚁群、遗传和模拟退火等智能优化算法，该算法完善了蚁群算法收敛速度慢和遗传算法、模拟退火算法易陷入局部最优解等缺陷[33]。PSO 算法凭借参数少、结构简单、效率高、容易实现以及能解决非凸问题等优点，已被广泛用来解决支持向量机(SVM)[34]、BP 神经网络[35]、极限学习(ELM)等算法的优化问题，并且优化效果显著。PSO 算法将优化问题的解定义为在有限维度空间内搜索粒子，每个粒子由一个位置矢量和速度矢量组成，所有粒子共同合作择优，通过自身最优值和粒子群的最优值向更好的位置搜索。每个粒子通过适应度函数计算适应值来衡量自身位置的优劣，同时粒子群中所有粒子都追随当前最优粒子在解空间中位置进行搜索。

假设有一个 D 维的搜索空间，群中粒子总数为 m，第i个粒子的位置表示为向量$\boldsymbol{X}_i=[x_{i1},x_{i2},\cdots,x_{iD}]^{\mathrm{T}}$，速度向量表示为$\boldsymbol{V}_i=[V_{i1},V_{i2},\cdots,V_{iD}]^{\mathrm{T}}$，该粒子自身搜索到的最优位置为$\boldsymbol{P}_i=[P_{i1},P_{i2},\cdots,P_{iD}]^{\mathrm{T}}$，整个种群搜索到的最优位置表示为$\boldsymbol{P}_g=[P_{g1},P_{g2},\cdots,P_{gD}]^{\mathrm{T}}$，这里$g$为粒子编号，$g\in(1,2,3,\cdots,m)$。初始化粒子群后 PSO 算法将计算每个粒子的适应值，通过不断更新迭代来搜索最优解。每进行一次迭代，粒子X_i通过个体最优值P_i和群体最优值P_g来更新自身的位置和速度，迭代公式如下：

$$V_i^{k+1}=\omega V_i^k+c_1r_1\left(P_i^k-X_i^k\right)+c_2r_2\left(P_g^k-X_g^k\right) \tag{6}$$

$$X_i^{k+1}=X_i^k+V_i^{k+1}(i=1,2,\cdots,m) \tag{7}$$

式中：k为迭代次数；ω为惯性系数，用来控制算法的收敛和搜索能力；r_1、r_2为 [0,1] 之间的随机数；c_1和c_2分别为加速因子，表示将粒子推向个体最优值P_i和群体最优值P_g的加速项权重。

1.3　PSO-XGBoost 模型构建

基于 XGBoost 原理与 PSO 算法理论，将 PSO 应用于 XGBoost 分类器的参数寻优，构建 PSO-XGBoost 矿井突水水源光谱分类预测模型(图 1)，流程为：

1)对测量得到的光谱数据进行多元散射校正、平滑去噪及标准化预处理；

2)对预处理后的光谱数据进行主成分分析，依照累计贡献率选取若干主成分，根据主成分分析结果重新计算数据集，以此来实现数据集降维；

3)按照 7∶3 的比例划分训练集与测试集，设定适当的适应度函数并初始化粒子个体最优值和全局最优值，对 learning_rate、max_depth 等参数进行

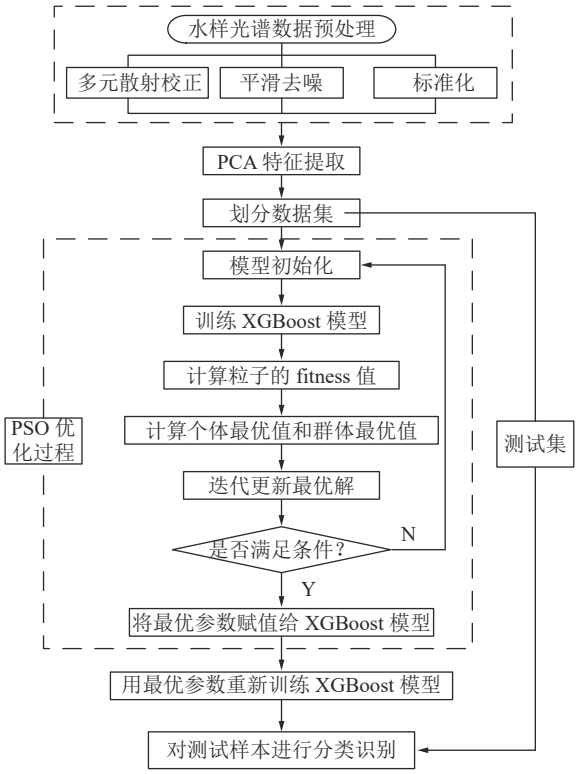


图 1　矿井突水水源识别模型流程

Fig.1　Flow of mine water inrush source identification model

寻优；

4）迭代更新粒子速度与位置，通过计算其适应度值，不断更新个体最优值与全局最优值至达到迭代终止条件；

5）选取最优参数值，构建参数优化的 XGBoost 分类模型，导入训练集进行模型训练学习。

1.4　模型评价指标

研究选择判识正确率和对数损失值对模型的分类判识效果进行评价。其中正确率 A 计算公式为

$$A=\frac{n_{\mathrm{p}}}{n}\times 100\% \tag{8}$$

其中：n_{p}为模型正确分类样本个数；n为样本总个数。模型正确分类的样本数越多，正确率越高，模型分类效果越好。

对数损失值类似于逻辑回归中的损失函数值，通过对错误分类结果的惩罚修正，实现对分类器的分类效果量化，其计算公式为：

$$L(Y,P(Y|X))=-\lg P(Y|X)=-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}y_{ij}\lg(p_{ij}) \tag{9}$$

其中，Y为输出变量；X为输入变量；L为损失函数；N为输入样本量；M为分类类别数；y_{ij}用于表示类别 j 是否是输入实例x_i的真实类别；p_{ij}为模型输入实例x_i属于类别 j 的概率。对数损失值越接近 0，表示损失越小，模型分类效果也就越好。

2　PSO-XGBoost 模型应用

2.1　矿区概况

老虎台煤矿开采于 1907 年，位于辽宁省抚顺煤田中部，井田东西部分别与龙凤矿报废井田和西露天矿井田相邻，南至煤层露头，北至最终开采境界线，面积约为 6.88 km^2（图 2）。矿区整体地势南高北低，南部最高为老虎台山，中部矿区地势平坦，浑河从矿区北部由东向西流过，是矿区水系的主干流。

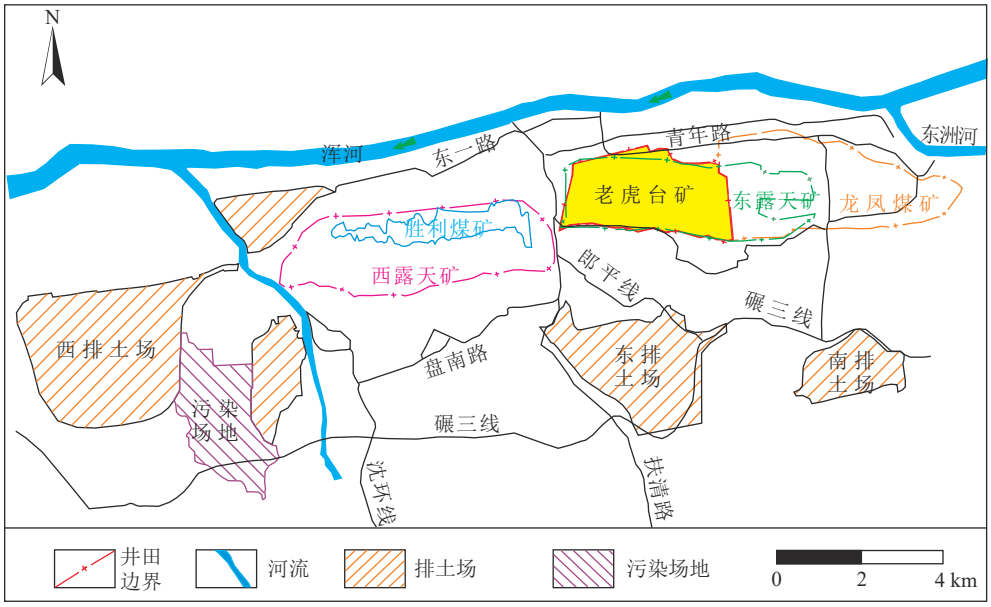


图 2　研究区位置

Fig.2　Location of the study area

矿区共发育 4 个含水层，自上而下依次为：第四系冲积砂及砾石孔隙含水层、古近系西露天组泥灰岩裂隙含水层、古近系栗子沟组和老虎台组凝灰岩、玄武岩弱含水层和白垩系龙凤坎组砂砾岩含水层，其中第四系冲积砂、砾石含水层为主要含水层，其余为弱含水层(图 3)。各含水层水文地质特征如下：

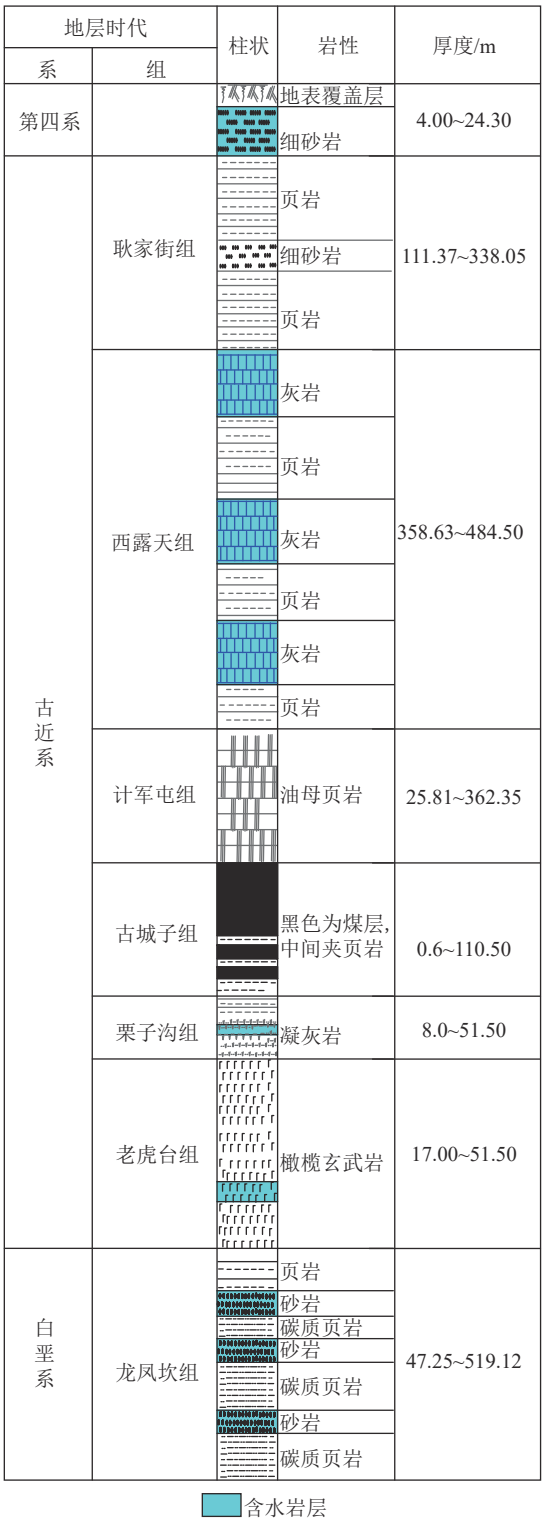


图 3　含隔水层垂向示意

Fig.3　Vertical of aquifer layer

1)第四系冲积砂及砾石孔隙含水层。第四系冲积层位于基岩剥蚀面之上，厚度 4 ~ 24.3 m，由粗细不等的砂卵石组成。含水层上部为黄色亚粘土及砂土覆盖，底部为卵石，单位涌水量 q=0.841 ~ 4.12 L/(s·m)，渗透系数 k=10.27 ~ 92.8 m/d。该层受大气降水补给，富水性强，是矿井主要含水层，水质类型为 HCO_3-Ca-Mg 型。

2)古近系西露天组泥灰岩裂隙含水层。该层全区发育，属泥灰岩绿色页岩系，为绿色含钙质页岩及绿色石灰岩互层，其单位涌水量 q=0.071 L/(s·m)，渗透系数 k=0.065 m/d，水质类型为 SO_4-HCO_3-Ca-Mg 型，属弱矿化淡水。

3)古近系栗子沟组和老虎台组凝灰岩、玄武岩弱含水层。由浅灰绿−暗灰绿色凝灰岩及橄榄玄武岩组成，位于煤层底板以下，弱含水。单位涌水量 q=0.516 ~ 0.000 015 l L/(s·m)，渗透系数 k=1.178 ~ 0.228 m/d。该层在煤层露头接受大气降水和第四纪冲积层孔隙水补给，采掘工作面揭露时，局部表现为滴水和淋水，有的地点为裂隙水。

4)白垩系龙凤坎组砂砾岩含水层。该层分布在井田西部，由不同岩石的角砾组成，岩石主要为花岗片麻岩，石英及玄武岩，微弱含水，其单位涌水量为 0.001 89 ~ 0.002 47 L/(s·m)，渗透系数为 0.000 462 3 ~ 0.007 84 m/d，平均水位标高 76.20 m。

2.2　数据采集与处理

在老虎台煤矿防治水专业人员的指导下，依据矿井突(淋)水特点和以往突水情况，科学选取了地表水体、第四系冲积砂及砾石孔隙含水层、古近系西露天组泥灰岩裂隙含水层和老虎台矿东部龙凤井田老空积水 4 种水样类型。每种类型水单独采集 10 组，总计 40 组水样，水样信息及编号见表 1，对采取的水样进行密封、避光运输和保存。采用 DR-3900 型分光光度计对采集到的水样进行紫外-可见光光谱测量，波长范围设置为 320 ~ 1 100 nm，采样间隔为 1 nm，在测量前用静置 10 min 的超纯水作为参比溶液进行基线校准，最终测得 40 组光谱数据，各水样光谱曲线如图 4 所示。

表 1　水样信息及编号

Table 1　Water sample information and number

编号	水样类型	采集位置
1	老空水	−580 m井下水泵房
2	地表水	老虎台矿西南泵站
3	西露天组泥灰岩裂隙水	−330 m东边界综采面
4	第四系孔隙水	55006东部地表水体

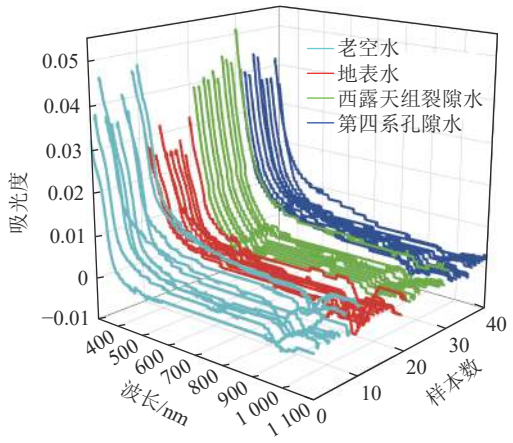


图 4　原始水样光谱曲线

Fig.4　Spectral curve of the original water sample

图 4 显示光谱数据测量波段内的最大值均集中在紫外光区，符合物质对光的波长选择性吸收的原理。由于本次研究集中在紫外–可见光波段，因此只选取 320 ~ 700 nm 的波段测量值作研究数据集。在光谱测量过程中，可能受到光源不稳定性、光路衰减和末端吸收、电路噪声及外界杂散光等多种因素的影响，使得光谱图像受到噪声干扰。对此选用平滑去噪和多元散射校正法对光谱数据进行预处理，以增强光谱数据特征并消除干扰。平滑去噪选用 Savitzky-Golay 卷积平滑法，先从 4 类水样中各取一组样本计算不同窗口宽度和多项式阶数取值下的均方差值（MSE），确定出最佳的 S-G 法参数取值，得到的评价见表 2。综合来看，当窗口宽度取 5，多项式阶数取 3 时有最小的 MSE 值，采用该参数组合对数据集进行去噪处理。

对去噪后的数据集进行多元散射校正处理，减少由于水样散射水平不同导致的光谱差异，将同一类水样的谱线进行聚拢，最终预处理完成得到的水样光谱曲线如图 5 所示。经过光谱预处理后，各水样光谱图像之间的差异得到了显著增强，谱线的灵敏度也有所提高，更易于后续模型的分类识别。

表 2　Savitzky-Golay 卷积平滑评价指标

Table 2　Evaluation indexes of Savitzky-Golay convolution smoothing

窗口宽度		3		5			
多项式阶数		1	2	1	2	3	4
MSE	1水样	2.32×10^{-8}	0	2.93×10^{-8}	1.75×10^{-8}	1.74×10^{-8}	0
	2水样	2.98×10^{-8}	0	4.85×10^{-8}	2.11×10^{-8}	1.95×10^{-8}	0
	3水样	3.97×10^{-8}	0	5.35×10^{-8}	3.01×10^{-8}	3.01×10^{-8}	0
	4水样	3.99×10^{-8}	0	5.32×10^{-8}	3.09×10^{-8}	2.83×10^{-8}	0
MSE均值		3.32×10^{-8}	0	4.61×10^{-8}	2.49×10^{-8}	2.38×10^{-8}	0

注：MSE为均方误差。

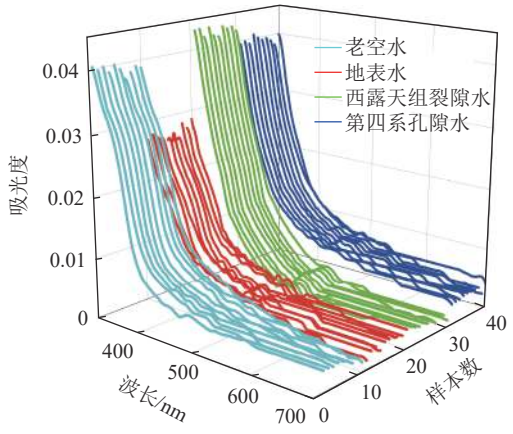


图 5　预处理后的水样光谱曲线

Fig.5　Spectral curve of pretreatment water samples

对预处理后的光谱数据 min-max 标准化处理，并采用主成分分析法（PCA）对其进行特征提取。低维度的数据能加快识别算法的运算速度，同时让数据的特征更明显，提升识别准确度。设置累计贡献率为 90%，获取 5 个主成分，其累计贡献率见表 3，降维处理后的各主成分荷载得分如图 6 所示。主成分分析后的总数据点数从 15 240 个减少到 200 个，数据量减少了 98.69%，极大地简化了数据处理工作量。基于 PCA 得到的特征向量重新计算并划分数据集，考虑到样本数量较少，采用随机划分训练集、测试集的方式不能保证数据分布的一致性，因此采用分层随机抽样的方法按照比例 7∶3 将数据集划分为训练集和测试集。

2.3　水源判别结果

对训练集运用 PSO 算法对 XGBoost 分类器的参数进行寻优，并将适应度判断标准设定为每次迭代中模型经过 15 次交叉验证得到的平均准确率，同时设置输出迭代过程中相应的对数损失值进行辅助判别。本次需要优化的参数有 7 类：包括 learning_

表 3　主成分及其累计贡献率

Table 3　Cumulative contribution of principal components

主成分个数	PCA1	PCA2	PCA3	PCA4	PCA5
累计贡献率/%	56.46	75.79	84.87	88.56	90.78

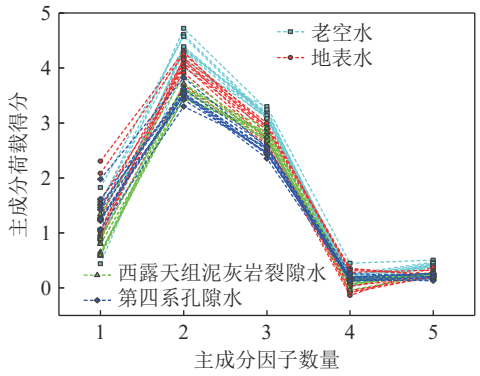
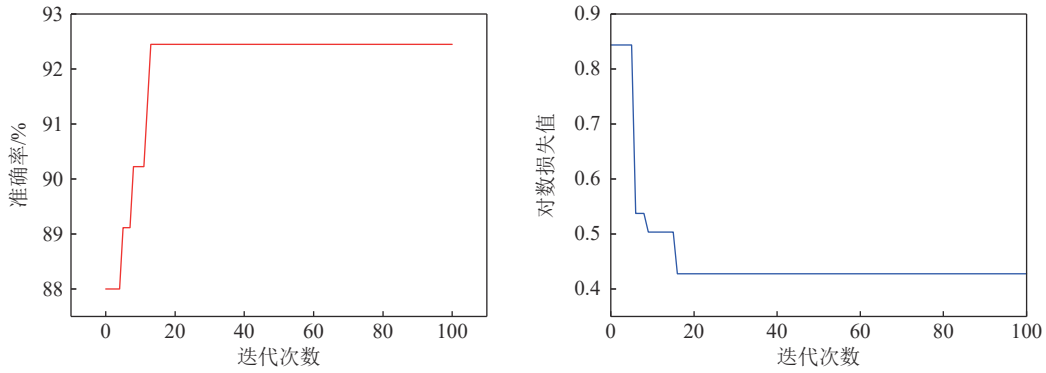


图 6　主成分荷载得分

Fig.6　Plot of principal component load scores

rate、n_estimators、max_depth、min_child_weight、gamma、subsample 和 colsample_bytree，其中 learning_rate 为学习率，控制每次迭代更新权重时的步长；n_estimators 代表树（弱分类器）的数量；max_depth 用于指定树的最大深度；min_child_weight 用于指定叶子节点最小权重和；gamma 值表示节点分裂所需的最小损失函数下降值；subsample 用于控制每棵树随机采样的比例；colsample_bytree 用于控制每棵树随机采样的列数占比。经过 100 次的迭代寻优运算，PSO 算法的准确率迭代至最大值 0.924，相应输出的对数损失值也达到最小值 0.427 6，并且在迭代 18 次时收敛，最优参数迭代寻优过程如图 7 所示，最终寻优得到的最佳参数取值如下：

参数	取值
learning_rate	0.1
n_estimators	100
max_depth	5
min_child_weight	1
gamma	0.01
subsample	0.845 7
colsample_bytree	0.322

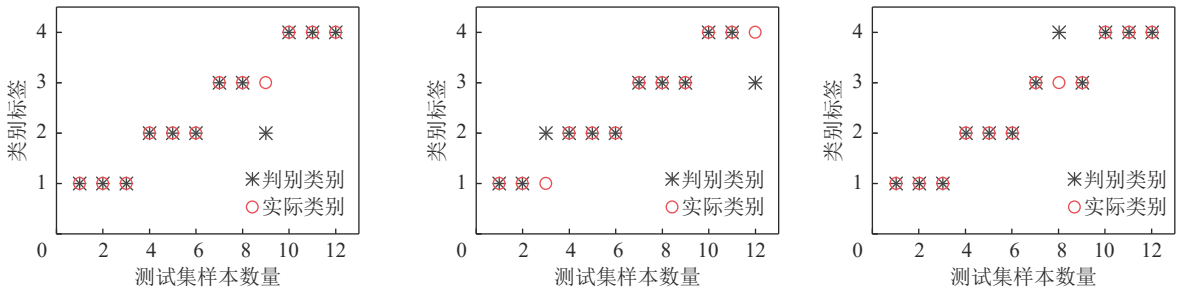


(a) 迭代准确率　(b) 对应准确率下的对数损失值

图 7　最优参数的 PSO 迭代寻优过程

Fig.7　Iterative PSO search process for optimal parameters

基于最优参数组合建立 PSO-XGBoost 分类预测模型，在训练集学习后对测试集进行分类判识，分类结果如图 8a 所示。通过图 8a 得出，PSO-XGBoost 算法分类性能良好，12 组测试集判别正确



(a) PSO-XGBoost　(b) PSO-SVM　(c) PSO-RF

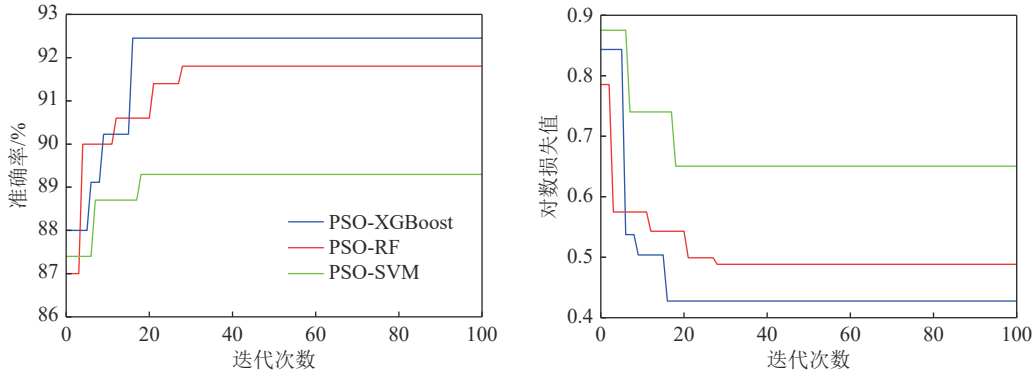
图 8　3 种不同算法测试集判识结果

Fig.8　Three different algorithm test sets to identify the results

11 组，仅在对第九份样本分类时将泥灰岩裂隙水误判为了地表水，模型的分类准确率达到了 91.67%。表明 PSO-XGBoost 模型对于水源光谱曲线有很好的分类性能。

3　不同模型判识结果对比

为进一步验证 PSO-XGBoost 算法模型在突水水源识别研究中的优越性，在测试集上分别选择 PSO 优化后的支持向量机（PSO-SVM）和随机森林（PSO-RF）两种不同分类方法与 PSO-XGBoost 进行对比，各模型的迭代寻优过程如图 9 所示。相较于 PSO-XGBoost 模型的寻优迭代过程，PSO-RF 模型在第 28 次迭代时收敛，最大准确率 91.8%，最小对数损失值 0.488 4；PSO-SVM 模型在第 18 次迭代收敛，最大准确率 89.3%，最小对数损失值 0.650 5。综合判识准确率和对数损失值得出，PSO-XGBoost 的迭代寻优结果最好。其中 PSO-SVM 模型中核函数选择为多项式核 Poly，优化后的参数取值为 C=9.18，degree=6.67；PSO-RF 模型优化参数取值为 n_estimators=129，max_depth=3.86，min_samples_splits=6，min_samples_leaf=6，max_features=0.8。最终 3 组对比算法得到的测试集分类结果如图 8 所示，其中 PSO-SVM 模型判错 2 组，准确率 83.33%，PSO-RF 模型判错一组，准确率和 PSO-XGBoost 模型相同，达到 91.67%。

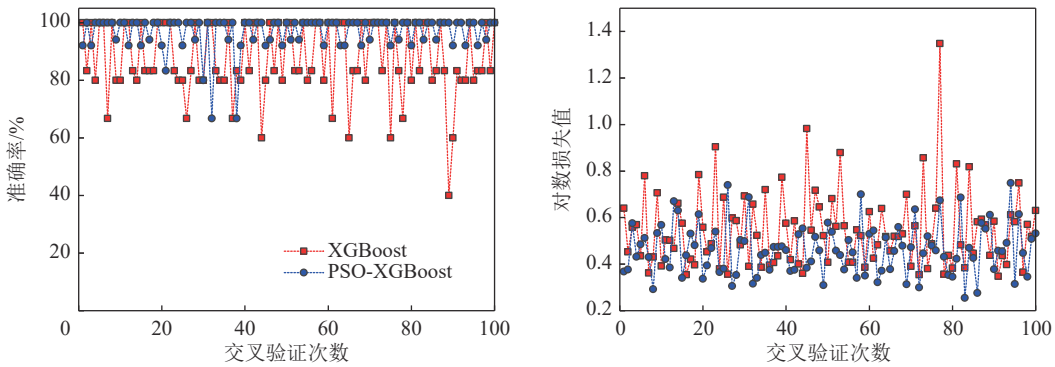


(a) 迭代准确率　　(b) 对应准确率下的对数损失值

图 9　3 种不同算法迭代寻优结果

Fig.9　Iterative optimization search results of three different algorithms

为消除训练集、测试集划分时存在的潜在特殊性和偶然性，同时更客观地评估模型的泛化能力，采用 100 次重复交叉验证的方法按照 8:2 比例的重新划分训练集、测试集，并选择平均准确率和平均对数损失值作为评价指标。首先将未优化的 XGBoost 模型与 PSO-XGBoost 模型进行对比，未优化的 XGBoost 模型参数设为默认参数：n_estimators=300，max_depth=5，learning_rate=0.1，subsample=1，min_child_weight=1，colsampe-bytree=1，对比结果如图 10 所示。可以看出，经过参数寻优后的 PSO-XGBoost 模型平均准确率 93.18%，高于未优化 XGBoost 模型平均准确率 87.76%，PSO-XGBoost 模型的平均对数损失值为 0.462 5，低于 XGBoost 模型的 0.545 3。此外，PSO-XGBoost 模型在判识平均准确率和平均对数损失值上的变化幅度更小，稳定性更好，表明 PSO-XGBoost 判识模型稳定性良好，PSO 优化效果显著。



(a) 准确率对比　　(b) 对数损失值对比

图 10　优化前后准确率对比结果

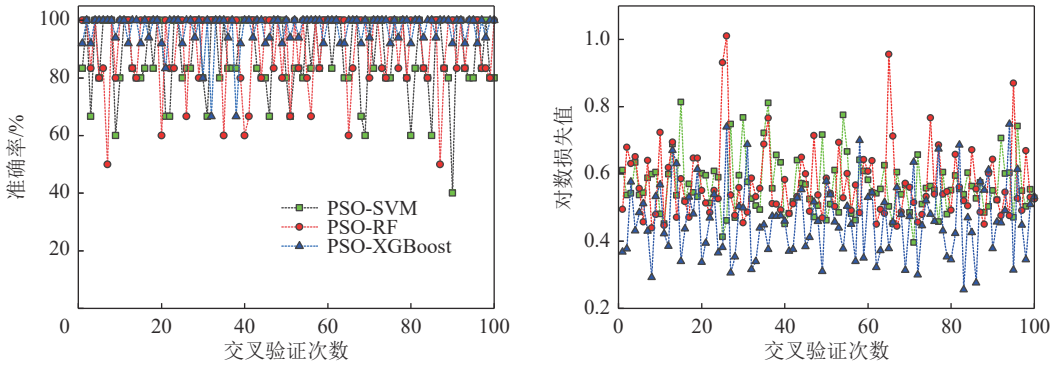
Fig.10　Comparison results of accuracy before and after optimization

进一步对 PSO-XGBoost、PSO-SVM、PSO-RF 3 种预测分类模型的准确性和对数损失值进行比较，对比结果见表 4，对 3 种不同模型进行 100 次重复交叉验证结果如图 11 所示。PSO-SVM 模型平均准确率 87.56%，平均对数损失值为 0.546 0，且稳定性最差；PSO-RF 模型虽然在测试集预测分类中得到了与 PSO-XGBoost 模型相同的准确率，但在重复多次交叉验证评价的平均准确率为 90.63%，平均对数损失值为 0.562 3，综合判识效果低于 PSO-XGBoost 的 93.18% 和 0.453 4，表明 PSO-RF 模型的稳定性不如 PSO-XGBoost 模型。通过对比得出，PSO-XGBoost 模型平均准确率达到三者中的最大值 93.18%，平均损失值也最小，证明该模型不但具有较强的分类预测和泛化能力，而且还具有良好的稳定性。因此，通过对比不同分类算法得出，基于 PSO-XGBoost 的矿井突水水源模型是高效可行的。

表 4　不同分类模型的性能对比

Table 4　Performance comparison of different classification models

算法	交叉验证平均准确率/%	平均对数损失值
PSO-XGBoost	93.18	0.453 4
PSO-SVM	87.56	0.546 0
PSO-RF	90.63	0.562 3



(a) 准确率对比　　(b) 对数损失值对比

图 11　三种算法测试分类准确率比较

Fig.11　Comparison of classification accuracy of s algorithms

4　结　　论

1）基于水样光谱数据，结合 XGBoost 原理与 PSO 理论，构建适用基于光谱数据的矿井突水水源分类预测的 PSO-XGBoost 模型，为矿井突水水源判识提供了新思路。

2）选用 PSO 对 XGBoost 算法参数寻优，选取最优参数组合，模型优化前后的分类准确率分别为 87.76% 和 93.18%，对数损失值分别为 0.545 3 和 0.462 5，PSO-XGBoost 模型相较于优化前的 XGBoost 具有更高的预测精度和更好的稳定性结果，表明 PSO 能显著提升 XGBoost 模型的分类性能。

3）测试结果表明，将 PSO-XGBoost 模型与 PSO-RF 和 PSO-SVM 分类学习模型进行对比，测试集在 PSO-SVM、PSO-RF、PSO-XGBoost 三类模型的平均判识准确率分别为 83.33%、91.67% 和 91.67%，在多次交叉验证测试评价中，3 类模型的平均判识准确率分别为 87.56%、90.63%、93.18%，平均对数损失值分别为 0.546 0、0.562 3、0.453 4。对比结果表明 PSO-XGBoost 模型在分类预测精度和泛化能力方面明显优于其他 2 种模型，因此，基于 PSO-XGBoost 模型判识矿井突水水源的方法是高效可行的。

参考文献(References):

[1] WU Qiang, MU Wenping, XING Yuan, *et al*. Source discrimination of mine water inrush using multiple methods: a case study from the Beiyangzhuang Mine, Northern China[J]. Bulletin of Engineering Geology and the Environment, 2019, 78(1): 469−482.

[2] 吴群英，彭　捷，迟宝锁，等. 神南矿区煤炭绿色开采的水资源监测研究[J]. 煤炭科学技术，2021，49(1)：304−311.
WU Qunying,PENG Jie,CHI Baosuo,*et al*. Researoh on water resources monitoring of green coal mining in Shenan Ming Anea[J]. Coal Science and Technology, 2021, 49(1): 304−311.

[3] 武　强，赵苏启，董书宁，等.《煤矿安全规程》(防治水部分)修改技术要点剖析[J]. 中国煤炭地质，2012，24(7)：34−37.
WU Qiang, ZHAO Suqi, DONG Shuning, *et al*. Analysis of the technical points of the revision of “Coal Mine Safety Regulations” (Part of Water Prevention and Control)[J]. China Coal Geology, 2012, 24(7): 34−37.

[4] 曾一凡，孟世豪，吕　扬，等. 基于矿井安全与生态水资源保护等多目标约束的超前疏放水技术[J]. 煤炭学报，2022，47(8)：3091−3100.

ZENG Yifan, MENG Shihao, LYU Yang, *et al*. Advanced drainage technology based on multi-objective constraint of mine safety and water resources protection[J]. Journal of China Coal Society, 2022, 47(8): 3091−3100.

[5] 曾一凡, 武　强, 杜　鑫, 等. 再论含水层富水性评价的“富水性指数法”[J]. 煤炭学报, 2020, 45(7): 2423−2431.
ZENG Yifan, WU Qiang, DU Xin, *et al*. Further research on “water-richness index method” for evaluation of aquifer water abundance[J]. Journal of China Coal Society, 2020, 45(7): 2423−2431.

[6] 董东林, 孙文洁, 朱兆昌, 等. 基于GIS-BN技术的范各庄矿煤12底板突水态势评价[J]. 煤炭学报, 2012, 37(6): 999−1004.
DONG Donglin, SUN Wenjie, ZHU Zhaochang, *et al*. Water-inrush assessment of coal 12 floor using a GIS-based bayesian network for Fangezhuang Coal Mine with collapse column[J]. Journal of China Coal Society, 2012, 37(6): 999−1004.

[7] 高小伟, 黄　欢. 基于水量水位变化及水化学特征的突水水源判别[J]. 煤炭技术, 2018, 37(9): 198−200.
GAO Xiaowei, HUANG Huan. Discrimination of water bursting source based on change of water quantity and water level and hydrochemical characteristics[J]. Coal Technology, 2018, 37(9): 198−200.

[8] 石　磊, 徐楼英. 基于水化学特征的聚类分析对矿井突水源判别[J]. 煤炭科学技术, 2010, 38(3): 97−100.
SHI Lei, XU Louying. Prediction of mine water inrush sources based on cluster analysis of hydrogeochem ical features[J]. Coal Science and Technology, 2010, 38(3): 97−100.

[9] 连会青, 刘德民, 尹尚先. 水化学综合识别模式在矿井水源判别中的应用[J]. 煤炭工程, 2012(8): 107−109.
LIAN Huiqing, LIU Demin, YIN Shangxian. Application of hydrochemistry comprehensive identification mode to distinguish mine water resources[J]. Coal Engineering, 2012(8): 107−109.

[10] 董东林, 张　健, 林　刚, 等. 矿井涌(突)水源混合水识别模型研究[J]. 煤炭工程, 2020, 52(12): 124−127.
DONG Donglin, ZHANG Jian, LIN Gang, *et al*. Identification model of the source of water-inrush[J]. Coal Engineering, 2020, 52(12): 124−127.

[11] 马　雷, 钱家忠, 赵卫东, 等. 基于GIS的矿井水害防治辅助决策支持系统[J]. 煤田地质与勘探, 2014, 42(5): 44−49.
MA Lei, QIAN Jiazhong, ZHAO Weidong *et al*. GIS-based decision-making support system for prevention and control of water hazards in coal mines[J]. Coal Geology & Exploration, 2014, 42(5): 44−49.

[12] 孙亚军, 杨国勇, 郑　琳. 基于GIS的矿井突水水源判别系统研究[J]. 煤田地质与勘探, 2007, 35(2): 34−37.
SUN Yajun, YANG Guoyong, ZHENG Lin. Distinguishing system study on resource of mine water inrush based on GIS[J]. Coal Geology & Exploration, 2007, 35(2): 34−37.

[13] TERRY Plank, H. LANGMUIR Charles. The chemical composition of subducting sediment and its consequences for the crust and mantle[J]. Chemical Geology, 1997, 145(3).

[14] 曾一凡, 梅傲霜, 武　强, 等. 基于水化学场与水动力场示踪模拟耦合的矿井涌(突)水水源判识[J]. 煤炭学报, 2022, 47(12): 4482−4494.
ZENG Yifan, MEI Aoshuang, WU Qiang, *et al*. Source discrimination of mine water inflow or inrush using hydrochemical field and hydrodynamic field tracer simulation coupling[J]. Journal of China Coal Society, 2022, 47(12): 4482−4494.

[15] 孙文洁, 杨　恒, 徐陈超, 等. 基于多元统计分析的东欢坨矿突水水源判别研究[J]. 煤炭工程, 2021, 53(9): 112−116.
SUN Wenjie, YANG Heng, XU Chenchao, *et al*. Discrimination on mine water inrush source in Donghuantuo mine based on multivariate statistical analysis[J]. Coal Engineering, 2021, 53(9): 112−116.

[16] 黄平华, 陈建生. 基于多元统计分析的矿井突水水源Fisher识别及混合模型[J]. 煤炭学报, 2011, 36(S1): 131−136.
HUANG Pinghua, CHEN Jiansheng. Fisher indentify and mixing model based on multivariate statistical analysis of mine water inrush sources[J]. Journal of China Coal Society, 2011, 36(S1): 131−136.

[17] 李　帅, 王　震, 史继彪, 等. 模糊数学在煤矿突水水源判别中的应用[J]. 煤矿安全, 2012, 43(7): 136−139.
LI Shuai, WANG Zhen, SHI Jibiao, *et al*. Application of Fuzzy Mathematics in Discriminating Sources of Water Inrush in Coal Mine[J]. Coal Mine Safety, 2012, 43(7): 136−139.

[18] 朱赛君, 姜春露, 毕　波, 等. 基于组合权-改进灰色关联度理论的矿井突水水源识别[J]. 煤炭科学技术, 2022, 50(4): 165−172.
ZHU Saijun, JIANG Chunlu, BI Bo, *et al*. Identification of water inrush source based on combined weight-theory of improved gray relational degree[J]. Coal Science and Technology, 2022, 50(4): 165−172.

[19] 王　欣, 葛恒清, 张凯婷, 等. 基于遗传BP神经网络的矿井突水水源识别[J]. 淮阴师范学院学报(自然科学版), 2017, 16(4): 307−311.
WANG Xin, GE Hengqing, ZHANG Kaiting, *et al*. Water source recognition of mine inflow based on the GA-BP neural network[J]. Journal of huaiyin teachers college (Natural science edition), 2017, 16(4): 307−311.

[20] 邵良杉, 李相辰. 基于MIV-PSO-SVM模型的矿井突水水源识别[J]. 煤炭科学技术, 2018, 46(8): 183−190.
SHAO Liangshan, LI Xiangchen. Indentification of mine water inrush source based on MIV-PSO-SVM[J]. Coal Science and Technology, 2018, 46(8): 183−190.

[21] 张瑞钢, 钱家忠, 马　雷, 等. 可拓识别方法在矿井突水水源判别中的应用[J]. 煤炭学报, 2009, 34(1): 33−38.
ZHANG Ruigang, QIAN Jiazhong, MA Lei, *et al*. Application of extension identification method in mine water-bursting source discrimination[J]. Journal of China Coal Society, 2009, 34(1): 33−38.

[22] 王　亚, 周孟然, 闫鹏程, 等. 基于极限学习机的矿井突水水源快速识别模型[J]. 煤炭学报, 2017, 42(9): 2427−2432.
WANG Ya, ZHOU Mengran, YAN Pengcheng, *et al*. A rapid identification model of mine water inrush based on extreme learning machine[J]. Journal of China Coal Society, 2017, 42(9): 2427−2432.

[23] 董东林, 陈昱吟, 倪林根, 等. 基于WOA-ELM算法的矿井突水

水源快速判别模型[J]. 煤炭学报, 2021, 46(3): 984−993.
DONG Donglin, CHEN Yuyin, NI Lingen, *et al*. Fast discriminant model of mine water inrush source based on WOA-ELM algorithm[J]. Journal of China Coal Society, 2021, 46(3): 984−993.

[24] 王甜甜, 靳德武, 刘　基, 等. 动态权-集对分析模型在矿井突水水源识别中的应用[J]. 煤炭学报, 2019, 44(9): 2840−2850.
WANG Tiantian, JIN Dewu, LUI Ji, *et al*. Application of dynamic weight-set pair analysis model in mine water inrush discrimination[J]. Journal of China Coal Society, 2019, 44(9): 2840−2850.

[25] 周孟然, 来文豪, 王　亚, 等. CNN在煤矿突水水源LIF光谱图像识别的应用[J]. 光谱学与光谱分析, 2018, 38(7): 2262−2266.
ZHOU Mengran, LAI Wenhao, WANG Ya, *et al*. Application of CNN in LIF fluorescence spectrum image recognition of mine water inrush[J]. Spectroscopy and Spectral Analysis, 2018, 38(7): 2262−2266.

[26] 范劭博, 张中杰, 黄　健. 决策树剪枝加强的关联规则分类方法[J]. 计算机工程与应用, 2023, 59(5): 87−94.
FAN Shaobo, ZHANG Zhongjie, HUANG Jian. Association rule classification method strengthened by decision tree pruning[J]. Computer Engineering and Applications, 2023, 59(5): 87−94.

[27] 杭　琦, 杨敬辉. 机器学习随机森林算法的应用现状[J]. 电子技术与软件工程, 2018(24): 125−127.
HANG Qi, YANG Jinghui. The current state of application of machine learning random forest algorithms[J]. The Application of Computer Technology, 2018(24): 125−127.

[28] 李继君, 薛　阳, 余桂希, 等. 基于支持向量机的煤矿井水害水源自动识别方法研究[J]. 华北科技学院学报, 2015, 12(2): 25−29.
LI Jijun, XUE Yang, YU Guixi, *et al*. Research on method of automatic recognition of water sources based on Support vector machine[J]. Journal of North China Institute of Science and Technology, 2015, 12(2): 25−29.

[29] 田东雨, 何玉珠, 宋　平. 基于灰狼优化算法的SVM的图像噪声识别[J]. 电子测量技术, 2019, 42(4): 90−94.
TIAN Dongyu, HE Yuzhu, SONG Ping. Approach for image noise recognition by optimizing SVM using grey wolf optimization algorithm[J]. Electronic Measurement Technology, 2019, 42(4): 90−94.

[30] 田　野, 闵锦涛. 基于PSO-XGBoost算法的多衰退特征锂离子电池SOH估计[J]. 电工材料, 2023, 184(1): 23−27.
TIAN Ye, MIN Jintao. SOH Prediction of lithium ion battery with multiple degradation characteristics based on PSO-XGBoost algorithm[J]. Electrotechnical materials, 2023, 184(1): 23−27.

[31] 张　涛, 张明辉, 李清伟, 等. 基于粒子群-支持向量机的时间序列分类诊断模型[J]. 同济大学(自然科学版), 2016, 44(9): 1450−1457.
ZHANG Tao, ZHANG Minghui, LI Qingwei, *et al*. Time series classification diagnosis model based on partical swarm optimization and support vector machine[J]. Journal of TongJi University(Natural Science), 2016, 44(9): 1450−1457.

[32] 周　旭, 朱　毅, 张九零, 等. 基于PSO-XGBoost的煤自燃程度预测研究[J]. 煤矿安全与环保, 2022, 49(6): 79−84.
ZHOU Xu, ZHU Yi, ZHANG Jiuling, *et al*. Study on prediction model of coal spontaneous combustion based on PSO-XGBoost[J]. Mining Safety Environmental Protection, 2022, 49(6): 79−84.

[33] 王雪阳, 史攀飞. 蚁群算法与模拟退火、遗传算法比较分析[J]. 无线互联网科技, 2015, 65(13): 126−127.
WANG Xueyang, SHI Panfei. Comparative analysis of ant colony algorithm with simulated annealing and genetic algorithm[J]. Wireless Internet Technology, 2015, 65(13): 126−127.

[34] 王海军, 乔　烨. PSO-SVM模型在葡萄酒品质分类中的应用研究[J]. 计算机与数字工程, 2012, 40(8): 146−148.
WANG Haijun, QIAO Ye. Wine Quality Classification model based on PSO-SVM[J]. Computer And Digital Engineering, 2012, 40(8): 146−148.

[35] 李祚泳, 汪嘉杨, 郭　淳. PSO算法优化BP网络的新方法及仿真实验[J]. 电子学报, 2008(11): 2224−2228.
LI Zuoyong, WANG Jiayang, GUO Chun. A new method of BP network optimized based on particle swarm optimization and simulation test[J]. Acta Electronica Sinica, 2008(11): 2224−2228.