DOI:10.19651/j.cnki.emt.2210581

# 基于 GWO-XGBoost 泥石流灾害预测*

王智勇 李丽敏 温宗周 尚艳芳 王莲霞
(西安工程大学电子信息学院 西安 710600)

**摘 要:** 针对引发泥石流灾害的致灾因子复杂多样而造成模型输入数据维度过大和极端梯度提升树容易陷入局部最优导致预测模型准确率不高的问题,提出一种基于 GWO-XGBoost 算法模型的泥石流灾害预测方法。首先,对传感器采集到的原始数据进行预处理,得到规范数据,然后通过线性判别分析法进行数据降维得到耦合性低且贡献率较高的致灾因子作为模型输入,对泥石流灾害是否发生进行预测;其次使用灰狼优化算法对模型超参数进行寻优;最后以磨子沟监测数据进行仿真验证。结果表明:经过预处理和线性判别分析法降维后的规范数据解决了模型输入的维数灾难问题,GWO-XGBoost 泥石流灾害预测模型的预测准确率为 96.64%,相较于随机森林模型、支持向量机模型和极端梯度提升树模型的预测准确率分别提高了 6.69%,5.13%和 3.86%,丰富了泥石流灾害预测方法并为相关决策部门提供了全新的思路。

**关键词:** 泥石流;预测模型;线性判别分析法;极端梯度提升决策树;灰狼优化算法
**中图分类号:** P642.23 **文献标识码:** A **国家标准学科分类代码:** 620.1030

# Debris flow disaster prediction based on GWO-XGBoost model

Wang Zhiyong Li Limin Wen Zongzhou Shang Yanfang Wang Lianxia
(School of Electronic Information, Xi'an Engineering University, Xi'an 710600, China)

**Abstract:** In view of the complexity and diversity of the disaster causing factors that cause debris flow disasters, resulting in the excessive dimension of the model input data and the problem that extreme gradient boosting is easy to fall into local optimization, resulting in the low accuracy of the prediction model. A debris flow disaster prediction method based on GWO-XGBoost model is proposed. First, the original data collected by the sensor is preprocessed to obtain the standard data, and then the dimension of the data is reduced by linear discriminant analysis, and the disaster causing factors with low coupling and high contribution rate are obtained as the model input to predict whether the debris flow disaster occurs; Secondly, grey wolf optimizer is used to optimize the super parameters of the model; Finally, Mozigou monitoring data are used for simulation verification. The results show that the normalized data after preprocessing and dimensionality reduction by linear discriminant analysis algorithm solves the problem of dimensionality disaster of model input. The prediction accuracy of GWO-XGBoost debris flow disaster prediction model is 96.64%, which is 6.69%, 5.13% and 3.86% higher than that of random forest model, support vector machine model and xgboost model respectively, It enriches the prediction methods of debris flow disasters and provides new ideas for relevant decision-making departments.

**Keywords:** debris flow; forecast model; linear discriminant analysis; extreme gradient boosting; gray wolf optimization algorithm

## 0 引 言

泥石流是山区主要的地质灾害之一[1]。我国国土面积大,自然环境复杂多样,山区地形多是峰高沟深、山势险峻的地貌特征。近年来伴随着全球气温持续攀升海平面持续增高,多地出现暴雨、冰雪融化、水库溃决的事件是频发泥石流灾害的主要原因之一[2]。泥石流灾害已经得到了国家相关部门的高度重视,为了能够有效的降低灾害带来的危害,加强防范就显的尤为突出重要,可以有效的减少人员伤亡及避免国家和人民财产损失[3]。因此目前泥石流灾害预

收稿日期:2022-07-06
* 基金项目:陕西省自然科学基础研究计划项目(2022JM-322)、陕西省技术创新引导专项(2020CGXNX-009)资助

报问题已成为自然灾害领域非常前沿和迫在眉睫的研究。提出合理可靠的泥石流灾害预测方案已经成为我国应对泥石流灾害做好防灾减灾工作的一个重要环节。

泥石流领域研究者根据泥石流的成灾特点进行深层次的研究并结合各种数学模型,从 20 世纪 90 年代以来不断地提出了多种泥石流灾害预测方法,极大程度的开阔了泥石流灾害预测方面新思路。郑国强等[4]首次将 Bayes 判别分析法运用在泥石流灾害预报中,预测方法简单并得到了较好的预测效果,但此方法过于依赖专家经验,当经验不足时误判性较大;李丽敏等[5]采用多传感器信息融合和径向基函数(radial basis function,RBF)神经网络模型对泥石流发生概率进行预测,很好的克服了以往泥石流数据因单一采集手段所造成的遗漏问题,但多种致灾因子间的相互关联性并没有被分析到;Xu 等[6]利用逻辑回归模型建立泥石流预测模型,并考虑到泥石流与降雨之间的关系,选择了 7 个重要的环境因子引入整个研究区域后,确实可行的提高了泥石流预测模型精度,但提前是要对不同地区域进行等级划分和权重分配,否则会影响预测的精度;周伟等[7]根据泥石流形成所需的地形地貌、物源条件,选取泥石流的致灾因子,基于 Fisher 判别法建立泥石流预报模型,弥补了降雨阈值模型仅靠降雨资料分析的不足提高了预测的精度,但该模型所需因子较多,搜集和处理数据过程十分繁琐,消耗时间很长;以往的泥石流预报多采用机器学习方法,但徐根祺等[8]把深度学习的方法使用到了泥石流灾害预报中,运用了快速的多主成分并行提取方法并结合宽度学习方式实现了泥石流灾害的概率预测,有效的降低了数据的维度问题,但是却没有考虑到宽度学习模型结构复杂且需要巨大的数据量来进行训练,通常情况下泥石流的数据难以采集。

同时,本文使用的极端梯度提升树(extreme gradient boosting,XGBoost)算法是传统梯度提升树算法(gradient boosting decision tree,GBDT)的改进,具有运算速度快、预测精度高等优势。孙朝云等[9]针对高速公路服务区交通量提出经粒子群优化改进的 XGBoost 模型,使预测精度得到提升;龚雪娇等[10]利用贝叶斯算法对 XGBoost 超参数进行寻优并应用在电网短期峰值负荷预测方面,取得了较好的泛化效果和预测结果;谭海旺等[11]使用 LSTM 算法优化 XGBoost 预测光伏发电功率,获得了接近真实值的预测效果。

为了解决在应用 XGBoost 算法时输入数据维度过大和提高预测准确率问题,本文首先对原始数据进行预处理,然后利用线性判别分析法(linear discriminant analysis,LDA)进行降维,解决模型输入维数过大问题;利用 XGBoost 预测精度高的优点对泥石流灾害是否发生进行预测,但该模型依旧存在超参数多复杂性大容易陷入局部最优,因此选择使用灰狼寻优算法(grey wolf optimizer,GWO)确定 XGBoost 模型超参数,提高模型参数寻优能力和避免陷入局部最优并进一步提高预测准确率。通过与其他预报模型进行对比验证,得出本文提出的模型具有较好预测精度,可以拓展泥石流灾害预测方法。

## 1 LDA 降维算法

LDA 是一种常见有监督学习数据降维算法,对于数据集的每一个样本都有类别输出。原理是将带标签的数据通过投影到维度更低的空间中,使投影后的数据类内方差达到最小,类间方差达到最大,得到较好的数据降维效果[12]。LDA 算法的基本原理如下。

已有数据集 $\boldsymbol{D}=(x_1,y_1),(x_2,y_2),\cdots(x_m,y_m)$,样本 $\boldsymbol{x}_i$ 为任意 $n$ 维向量,类别 $y_i\in C_1,C_2,\cdots C_k$,定义 $N_{j(j\in1,2,\cdots k)}$ 是第 $j$ 类的样本个数,$X_{j(j\in1,2,\cdots k)}$ 是第 $j$ 类样本集合,$\mu_{j(j\in1,2,\cdots k)}$ 是第 $j$ 类样本的均值,$\sigma_j$ 是第 $j$ 类样本均方差,$\mu_j$ 和 $\sigma_j$ 表达式如式(1)、(2):

$$\mu_j=\frac{1}{N_j}\sum_{x\in X}x \tag{1}$$

$$\sigma_j=\sum_{x\in X}(x-\mu_j)(x-\mu_j)^{\mathrm{T}} \tag{2}$$

式中:$j\in\{0,1,\cdots,k\}$,将其降到一维。寻找一个向量 $\boldsymbol{w}$,不同样本中心投影之后得:

$$\mu_{wj}=\boldsymbol{w}\mu_j \tag{3}$$

定义类内散度矩阵如式(4)所示,表示各类数据点聚集程度:

$$S_W=\sum_{j=1}^{k}\sum_{x\in X}(x-\mu_j)(x-\mu_j)^{\mathrm{T}} \tag{4}$$

定义类间散度矩阵如式(5)所示,表示不同类分散程度:

$$S_b=\sum_{j-1}^{k}N_j(\mu_j-u)(\mu_j-u)^{\mathrm{T}} \tag{5}$$

设投影到的低维空间维度为 $d$,对应的基向量为 $w_1$,$w_2,\cdots w_d$,构成投影矩阵 $\boldsymbol{w}$ 函数 $J$:

$$J=\frac{\boldsymbol{w}^TS_b\boldsymbol{w}}{\boldsymbol{w}^TS_w\boldsymbol{w}} \tag{6}$$

运用拉格朗日乘子法可得:

$$S_w^{-1}S_b\boldsymbol{w}=\lambda\boldsymbol{w} \tag{7}$$

此时可以得到矩阵的最大特征值。而投影方向就是这个特征值对应的特征向量。转化样本集的每一个样本,得到新样本 $P_i=\boldsymbol{W}^{\mathrm{T}}x_i$。最后得到降维后的新样本集 $\overline{D}=\{(p_1,y_1),(p_2,y_2),\cdots(p_m,y_m)\}$。

## 2 XGBoost 模型

XGBoost 是一种适合解决回归或预测的机器学习算法,由多个弱学习器叠加训练而成的 Boosting 类集成算法。具有稳定性强、预测性能优异等特点。应用在泥石流灾害预测时,XGBoost 算法是将预测结果与真实值之间残差进行不断拟合,逐步迭代直到符合停止条件,最终对所有树拟合的结果累加求和,得到预测结果[13]。

令数 $\boldsymbol{D}=(x_i,y_i)(|D|=n,x_i\in R^m,y_i\in R)$,$m$ 表

示特征维数，$n$ 表示样本数量。假设某个模型由 $K$ 颗树组成，则：

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{8}$$

式中：$\hat{y}_i$ 为结果输出；$x_i$ 为输入的第 $i$ 个样本；$f_k(x_i)$ 为第 $k$ 棵子树的输出；$F$ 为回归树空间。

XGBoost 算法的目标函数由训练误差项与约束正则项两部分组成：

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{9}$$

式中：$\hat{y}_i$ 为预测值；$y_i$ 为真实值；$k$ 为子树个数；$f_k$ 为第 $k$ 棵子树的个数的输出结果值。其中约束正则项 $\Omega(f)$ 的具体形式为：

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \| \boldsymbol{\omega} \|^2 \tag{10}$$

式中：$T$ 为子树叶子节点的个数；$\boldsymbol{\omega}$ 为叶子节点的分数组成集合；$\gamma$ 和 $\lambda$ 为系数。

XGBoost 模型采用的是加法训练方式，设第 $t$ 次迭代训练时的第 $i$ 棵树输出为 $\hat{y}_i^{(t)}$：

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{11}$$

式(11)表示模型最终输出结果是前 $t-1$ 棵树与第 $t$ 颗树的输出结果相加值。于是在第 $t$ 次时，目标函数可以写为：

$$L^{(t)} = \sum_{i=1}^{N} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)) \tag{12}$$

通过二阶泰勒展开目标函数得到的近似结果为：

$$L^{(t)} \cong \sum_{i=1}^{N} [l(y_i, \hat{y}_i^{(t-1)}) + g f_t(x_i) + \frac{1}{2} h f_t^2(x_i)] + \Omega(f_t) \tag{13}$$

式中：$g$ 为误差函数一阶导数；$h$ 为误差函数二阶导数。

泥石流灾害预报模型可以看做在解决二分类问题。模型的输入是引发泥石流的致灾因子样本数据，对应输出结果为泥石流是否发生，标记为 0 和 1 两个类别分别代表无/有泥石流灾害发生。

$$l(y, \hat{y}^{t-1}) = -\sum_{c=1}^{M} y_c \log(p_c) \tag{14}$$

式中：$l(y, \hat{y}^{t-1})$ 表示实际类别和预测类别之间的差异大小。$M$ 表示类别数，$y_c$ 为指示变量，当预测值与实际值相同时为 1，不同时为 0；$p_c$ 为类别 $c$ 的预测概率。

## 3　GWO-XGBoost 模型

灰狼算法是一种根据灰狼种群机制和掠夺行为推演得到的，具有搜索速度快、易得到全局最优解和较强稳定性等优点[14]。在 GWO 算法中等级最高的狼被标记为 α 负责捕猎(寻优)过程中带领整个狼群并且决定抓捕方向，剩余狼群按等级被标记为 β、δ 和 ω。等级较低的狼群要服从等级较高的狼群指挥从而进行群体捕猎行动。GWO 优化过程具体步骤如下：

1)等级划分

对整个狼群通过适应度进行 4 个等级划分，分别为 $\alpha$、$\beta$、$\delta$ 和 $\omega$。GWO 算法的优化过程均是在 $\alpha$、$\beta$ 和 $\delta$ 三匹狼的领导下进行的。

2)围捕猎物

灰狼围捕猎物时，是先一步步接近猎物，然后包围。该行为用数学模型描述表示如下：

$$\boldsymbol{D} = |\boldsymbol{C} \cdot \boldsymbol{X}_p(t) - \boldsymbol{X}(t)| \tag{15}$$

$$\boldsymbol{X}(t+1) = \boldsymbol{X}_p(t) - \boldsymbol{A} \cdot \boldsymbol{D} \tag{16}$$

其中，$\boldsymbol{A}$ 和 $\boldsymbol{C}$ 是协作系数向量；$t$ 是指迭代的次数；$\boldsymbol{X}_p$ 表示猎物的位置向量；$\boldsymbol{X}$ 是代表当前灰狼的位置向量。$\boldsymbol{A}$、$\boldsymbol{C}$ 的值可以由下式得到：

$$\boldsymbol{A} = 2a \cdot r_1 - a, a = 2 - 2t \cdot \frac{1}{G_{\max}} \tag{17}$$

$$\boldsymbol{C} = 2r_2 \tag{18}$$

式中：$a$ 为收敛因子，初始值为 2，收敛因子会随着迭代次数线性下降递减至 0，$r_1$ 和 $r_2$ 是[0,1]之间均匀分布的随机数，$G_{\max}$ 是最大迭代次数。

3)狩猎

狩猎过程是具有较强搜索能力的 $\alpha$、$\beta$ 和 $\delta$ 灰狼带领下开展的。其他灰狼的位置则是在 $\alpha$、$\beta$ 和 $\delta$ 指引下随机更新。狩猎过程用数学模型表示如下：

$$D_\alpha = |C_1 \cdot X_\alpha|, D_\beta = |C_2 \cdot X_\beta|, D_\delta = |C_3 \cdot X_\delta| \tag{19}$$

$$X_1 = X_\alpha - A_1(D_\alpha), X_2 = X_\beta - A_2(D_\beta) \tag{20}$$

$$X_3 = X_\delta - X_3(X_\delta) \tag{21}$$

$$X(t+1) = \frac{X_1 + X_2 + X_3}{3} \tag{22}$$

式(22)即为猎物的位置(最优解)。

4)攻击猎物

狼群狩猎的最后一步是对猎物进行攻击并捕食，计算最优解。判断结果是否达到要求，如果达到最大迭代次数，则运行结束并输出最优解；否则返回狩猎过程继续进行寻优计算。

在 XGBoost 模型进行泥石流预测时，不恰当的参数设置会对模型预测结果产生很大的影响[15]；因此，本文选择 LDA 进行数据降维并使用 GWO 算法对 XGBoost 模型重要超参数进行寻优设置，组合模型一方面解决 XGBoost 模型容易陷入局部最优问题，另一方面进一步提高预测结果的准确率。基于 GWO-XGBoost 的预测流程如图 1 所示。

GWO-XGBoost 预测具体实现步骤如下：

1)对原始数据进行预处理并使用 LDA 算法进行数据降维；

2)对 GWO 算法进行初始化参数设置，参数包括灰狼数量 $N$，以及 $a$、$A$ 和 $C$ 和最大迭代次数 $t$ 等多个参数，以及灰狼种群 $X = (X_1, X_2, \cdots X_N)$ 和狼群个体的位置 $X_i = (X_{i1}, X_{i2}, \cdots X_{iD})$，其中 $i \in \{1,2,3,\cdots N\}$；

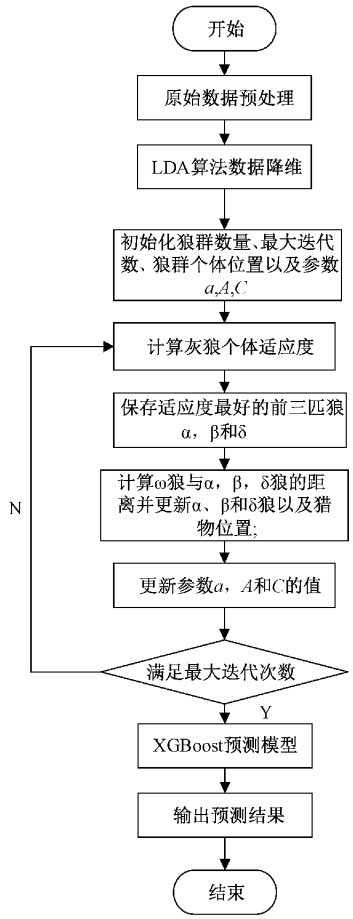


图 1　GWO-XGBoost 预测流程

3)计算灰狼个体适应度;

4)保存适应度最好的前三匹狼 $\alpha$、$\beta$ 和 $\delta$;

5)根据式(17)计算得到各 $\omega$ 狼与 $\alpha$、$\beta$、$\delta$ 各个狼的距离,并依据式(19)～(22)更新 $\alpha$、$\beta$ 和 $\delta$ 狼以及猎物位置;

6)更新算法参数 $a$,$A$,$C$;

7)判断是否达到最大迭代次数,若达到则保存最优解,否则回到步骤 3)

8)根据以上步骤确定 GWO 优化算法对 XGBoost 模型超参数寻优的结果,得到泥石流灾害发生预测值。

## 4　仿真验证与结果分析

### 4.1　泥石流研究区概况

磨子沟地处于陕西省安康市紫阳县城关镇太平乡沟长约 3 km,沟谷两侧山地多呈陡峭之势且陡坡范围大约在 25°～45°。该地区全年降雨大多集中在 6 月～9 月,全年平均降雨量约为 1 000 mm,日平均降雨量最高时约为 185 mm 巨大的降雨量为泥石流灾害发生提供了充足水源。2010 年 7 月发生特大暴雨造成泥石流灾害山体多处出现塌方情况[16]。由于该地区温度变化差异大且十分潮湿因此极易造成土质疏松,并且矿产丰富人类活动强度大,上述因素都是导致泥石流灾害成为该区域安全隐患的原因之一[17]。

根据实地考察以及分析泥石流形成的种种因素,借助该地区已存在的科研项目平台,在磨子沟部分山体区域布设的雨量监测、土壤含水率监测、孔隙水压力监测、倾角坡度监测、泥位监测等相关传感器,获得较为可靠的原始数据作为实验依据。

### 4.2　泥石流预测总体框架

本文依据在磨子沟地区布设的各种传感器监测采集到的可靠数据为基础,选取降雨量(mm)、土壤含水率(%)、孔隙水压力(kPa)、山坡坡度(°)、相对高差(m)、泥位电信号(mA)、植被覆盖率(%)、次声(Hz)8 个泥石流诱发条件作为泥石流预测初始致灾因子,然后经过数据预处理和降维过程,降低原始数据对预测精度影响;将处理过后的数据作为 XGboost 预测模型输入,GWO 优化算法对模型超参数进行寻优并对比布谷鸟(CS)和遗传(GA)优化算法超参数寻优的结果,解决模型容易陷入局部最优问题,相同条件下对比 RF、SVM 和未经优化的 XGBoost 预测模型精度。本文具体技术路线如图 2 所示。采集到的磨子沟部分泥石流数据如表 1 所示。

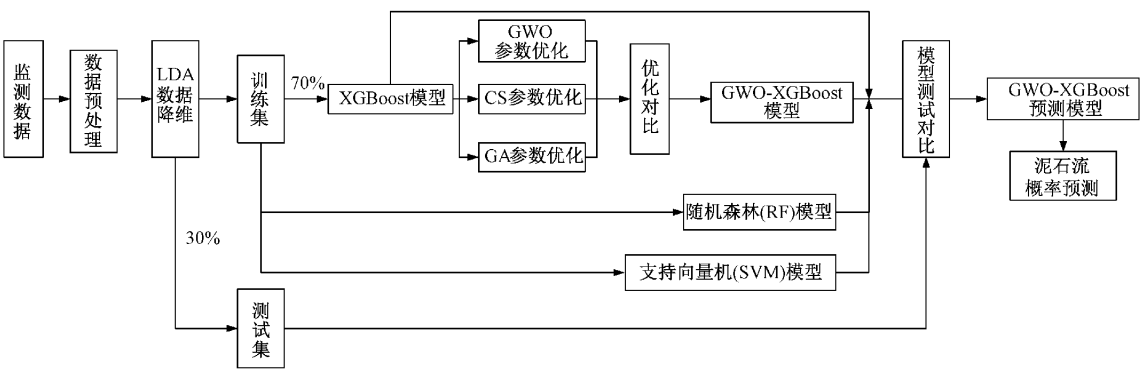


图 2　技术路线图

### 4.3　数据预处理

由于研究区域环境因素复杂原因,传感器收集的泥石流致灾因子数据并不适用直接进行预测模型的训练[18]。因此需要对数据进行预处理操作。选取 600 组泥石流数

**表 1　磨子沟地区部分泥石流数据**

| 编号 | 类别 | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 降雨量/mm | 土壤含水率/% | 孔隙水压力/kPa | 山坡坡度/(°) | 相对高差/m | 泥位电信号/mA | 植被覆盖率/% | 次声/Hz |
| 1 | 20.4 | 0.05 | −0.7 | 26.4 | 382 | 7.5 | 3.12 | 0.5 |
| 2 | 25.9 | 0.07 | −0.33 | 35.1 | 412 | 7.8 | 1.23 | 2.4 |
| 3 | 30.3 | 0.11 | 0.22 | 27.4 | 490 | 8.2 | 0.59 | 2.6 |
| 4 | 32.5 | 0.12 | 0.32 | 32.2 | 580 | 8.6 | 4.19 | 3.1 |
| 5 | 35.4 | 0.19 | 0.34 | 30.8 | 489 | 7.4 | 0.54 | 2.9 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 95 | 131.7 | 0.3 | 0.75 | 38.1 | 465 | 14.4 | 6.14 | 5.4 |
| 96 | 136.2 | 0.31 | 0.77 | 32.1 | 409 | 14.5 | 2.11 | 5.7 |
| 97 | 137.3 | 0.32 | 0.77 | 33.2 | 392 | 14.6 | 3.21 | 5.5 |
| 98 | 142.5 | 0.33 | 0.78 | 34.3 | 580 | 14.8 | 0.89 | 5.6 |
| 99 | 145.2 | 0.34 | 0.79 | 35.6 | 477 | 15.2 | 3.45 | 6.8 |
| 100 | 147.3 | 0.34 | 0.8 | 36.2 | 399 | 15.7 | 1.58 | 7.1 |
| 101 | 148.5 | 0.35 | 0.82 | 36.8 | 423 | 15.9 | 5.16 | 7.4 |

据作为研究样本，其中包括已经发生的泥石流灾害样本数据也有未发生的泥石流灾害样本数据。预处理具体手段包括缺失值处理、异常值剔除和归一化。

1)缺失值处理。数据遗漏是很正常的情况，对传感器遗漏的数据按属性进行统计，得到缺失率 $q$，若 $q\geqslant 90\%$，则直接删除该列；若处于 $40\%\leqslant q<90\%$，则使用相邻属性加权填充；若 $20\%\leqslant q<40\%$，则使用均值进行填充；若 $q<20\%$，则使用众数填充。

2)异常值剔除。对于样本集数值距离均值达到 3 倍以上或达到 5 倍标准差的均异常值，则直接剔除。

3)归一化。由于不同种类的数据量纲是不一样的，数据的复杂多样化会影响预测模型的准确率，所以归一化处理是必要的。可通过式(23)进行归一化处理：

$$R'=\frac{R-R_{\min}}{R_{\max}-R_{\min}} \tag{23}$$

其中，$R'$为归一化处理后的数据；$R$ 为原始数据；$R_{\max}$ 和 $R_{\min}$ 为最大值和最小值。

对预处理之前和预处理之后的原始数据，均采用 GWO-XGBoost 模型进行仿真训练，得到数据预处理前后的结果对比如表 2 和图 3 所示。

**表 2　数据预处理前后结果对比**

| | 增量周期数/个 | 准确率/% |
|---|---|---|
| 数据与处理前 | 1 700 | 79.82 |
| 数据与处理后 | 600 | 94.68 |

注：仿真均在 CPU2.4 GHz，内存 4 G 机器上使用 Matlab 进行。

由表 2 和图 3 分析可得，未预处理前的数据，当模型增量周期数达到 1 700 个左右时，模型预测精度才趋于稳定，

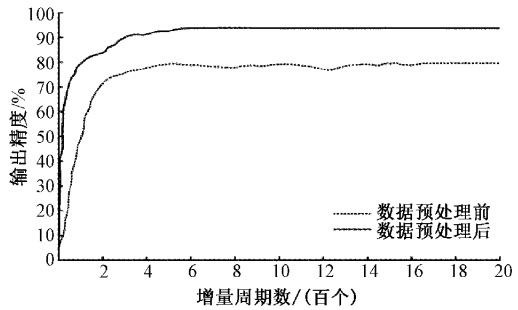


图 3　数据预处理前后结果对比

并最终稳定在 79.82%。但是对于经过预处理后的数据，当模型增量周期数达到 600 个左右时，模型预测精度稳定在 94.68%。表明数据经过预处理操作是可以更快趋于稳定且有效提高模型预测精度。

### 4.4　LDA 数据降维

本文利用磨子沟监测点前期布设的传感器，收集到 8 类引发泥石流的初始致灾因子数据。经过预处理后的数据直接作为预测模型输入，依旧存在模型输入维度过大和样本特征变量之间存在一定程度相关性问题，影响预测精度问题，因此本文选择 LDA 算法，解决上述问题。LDA 是一种有监督学习降维算法，对特征空间降维时，将预处理后的 8 维特征数据样本集作为输入，对应样本泥石流灾害结果作为输出标签，训练得到样本特征变量的独立辨识力，如图 4 所示。

图 4 中，数据集由原来的 8 维经 LDA 处理后降成 $(k-1)$维($k$ 为样本类别)，因为样本只有 0 和 1 两种类别，所以只能降至一维，极大程度降低样本数据复杂度。降维后特征变量累计辨识力足以达到选取的精度标准，避免特征值之间相互影响并解决维度灾难问题，使数据更加适用

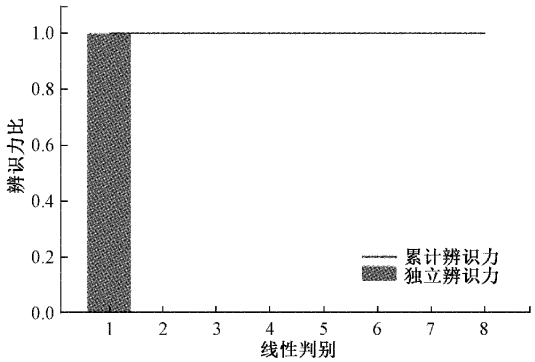


图 4 LDA 特征降维

于模型输入提高预测模型的精度。

### 4.5 XGBoost 超参数寻优

XGBoost 模型的 3 类超参数分别为常规超参数、任务超参数以及提升器超参数需要被设定。一般情况下对于常规超参数和任务参数都使用默认值,所以只需要对提升器超参数进行适当调整,达到优化模型性能目的[19]。

需要确定的主要寻优参数为:学习速率(eta)、树的最大深度(max_depth)、使用树的数量(n_estimator)、最小叶子节点样本权重和(min_child_weight)4 个参数。学习速率控制模型的训练效率;树的最大深度和最小叶子节点样本权重用来防止模型过度拟合或欠拟合。本文选用 GWO 优化算法并对比布谷鸟算法(cuckoo search,CS)和遗传算法(genetic algorithm,GA),在相同条件下对 XGBoost 模型超参数进行寻优。3 种寻优算法寻优结果如表 3 所示。

**表 3 3 种寻优算法寻优结果**

| 超参数名称 | 取值范围 | GWO | CS | GA |
|---|---|---|---|---|
| eta | [0,1] | 0.38 | 0.5 | 0.1 |
| max_depth | [1,10] | 6 | 3 | 5 |
| min_child_weight | [1,20] | 1 | 3 | 1 |
| n_estimator | [1,400] | 331 | 241 | 105 |

分析表 3 可得,由不同寻优算法所得到的参数值相差甚大,这是 CS 和 GA 在寻优过程中陷入局部最优导致。综合考虑,为了证明 GWO 优化算法的优越性,选择均方误差和寻优时间作为寻优的判定指标。3 种寻优算法寻优性能对比如表 4 所示,GWO 在均方误差和时间方面上均具有优势,均方误差为 0.008 24,相较于布谷鸟和遗传寻优算法分别低了 0.006 39 和 0.013 43,寻优时间分别快了 6.4 s 和 23.41 s。结合表 3 和 4 分析可得 GWO 算法相较于其他两种寻优算法具有更好更快的参数寻优能力。

### 4.6 GWO-XGBoost 预测模型

将经过预处理和 LDA 降维后的数据按照训练集∶测试集=8∶2 比例随机取样进行划分,其中训练集数据作为 XGBoost 预测模型的输入,并使用 GWO 优化算法对模型超参数寻优进行模型训练,然后使用测试集数据对训练好

**表 4 3 种寻优算法性能对比**

| 寻优算法 | 均方误差 | 寻优时间/s |
|---|---|---|
| 灰狼 | 0.008 4 | 3.12 |
| 布谷鸟 | 0.014 63 | 9.52 |
| 遗传 | 0.021 67 | 26.53 |

的模型进行验证,相同条件下和随机森林(random forest,RF)、支持向量机(support vector machine,SVM)与未优化的 XGBoost 预测模型进行对比。

本文所构建的 GWO-XGBoost 模型的预测效果可以由总体分类准确率 TR 和 AUC 值反映。设 TP 和 TN 分别表示泥石流灾害发生样本和泥石流灾害不发生样本预测正确的数量,FN 和 FP 均表示预测错误的数量。模型预测分类的结果可由混淆矩阵,如表 5 所示。

**表 5 混淆矩阵**

| 真实类别 | 预测类别 | | |
|---|---|---|---|
| | 预测发生泥石流 | 预测未发生泥石流 | 合计 |
| 真实发生泥石流 | TP | FN | TP+FN |
| 真实未发生泥石流 | FP | TN | FP+TN |
| 合计 | TP+FP | FN+TN | |

根据模型分类准确率评估指标:TR=(TP+TN)/(TP+TN+FP+FN)公式得到各模型预测平均准确率如表 6 所示。

**表 6 各模型预测平均准确率对比**

| 模型 | 准确率/% |
|---|---|
| RF | 89.95 |
| SVM | 91.51 |
| XGBoost | 92.78 |
| GWO-XGBoost | 96.64 |

分析表 6 可得,平均准确率方面,GWO-XGBoost 比 RF 高出 6.69%,比 SVM 高出 5.13%,比 XGBoost 高出 3.86%。由此可知,GWO-XGBoost 预测泥石流灾害发生平均准确率更高。

此外,通过公式 TPR=TP/(TP+FN)和 FPR=FP/(FP+TN)计算得 ROC 曲线对模型进行验证,其中 TPR(true positive rate)是真阳性率;FPR(false positive rate)是假阳性率,预测模型的性能好坏取决于 ROC 曲线下方包围面积 AUC(area under curve)值,面积越大说明预测模型效果越好[20]。本文提出的 GWO-XGBoost 预测模型与 RF、SVM 和未经优化的 XGBoost 绘制的 ROC 曲线如图 5 所示且 4 种模型 AUC 均值对比如表 7 所示。

分析表 7 可得,GWO-XGBoost 预测模型的 AUC 均

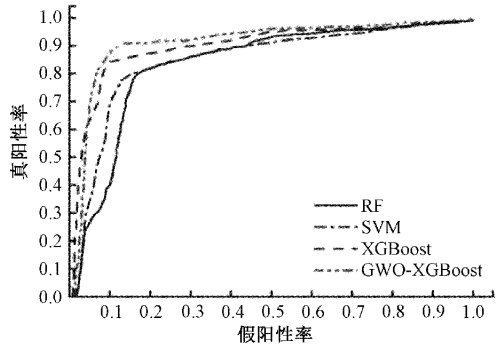


图 5　4 种模型 ROC 曲线对比

**表 7　4 种模型 AUC 均值对比**

| 模型 | AUC 均值 |
|---|---|
| RF | 89.16 |
| SVM | 90.63 |
| XGBoost | 91.84 |
| GWO-XGBoost | 95.11 |

值为 95.11%，比 RF 模型高出 5.95%，相较于 SVM 模型高出 4.48%以及与 XGBoost 模型相比高出 3.27%。实验对比结果表明，本文提出的预测模型比较 RF、SVM 以及未经优化的 XGBoost，具有更好的预测精度，可见将 GWO-XGBoost 组合模型运用在泥石流灾害预报方面效果好。

## 5　结　　论

本文采用 LDA 算法和经 GWO 优化算法进行参数寻优的 XGBoost 泥石流灾害预测模型，对泥石流发生概率进行预测。选择磨子沟为研究区域，分析了泥石流致灾因子与灾害发生概率之间关系。同时将本文所提出的模型与 RF 模型、SVM 模型以及未经优化的 XGBoost 模型进行比较。通过验证，证明了本文提出 GWO-XGBoost 模型在泥石流预测方面的准确性。具体结论如下：

1)由于研究区域自然环境复杂，传感器收集到的数据存在缺失情况，预测准确率会受影响，所以需要预处理，对经过预处理的数据使用 LDA 算法进行降维，将原始的 8 维数据降至 1 维，降低致灾因子之间的耦合性，避免模型输入维度灾难问题使数据更加适用于模型输入进一步提高准确率。

2)将极端梯度提升树预测方面准确性好应用于泥石流灾害方向。与传统 RF 和 SVM 两种模型比较，展现出更好的预测性能和泛化能力。同时选用 GWO 优化算法对 XGBoost 中的超参数进行寻优，组合模型一方面避免了 XGBoost 模型容易陷入局部最优问题，另一方面进一步提高模型预测准确率。

3)通过验证样本集得出本文提出模型具有较高的平均准确率和 AUC 值，说明将该模型应用于泥石流灾害预测是合理可行的并且丰富了我国泥石流灾害预测方法。

4)本文研究工作具体针对泥石流灾害空间预测开展，然而如何计算临灾时间，如何保护人民生命安全和财产还有待进一步探究。同时各地区泥石流灾害形成原因复杂多样，因此本文提出的预测模型还需进一步优化和完善才能更好应用在其它地区。

## 参考文献

[1] 廖学海，陈洪凯，蒋鸿正. 山区公路泥石流病害与防治研究[J]. 水利水电科技进展，2021，41(6)：109.

[2] 吴悦，任涛，杨卓静，等. 应用于地质灾害的泥石流监测分析仪[J]. 电子测量技术，2013，36(3)：67-70.

[3] 李涌波，陈实，陈敏，等. 多方数据融合的山洪灾害模型应用研究[J]. 电子测量技术，2021，44(1)：92-97.

[4] 郑国强，张洪江，刘涛，等. 基于 Bayes 判别分析法的密云县山洪泥石流预报模型[J]. 水土保持通报，2009，29(1)：83-87，107.

[5] 李丽敏，温宗周，李璐，等. 基于多参数融合和 RBF 神经网络的泥石流预报[J]. 西安工程大学报，2017，31(1)：77-81.

[6] XU W B，YU W J，JING S C，et al. Debris flow prediction models based on environmental factors and susceptible subarea classification in Sichuan，China[J]. Natural Hazards，2013，67(2)：869-878.

[7] 周伟，邓玖林. 基于 Fisher 判别法的台风暴雨泥石流预测模型[J]. 水科学进展，2019，30(3)：392-400.

[8] 徐根祺，李丽敏，温宗周，等. 基于宽度学习模型的泥石流灾害预报[J]. 山地学报，2019，37(6)：868-878.

[9] 孙朝云，吕红云，杨荣新，等. 改进粒子群优化 XGBoost 模型的高速公路服务区交通量预测[J]. 北京交通大学学报，2021，45(5)：74-83.

[10] 龚雪娇，朱瑞金，唐波. 基于贝叶斯优化 XGBoost 的短期峰值负荷预测[J]. 电力工程技术，2020，39(6)：76-81.

[11] 谭海旺，杨启亮，邢建春，等. 基于 XGBoost-LSTM 组合模型的光伏发电功率预测[J]. 太阳能报，2022，43(8)：75-81.

[12] 杨明莉，范玉刚，李宝芸. 基于 LDA 和 ELM 的高光谱图像降维与分类方法研究[J]. 电子测量与仪器学报，2020，34(5)：190-196.

[13] 王媛彬，李媛媛，韩骞，等. 基于 PCA-BO-XGBoost 的矿井回采工作面瓦斯涌出量预测[J]. 西安技大学学报，2022，42(2)：371-379.

[14] 肖艳丽，向有涛. 企业债券违约风险预警——基于 GWO-XGBoost 方法[J]. 上海金融，2021(10)：44-54.

[15] 洪洋，汪家兴，宁宇航. 基于 GWO-RBF 神经网络的天线建模参数预测[J]. 电子测试，2021(19)：60-62.

[16] 张亮，陈练武. 陕西紫阳县磨子沟泥石流特征及危险性分析[J]. 四川地质学报，2017，37(1)：104-107.

[17] 李丽敏,程少康,温宗周,等. 基于改进 KPCA 与混合核函数 LSSVR 的泥石流预测[J]. 信息与控制,2019,48(5):536-544.

[18] 李丽敏,张俊,温宗周,等. 基于布谷鸟优化轻量梯度提升机的泥石流预测[J]. 科学技术与工程,2021,21(30):13177-13184.

[19] 徐伟,夏志祥,行鸿彦. 基于集成经验模态分解和极端梯度提升的雷电预警方法[J]. 仪器仪表学报,2020,41(8):235-243.

[20] 陶世银,贺敬安. 基于 XGBoost 与特征重要性筛选的闪电预报模型构建研究[J]. 国外电子测量技术,2022,41(1):99-105.

## 作者简介

**王智勇**,硕士研究生,主要研究方向为人工智能及其应用。
E-mail:783305083@qq.com

**李丽敏**,博士,副教授,主要研究方向为智能算法及其在地质灾害监测领域的应用研究。
E-mail:2364225096@qq.com