文章编号:1671-4598(2023)04-0225-07 DOI:10.16526/j.cnki.11—4762/tp.2023.04.035 中图分类号:TP389.1 文献标识码:A

基于 Xgboost 优化的 KELM 滑坡预报模型研究

李 璐[1],徐根祺[2],杨 倩[1],王艳娥[1],赵正健[1]
(1. 西安思源学院 理工学院,西安 710038;2. 西安交通工程学院 机械与电气工程学院,西安 710030)

摘要:针对极限学习机对滑坡预测准确性低及在训练过程中模型不稳定的问题,引入 RBF 高斯核函数并使用极限梯度提升树算法 Xgboost 对 KELM 进行优化,建立了 Xgboost 优化后的 Xgboost-KELM 预测模型;首先采用高斯核 RBF 作为极限学习机的核函数,解决隐藏节点随机映射问题,增加模型稳定性及适用性;其次将清洗后的监测数据作为模型输入,并使用 Xgboost 寻优算法对核函数中的超参数进行优化,通过 4 组测试集进行 Xgboost-KELM 建模,依据均方误差迭代曲线得出最佳超参数;最后使用两组 10%样本集验证模型评价指标及稳定性,实验结果 AUC 均值对比模型至少提高 3 个百分点,Precision、Accuracy 及 Recall 至少高于对比模型 1.7 个百分点,同时 Xgboost-KELM 模型的方差及偏差都较小,证明该模型稳定性较好,实验结果说明 Xgboost-KELM 模型具有较好的预测效果,在滑坡灾害预测中有较好的预测能力。

关键词:高斯核 RBF;KELM;Xgboost 超参数;滑坡灾害;预报模型

Research on the Optimization KELM Landslide Prediction Modeling Based on Xgboost

LI Lu[1], XU Genqi[2], YANG Qian[1], WANG Yane[1], ZHAO Zhengjian[1]
(1. Xi'an Siyuan University of Science and Technology, Xi'an 710038, China;
2. School of Mechanical and Electrical Engineering, Xi'an Transportation Engineering College, Xi'an 710030, China)

Abstract: To solve the problems of low accuracy for an extreme learning machine (ELM) in landslide prediction, and the instability of the model in the training process, a radical basis function (RBF) Gaussian kernel function is introduced to optimize the Kernel extreme learning machine (KELM) by using a xtreme gradient boosting (Xgboost) algorithm, and the Xgboost KELM prediction model of the optimizated Xgboost is established; Firstly, as the kernel function of the limit learning machine, the Gaussian kernel RBF is used to solve the random mapping of the hidden nodes and increase the stability and applicability of the model; Secondly, the cleaned monitoring data is used as the model input, and the Xgboost optimization algorithm is used to optimize the super parameters in the kernel function, the Xgboost KELM modeling is conducted through four groups of test sets, and the best super parameters are obtained according to the iteration curve of the mean square error; Finally, two groups of 10% sample sets are used to verify the model evaluation indicators and stability. The experimental results show that compared with the GA and GC models, the AUC mean of the Xgboost-KELM model is increased by 3 percentage points, and the performance indexes of precision, accuracy and recall by at least 1.7 percentage points. At the same time, the variance and deviation of the Xgboost KELM model are small, which proves that the model is stable and has a good prediction ability in landslide disaster prediction.

Keywords: Gaussian kernel RBF; KELM; Xgboost super parameter; landslide disaster; forecasting model

0 引言

滑坡灾害的发生是在自然演变或人为因素的影响下,一种复杂的非线性动力学演化过程[1]。滑坡灾害由于它本身高频发生、分布区域广泛及破坏力极强,对山区人民生命、财产有极大的威胁,对防灾减灾工作提出严峻的考验[2]。近年来国家相关部门和单位陆续出台相关政策,滑坡等地质灾害的课题也成为热门课题,相关学者对其研究也取得不错的成效。滑坡早期预报可以有效减少灾害的损失,影响滑坡发生的主要条件[3-4]:岩土条件、地质构造及地面地形的条件。近年来对滑坡的研究主要包括:1)通过地面调查结合遥感技术观察并分析滑坡全域地形地貌,分析成灾机理[5];2)对滑坡易发区域地质构造分析滑坡形成的特征及成因,对滑坡滑动面失稳、变形及位移建模[6-7];3)地面外界因素如暴雨、坡体、地震等综合影响,进行滑坡灾害的分析及建模[8-9]。

随着机器学习理论的发展,非线性模型被广泛应用在滑坡灾害预测的理论中[10],赵晓萌[11]等从降雨量特征结合机器学习相关知识进行预测模型的建立;赵彬如[12]等从水文、气

收稿日期:2022-12-29; 修回日期:2023-01-04。
基金项目:陕西省教育厅科研计划资助项目(2022JK0515);陕西省自然科学基础研究计划项目(2023-JC-YB-464)。
作者简介:李 璐(1994-),女,山西长治人,硕士,助教,主要从事地质灾害预报、人工智能算法方向的研究。
徐根祺(1985-),男,陕西西安人,硕士,副教授,主要从事地质灾害预测预报方向的研究。
引用格式:李 璐,徐根祺,杨 倩,等.基于 Xgboost 优化的 KELM 滑坡预报模型研究[J].计算机测量与控制,2023,31(4):225-231.

象阈值进行滑坡预测的研究，预测效果较好，但只适用于降雨型的滑坡灾害预测；胡欣等[13]将 SVM-BP 模型应用于降雨型的滑坡灾害安全评价的模型研究中，随后马娟、杨宗佶等[14-15]将多参数预警模式应用于滑坡预测中，同时李丽敏等[16-17]也将多影响因子作为滑坡位移预测模型的输入，针对滑坡灾害建模预报灾害的发生。针对滑坡体灾害概率预测及预警的相关研究目前存在以下不足[18]：地域成灾机理复杂，成灾因子单一，导致预测预警准确率低；预测模型容易陷入局部最优及模型本身的稳定性及适应性不足。本文针对这些问题分析滑坡全域地形地貌成灾机理并筛选因子，构造滑坡灾害预测模型，引入核极限学习机（KELM，kernel extreme learning machine）应用于滑坡预测预报中，改善极限学习机（ELM，extreme learning machine）奇点容易产生及隐含层数量确定问题，增加模型本身的稳定性及适用性，并使用高斯核函数（RBF，radical basis function）作为核函数进行模型建立。以研究区山阳县山区 6 种影响因子作为滑坡预测模型的输入，通过极限梯度提升树算法 Xgboost（Xtreme gradient boosting）确定 KELM 模型中的参数惩罚系数及宽度参数，解决梯度决策树算法容易局部最优的问题，提高滑坡预测的精度。最后通过与其他预测算法进行比对，AUC 均值、Precision、Accuracy 及 Recall 都明显提升，体现出超参数优化后的模型在滑坡预测中较高的预测能力，给地质灾害研究带来活力及新思路。

1 Xgboost-KELM 模型建立

1.1 KELM 模型

ELM 是一种单隐层的前馈神经网络，对于线性问题计算速度快且拟合能力也较好，但由于其隐藏节点存在随机映射特点，所以训练的模型稳定性和泛化性能力较差[19-20]。KELM 在其基础上将核函数引入解决随机映射问题。针对于组不同的样本，ELM 隐含层节点数用表示，激励函数用 $g(\cdot)$ 表示：

$$\boldsymbol{H}\boldsymbol{\beta}=\boldsymbol{U} \tag{1}$$

$$\boldsymbol{H}=\begin{bmatrix} g(\omega_1 x_1+b_1) & g(\omega_M x_1+b_M) \\ g(\omega_1 x_N+b_1) & g(\omega_M x_N+b_M) \end{bmatrix}_{N\times M} \tag{2}$$

$\boldsymbol{H}$ 为隐含层的输出矩阵，$\boldsymbol{\beta}$ 为输出权值矩阵，$\boldsymbol{U}$ 为目标输出矩阵。

ELM 学习过程中为提升稳定性及泛化能力，引入正则化系数 C 利用最小二乘法求解最优值[21]。

$$\boldsymbol{\beta}=\boldsymbol{H}^{+}\boldsymbol{U}=\boldsymbol{H}^{T}\left(\frac{\boldsymbol{I}}{C}+\boldsymbol{H}\boldsymbol{H}^{T}\right)^{-1}\boldsymbol{U} \tag{3}$$

$\boldsymbol{H}^{+}$ 为 $\boldsymbol{H}$ 的广义逆。针对 ELM 解决多分类问题，依据 KKT 优化拉格朗日函数得到求解方案。当特征映射 $h(x)$ 未知的情况得出 ELM 的输出函数：

$$f_{\mathrm{ELM}}(x)=h(x)\boldsymbol{\beta}=h(x)\boldsymbol{H}^{T}\left(\frac{\boldsymbol{I}}{C}+\boldsymbol{H}\boldsymbol{H}^{T}\right)^{-1}\boldsymbol{U} \tag{4}$$

对角矩阵用 $\boldsymbol{I}$ 表示，正则化系数用 C 表示，输出向量用 $\boldsymbol{U}$ 表示。依据 Mercer's 条件，用核矩阵 $\boldsymbol{\Omega}_{\mathrm{ELM}}$ 代替随机矩阵 $\boldsymbol{HHT}$ 得出：

$$K(x_i,x_j)=h(x_i)h(x_j)=\boldsymbol{\Omega}_{\mathrm{ELM}}(i,j) \tag{5}$$

得到 KELM 的输出模型为：

$$f(x)=h(x)\boldsymbol{H}^{T}\left(\frac{\boldsymbol{I}}{C}+\boldsymbol{H}\boldsymbol{H}^{T}\right)^{-1}\boldsymbol{U}=\begin{bmatrix} K(x,x_1) \\ \vdots \\ K(x,x_N) \end{bmatrix}^{T}\left(\frac{I}{C}+\boldsymbol{\Omega}_{\mathrm{ELM}}\right)^{-1}\boldsymbol{U} \tag{6}$$

KELM 中核函数会直接影响模型本身的性能。图 1 为 RBF 高斯核函数曲线图，x 为分布中心偏移程度，$sigma$ 为分布的宽窄程度，当样本值接近分布中心 x，对核函数值影响较大，反而影响较小，这与内积原理相似，可以通过距离衡量样本的相似性，高斯径向基核函数因其是局部核函数，所以局部学习能力强且只受相离较近样本点的影响，所以其对于一定范围内距离的样本可以在特征空间种线性分离预测较准确。文中选取 RBF 为核函数：

$$K(x,x_1)=\exp\left(-\frac{\|x_i-x_j\|}{2\sigma^2}\right) \tag{7}$$

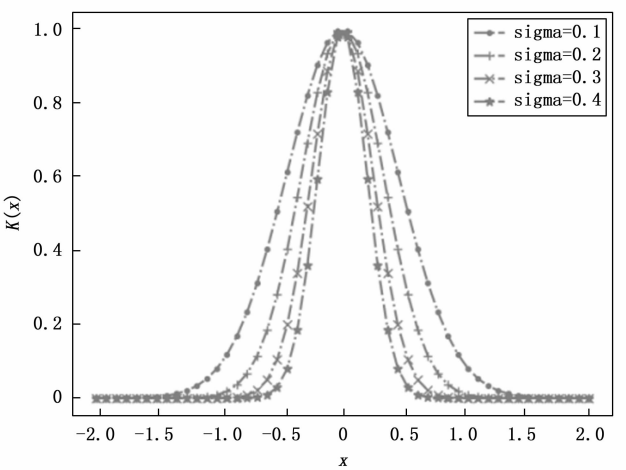

图 1 RBF 高斯核函数

核极限学习机在 ELM 的基础上引入了核函数，通过非线性映射到高维特征空间的方式将线性不可分的问题进行划分，进而提高了 ELM 性能。但由于核函数的引入，使得 KELM 算法对参数的选择非常敏感，所以引入 Xgboost 算法对 KELM 正则化系数 C 和核函数进行寻优，减少人为调参的复杂度，提高模型的准确性。

1.2 Xgboost 优化算法

Xgboost[22-23]优化了梯度提升决策树，将原来针对错误样本分类中权值分配的不断迭代，改变为贪心策略，训练最佳方向是损失函数梯度下降的方向，最后通过加权求和的方法得出。并使用增加正则化优化了损失函数求解的方法，模型复杂度得到一定降低并防止了过拟合现象，模型精度得到一定程度的提升。

损失函数增加正则化，求解最优模型使用结构风险最小的思想[24]设置目标函数来寻优迭代模型，如公式（8）：

$$OBj=\sum_{i=1}^{I}l(Y_i,f_N(X_i))+\sum_{1}^{N}\Omega(h_n(X)) \tag{8}$$

i 为样本索引符号，I 为样本总量，l 为决策树的叶子节点数目，Y_i 为训练样本实际值。根据残差公式 $f_n(X_i)=f_{n-1}(X_i)+h_n(X_i)$，可得出 $f_N(X_i)$ 决策树训练（$N=1,2,\cdots,N$）积累的残差及 $\Omega(h_n(X))$ 针对 $\Omega(h(X))$ 在 n 次训练叶子节点得分，叶子节点得分 ω 用 $L2$ 正则化的表示：

$$\Omega(h_n(X))=\gamma L+\frac{1}{2}\lambda\sum_{l=1}^{L}\omega_l^2,(l=1,2,\cdots,L) \quad (9)$$

使用泰勒展开公式进行求解得到二次项，如式（10）：

$$f(x+\Delta x)\approx f(x)+f'(x)+\frac{1}{2}f''(x)\Delta x^2 \quad (10)$$

并规定：

$$p_i=\partial_{f_{n-1}(X_i)}l(y_i,f_{n-1}(X_i))$$
$$q_i=\partial^2_{f_{n-1}(X_i)}l(y_i,f_{n-1}(X_i)) \quad (11)$$

最终得出目标函数如（12）所示：

$$OBj_n=\sum_{i=1}^{I}\left[p_ih_n(X_i)+\frac{1}{2}q_ih_n^2(X_i)\right]+\Omega(h_n(X) \quad (12)$$

用该目标函数进行模型最优解的求解，可以自己定义损失函数，对于模型构建的灵活性有极大的提升。由于 Xgboost 基于思想通过不断最小化目标函数迭代生成决策树，其预测偏差得到不断降低，综合两方面因素，Xgboost 泛化误差得到整体降低，模型精度势必得到一定程度的提升。

2　模型应用

2.1　预测总流程

本文选取裂缝位移 ΔX、岸坡水文地质条件 H、土壤含水率 D、土压力 ΔF、斜坡倾角 θ、及降雨量 R 6 个滑坡诱发条件作为滑坡预测影响因子，监测数据通过数据预处理作为模型输入数据，通过 Xgboost 优化超参数 C 和并训练出 Xgboost-KELM 模型，分别与 GA 及 GC 优化的 KELM 模型比对，最后通过验证集验证模型预测的结果。文中具体预测路线如图 2 所示。

2.2　样本数据选取

样本数据使用陕西省山阳县 2018 年 3 月到 2020 年 3 月的 10 个监测点数据作为数据集，样本集分为 80％测试集和 20％的两个验证集。模型输入数据：裂缝位移 ΔX、岸坡水文地质条件 H、土壤含水率 D、土压力 ΔF、斜坡倾角 θ、及降雨量 R 6 个滑坡体影响因子。预报模型训练及验证数集总共 1 280 组数据集。

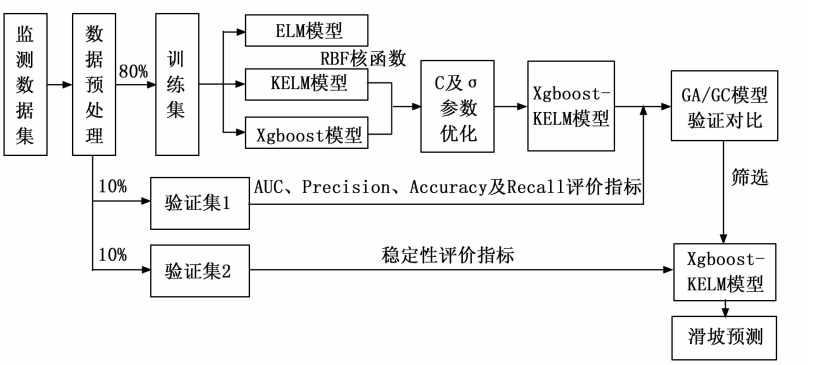

图 2　预测路线图

表 1　监测数据

监测点	样本编号	裂缝位移 ΔX /cm	岸坡水文地质条件 H/％	土壤含水率 D/％	土压力 ΔF / kPa	斜坡倾角 θ/％	降雨量 R /(mm /24 h)
1～10	1	2.80	0.24	10.1	9.2	0.25	12.1
	2	4.99	0.36	10.7	3.7	0.32	12.2
	3	5.7	0.47	10.0	6.7	0.39	22.1
	4	6.8	0.72	10.3	8.2	0.58	43.4
	5	10.2	0.41	11.2	8.6	0.43	102.7
	6	14.2	0.45	12.8	12.1	0.45	105.5
	……	……	……	……	……	……	……
	1 275	18.1	0.45	38.2	14.5	0.56	22.2
	1 276	19.4	0.56	42.5	19.4	0.89	46.5
	1 277	22.5	0.49	37.2	31.3	0.73	103.3
	1 278	22.5	0.63	37.3	26.7	0.88	136.2
	1 279	26.4	0.71	36.4	32.1	0.88	88.1
	1 280	26.6	0.79	38.7	35.1	0.92	68.7

2.3　数据预处理

监测数据来源于不同的采集传感器，由于环境的影响会出现一些缺失、离群或维度不统一的数据，对于模型的建立有极大的消极影响，因此需要对监测数据进行预处理。

2.3.1　异常值处理

监测数据中存在一部分偏离传感器本身范围的值或者偏离观测值较大的值，如果不处理会影响数据本身预测的准确性，如果距离达到 5 倍或者相距均值距离≥3 倍标准差的数据为离群点。

2.3.2　缺失值的处理

监测数据通过多传感器传输，传输过程中经常会出现遗漏或者离群点情况，会损失有效信息，导致属性值确实。按照属性因素方法进行统计得出缺失率 q，本文划分两种类别数据的缺失值，如表 2 所示。

表 2　数据缺失值

类型	类别型	数值型
$q\geqslant 90\%$	缺失值属性剔除	缺失值属性剔除
$40\%\leqslant q<90\%$	缺失值属性作为一种新的类别	相邻属性加权值填充
$20\%\leqslant q<40\%$	多重插补	均值填充
$q<20\%$	同类均值插补	众数填充

2.3.3　数据归一化

采集的监测数据种类及数量都较大，多传感器数据量纲不同有较大的差异，原始数据直接建模对于预测的准确性影响极大，归一化处理公式如下：

$$R=\frac{R-R_{\min}}{R_{\max}-R_{\min}} \quad (13)$$

式中，R 为某因素归一化处理后的数据 $R_{\min}$，和 $R_{\max}$ 为某因素数据中最小值及最大值。图 3 为选取 4 个传感器中部分监测数据数据归一化前后的数据分布图，图中可以看出归一化前的数

据跨度分布比较大，归一化后的数据在［0～1］量纲内，避免了数据本身量纲问题对预测模型的影响。

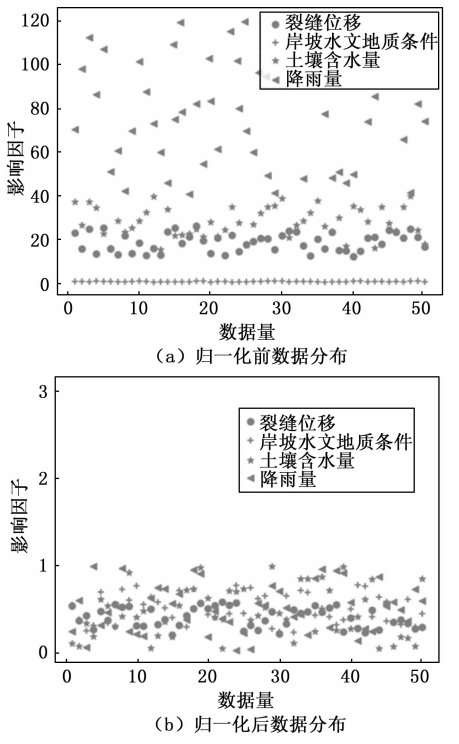

图 3 监测数据归一化分布

2.4 Xgboost 寻优的 KELM 模型

选取高斯径向基 RBF 核函数，根据（6）和（7）两式可知需要优化的参数有正则化系数 C 和高斯径向基 RBF 核函数中的核参数 σ，其中 $C\in$（0，a］，$\sigma\in$（0，b］。正则化系数 C 对模型方差、偏差、训练误差及测试误差都有较大的影响，具体影响如表 3 所示。核参数 σ 主要影响模型的拟合程度，当 $\sigma>0$ 且较小时容易出现过拟合，相反 $\sigma>0$ 较大时容易出现欠拟合。Xgboost 优化的核极限学习机就是利用 Xgboost 优化算法对 KELM 中的参数进行择优选取，从而提升 KELM 的性能，提高分类准确度，限制模型复杂度，缓解模型过拟合问题。为了避免多次枚举造成运算量过大，采用贪心算法寻求最优数结构，当 Gain 信息增益达到树的深度限制或 Gain＜0 数停止分割，防止过拟合的前提下达到速度快拟合效果好。

表 3 正则化参数 C 对模型指标影响

正则化参数 C	模型复杂度	方差	偏差	训练误差	测试误差
大	低	低	高	高	高
中	中	中	中	中	低
小	高	高	低	低	高

寻优具体流程如下：

STEP 1：数据标准化处理，划分训练样本；

STEP 2：初始化 Xgboost 算法参数；

STEP 3：Xgboost 损失函数使用 KELM 模型中均方误差（MSE，mean-square error）代替，选择 Xgboost 目标函数下降最大点作为最佳切分点，对 RBF 核函数参数 σ 及正则化系数 C 进行初始化；

STEP 4：将样本特征顺序排列，列出所有划分特征、特征值及 score 分值，根据 TOP1 分裂子树，同时计算分割的叶子节点的权重向量及信息增益；

STEP 5：判断是否达到数的深度限值或增益小于 0，更新并保存最终叶子节点的权重值与增益值；

STEP 6：判断最大迭代次数的条件是否达到，如果满足条件则确定此时的 σ 及 C 参数为最优参数，基于最优参数构建 Xgboost-KELM 最优模型；不满足条件则返回执行 STEP 4。

Xgboost 参数设置：

1）通用参数

booster 基学习器：*gbtree*（树模型）；

多线程：*nthread*；

迭代次数 *nround*：1 000。

2）Booster 初始参数

eta：0.1，通过调整学习率调整模型最佳收敛速度；

min _ child _ weight：0，设置最小叶子节点样本的权重和避免出现过拟合现象；

max _ depth：8，通过改变树的最大深度控制子节点分裂，避免出现过拟合现象；

gamma：节点分裂的依据，后损失函数下降值大于该值分裂；

subsample：0.7，对树随机行比例划分采样控制；

n _ estimator：120，控制最大树的数量，防止数量大过拟合，数量小欠拟合；

lambda：1，L2 正则化项；

Alpha：1，L1 正则化项。

3）目标参数

objective：*multi*：*softmax* 多分类问题；

eval _ metric：［*error*，*auc*，*mse*］，回归模型。

2.5 仿真验证及结果分析

使用训练集进行训练 Xgboost 模型，当 Xgboost 找寻最优的分裂节点时，可以基于 KELM 损失函数迭代确定 Xgboost 最佳参数。图 4 为模型筛选最佳参数过程曲线图，a、b 表示最大树数量与分类准确率关系，a 为整个范围最大树数量分类过程，最大树数量 *n _ estimator* 范围［1～1 000］，变化曲线先逐渐增加，增加至 92 棵树时上升缓慢，之后提升不明显渐渐趋于稳定；b 为 90～150 棵树分类过程，随之 *n _ estimator* 数量增加，会存在不到 1% 的下降趋势，当数量达到 102 棵树时分类准确率达到最高 86.34%，取该数量为模型 *n _ estimator* 最佳参数。c 为树最大深度分类变化过程，*max _ depth* 范围为［1～10］，迭代过程可以看出深度为 6 时分类准确度最高达 86.42%，*max _ depth* 最佳参数值取 6；d 为最小叶子权重变化过程，叶子节点的权重小于 *min _ child _ weight* 则停止拆分树，*min*

_child_weight 范围为［0～10］，曲线变化过程可以得出权重为 2 时分类准确率达到最高 88.25%，*min_child_weight* 最佳值取 2。根据训练集训练结果得出 Xgboost 最终参数如表 4 所示。

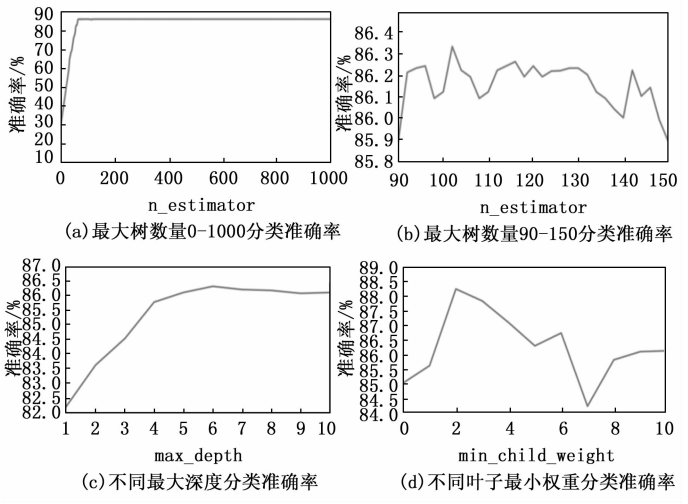

图 4　不同参数分类准确率

表 4　Xgboost 模型最终参数

参数	参数值
booster	*gbtree*
objective	*multi:softmax*
eval_metric	*mse*
lambda	8
Alpha	0.5
n_estimator	102
max_depth	6
subsample	0.7
gamma	1
eta	0.1
min_child_weight	2

为了获取 KELM 最优核参数 σ 及正则化 C 参数最佳组合，将测试集随机分为 4 组作为训练样本建立模型，参数的选值先从一个比较大的区间范围进行搜索，4 组测试集验证后得出小范围［500，1 000］×［0.2，0.3］，之后进行多次迭代训练，当均方误差达到最小值时，取此时参数核参数 σ 及正则化 C 为最佳组合参数。图 5 为 4 组测试集训练过程的均方误差迭代曲线，数据集 4 收敛速度最慢，数据集 2、3 均方误差较低但收敛速度在迭代次数 12 出现饱和，而数据集 3 均方误差及收敛速度相比其他数据集表现最佳，均方误差最低达到并稳定于 1.187×10^{-3}。且在 7 次迭代收敛饱和。所以选取数据集 3 训练参数为模型最佳参数，$\sigma=0.262\ 4$，$C=703.24$。

为验证模型的稳定性及适应能力，通过模型在新样本集中的适应度、准确性及方差偏差验证。方差表示模型每次预期结果与实际结果的误差的稳定情况；偏差值表示每次预期结果与实际结果的偏差。模型误差包括方差、偏差及其他无法避免误差，图 6 为偏差及方差示意图。预测模型最佳选择顺序：1）方差小，偏差小；2）方差小，偏差大；3）方差大，偏差小；4）方差大，偏差大。

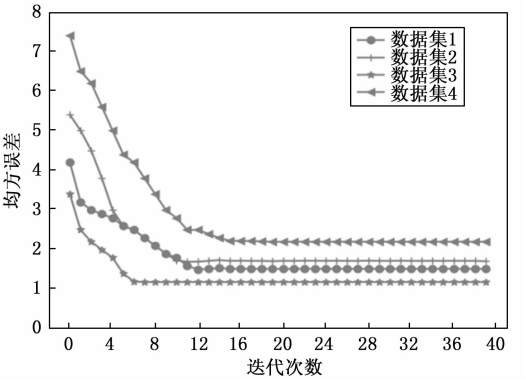

图 5　不同测试集的均方误差

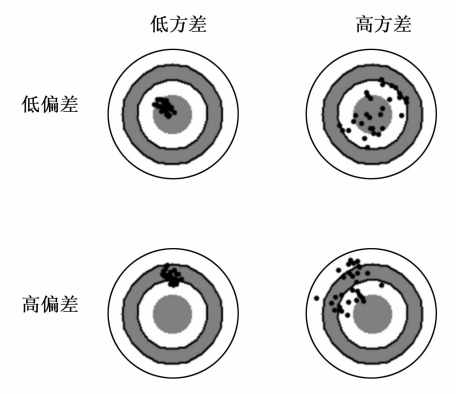

图 6　偏差及方差示意图

精确率：在预期的正样本中实际结果也为正样本的占比。

$$precision = TP/(TP+FP) \tag{14}$$

准确率：准确率表示所有的预测样本中，预测正确的占比。

$$Accuracy = (TP+TN)/(TP+FP+FN+TN) \tag{15}$$

召回率：预测结果准确的正样本占所有正样本的比例。

$$recall = TP/(TP+FN) \tag{16}$$

AUC：通过计算 *ROC* 曲线与坐标轴围成的面积得到，介于［0.5，～1］之间，预测的真实性取决与 *AUC* 值接近 1 的程度，靠近 1 真实性高反之则反。

$$AUC = 1-\frac{\frac{FP}{FN+TN}-\frac{FN}{TP+FP}}{2} \tag{17}$$

真正率：预期正样本数/实际正样本数。

$$TPR = \frac{TP}{TP+FN} \tag{18}$$

假正率：预期为正的负样本数/实际负样本数。

$$FPR = \frac{FP}{FP+TN} \tag{19}$$

实验采用 KELM 作为滑坡灾害预测模型，并用 Xg-

boost 优化算法中的超参数寻优。使用同一个验证集验证 GC 及 GA 优化 KELM 模型预测效果比对。表 5 为在验证集 1 中 128 个数据中各模型输入相同数据得出的预测结果的混淆矩阵对比表。

表 5 不同模型混淆矩阵对比表

预测模型	GA-KELM	GC-KELM	Xgboost-KELMTP
TP(真阳)	115	118	121
FN(假阴)	13	10	7
FP(假阳)	6	4	2
TN(真阴)	10	13	16

图 7 为 3 种模型同一评价指标的对比图，柱状图的差异可以对比得出各个预测模型对应评价指标的好坏，从而反映预测模型性能的优劣。4 个评估参数中，Xgboost 优化 KELM 均明显优于 GA 和 GC 优化的模型。本文的 Xgboost-KELM 模型 AUC 均值为 0.985，相比 GC 优化高约 3 个百分点，比 GA 优化高 6 个百分点。其他指标 Precision、Accuracy 及 Recall 高于另外两个模型百分比［1.7～7］范围之间。实验结果说明 Xgboost-KELM 模型具有较好的预测效果，在滑坡灾害预测中有较好的预测能力。

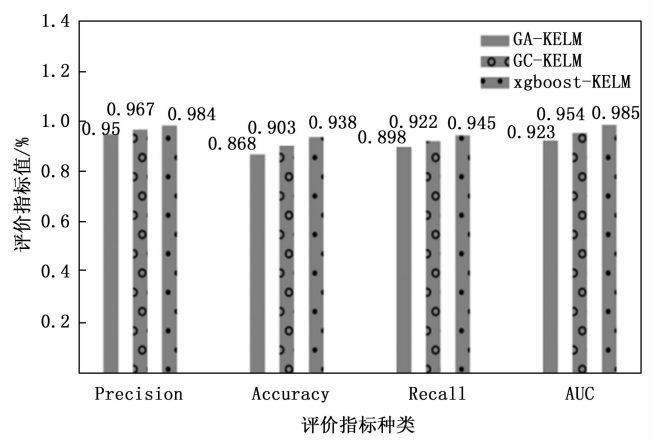

图 7 评价指标对比图

图 8 为验证集 2 中打乱随机抽取的 128 个监测数据使用 Xgboost 优化 KELM 模型后的实际发生概率与预测发生概率比对图。图中实际值与预测值基本吻合，拟合情况较好。极限的几个数据 27、38、89 及 111 发生概率存在一些差异，但是对应风险等级都属于同等级风险。准确率达到 98%。引入 Xgboost 对 KELM 参数值进行优化，实际预测精度较为理想。并使用验证集 2 进行各模型稳定性计算，Xgboost-KELM 有较小的方差且偏差也较小，属于最稳定的“方差小、偏差小”模型，稳定性能最强，而 GC-KELM 介于“方差大、偏差小”与“方差大、偏差大”之间，模型不稳定，GA-KELM 属于“方差大、偏差大”，模型最不稳定。

3 结束语

本文针对山阳县研究区域的数据使用基于 Xgboost 优化 KELM 模型建立滑坡灾害预测模型，通过仿真研究分析

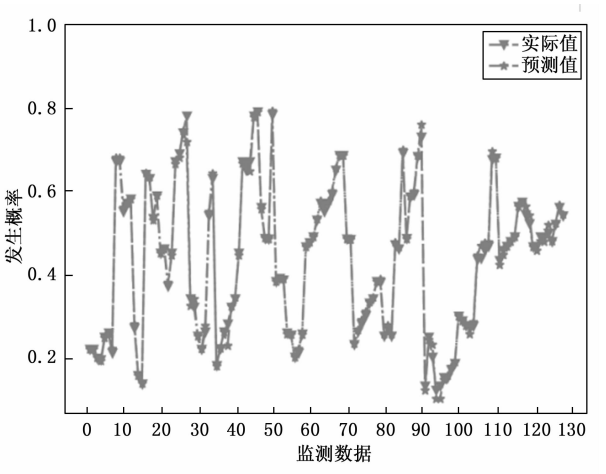

图 8 Xgboost 优化模型预测结果

滑坡影响因子与发生概率之间关系，并与 GA-KELM 及 GC-KELM 建模结果进行比较。

1）选用 RBF 高斯核函数的极限学习机模型较好地解决了 ELM 适用性及稳定性不佳的问题；

2）使用 KELM 模型中均方误差 *MSE* 作为损失函数，训练模型并选择目标函数下降最大点作为最佳切分点，确定最佳参数；并使用 Xgboost 寻优算法对核函数中的正则化系数 C 和核函数 σ 寻优，通过 4 组测试集的均方误差迭代曲线得出最佳超参数，建立 Xgboost-KELM 模型，并与 GA 及 GC 优化 KELM 建模进行比较；

3）通过几种模型的样本集 1 对比验证，结果表明 Xgboost-KELM 具有较高的 Precision、Accuracy 及 Recall 和 AUC 值，同时使用新样本集 2 验证稳定性及泛化能力，结果表明该模型稳定性较好，进一步证明该模型的应用能有效提高滑坡灾害的预报概率，对滑坡灾害提前预警，降低自然灾害造成的损失具有重要意义。

参考文献：

[1] SHOFIYATUL A，HENI M，INDRIA K，et al. Landslide disaster risk strategy：lesson learned from the community in the northwest part of Bromo Volcano Flank［C］// International Conference on Environmental and Energy Policy（ICEEP 2021），2021.

[2] ANIK S，SUNIL S. Integrating the artificial intelligence and hybrid machine learning algorithms for improving the accuracy of spatial prediction of landslide hazards in Kurseong Himalayan region［J］. Artificial Intelligence in Geosciences，2022，3（1）：14－27.

[3] 周 萍，邓 辉，张文江，等. 基于信息量模型和机器学习方法的滑坡易发性评价研究——以四川理县为例［J］. 地理科学，2022，42（9）：1665－1675.

[4] 张志兼，黄 勋，蔡雨微，等. 三峡库区武隆段滑坡灾害驱动因子演变格局与人类活动的影响［J］. 中国地质灾害与防治学报，2022，33（3）：39－50.

[5] 梁永平，赖国泉，严丽萍. 无人机低空遥感技术在滑坡应急测

绘及治理中的应用实践［J］. 测绘与空间地理信息，2022，45（5）：24－25，31.
[6] 孙　东，殷志强，李大猛，等. 美姑河流域地质构造对大型滑坡孕育的控制作用［J］. 中国地质灾害与防治学报，2019，30（6）：49－58，67.
[7] 崔圣华，裴向军，黄润秋，等. 汶川地震黄洞子沟右岸大型滑坡地质构造特征及成因［J］. 工程地质学报，2019，27（2）：437－450.
[8] 田述军，付国训，程小松. 汶川大地震同震滑坡复活变形特征研究［J］. 灾害学，2023，38（1）：50－56.
[9] 王　欣，方成勇，唐小川，等. 泸定 Ms 6.8 级地震诱发滑坡应急评价研究［J］. 武汉大学学报（信息科学版），2023，48（1）：25－35.
[10] ZHANG Y G，TANG J，CHENG Y M，et al. Prediction of landslide displacement with dynamic features using intelligent approaches［J］. International Journal of Mining Science and Technology，2022，32（3）：539－549.
[11] 赵晓萌，卫星君，王　娜，等. 降雨型滑坡灾害的特征聚合决策树预测模型［J］. 灾害学，2020，35（1）：27－31.
[12] 赵彬如，陈恩泽，戴　强，等. 基于水文-气象阈值的区域降雨型滑坡预测研究［J］. 测绘学报，2022，51（10）：2216－2225.
[13] 胡　欣，熊帮彬，王会峰，等. 基于 SVM-BP 降雨型黄土滑坡灾害安全评价模型研究［J/OL］. 中国公路学报：1－13［2022-11-28］.
[14] 马　娟，赵文祎，齐　干，等. 基于普适型监测的多参数预警研究——以三峡库区卡门子湾滑坡为例［J］. 西北地质，2021，54（3）：259－269.
[15] 杨宗佶，王礼勇，石莉莉，等. 降雨滑坡多指标监测预警方法研究［J］. 岩石力学与工程学报，2020，39（2）：272－285.
[16] 李丽敏，郭　伏，温宗周，等. 基于长短时记忆与多影响因子的滑坡位移动态预测［J］. 科学技术与工程，2020，20（33）：13559－13567.
[17] 杨忠平，李绪勇，赵　茜，等. 关键影响因子作用下三峡库区堆积层滑坡分布规律及变形破坏响应特征［J］. 工程地质学报，2021，29（3）：617－627.
[18] GAO H，HE L，HE Z W，et al. Early landslide mapping with slope units division and multi-scale object-based image analysis—A case study in the Xianshui River basin of Sichuan，China［J］. Journal of Mountain Science，2022，19（6）：1618－1632.
[19] 欧阳高明，宋加平，李　灿，等. 基于 ELM 人工神经网络的滑坡失稳预测模型［J］. 广东水利水电，2022（8）：12－16，23.
[20] 高　峰，吴晓东，周科平. 基于主成分分析和 PSO-ELM 算法的排土场稳定性预测模型［J］. 黄金科学技术，2021，29（5）：658－668.
[21] 孙世政，刘照伟，张　辉，等. 基于 HHO-KELM 的 FBG 流量温度复合传感解耦［J］. 光学精密工程，2022，30（11）：1290－1300.
[22] 史佳琪，张建华. 基于多模型融合 Stacking 集成学习方式的负荷预测方法［J］. 中国电机工程学报，2019，39（14）：4032－4042.
[23] 赵晓东，徐振涛，刘　福，等. 基于极端梯度提升算法的滑坡易发性评价模型［J］. 科学技术与工程，2022，22（23）：10347－10354.
[24] 林报嘉，刘晓东，杨　川，等. XGBoost 机器学习模型与 GIS 技术结合的公路崩塌灾害易发性研究［J］. 公路，2020，65（7）：20－26.

（上接第 224 页）

[19] 白洪涛，栾　雪，何丽莉，等. 基于缺失森林的医疗大数据缺失值插补［J］. 吉林大学学报（信息科学版），2022，40（4）：616－620.
[20] 李昕聪，刘俊岩，张启元，等. 融合自适应滤波和归一化 PGC－Arctan 的激光干涉测振信号解调算法研究［J］. 电子测量技术，2022，45（13）：115－122.
[21] 谭　阳，武小红，武　斌，等. GK 可能 C 均值模糊聚类的白菜红外光谱分析［J］. 光谱学与光谱分析，2022，42（5）：1465－1470.
[22] 林广朋. 基于 LDA 模型的网络信息内容安全分类系统设计［J］. 长江信息通信，2022，35（7）：53－55.
[23] 王孟妍，崔学荣，张国平，等. 基于 RSSI 的 WiFi 指纹定位数据融合算法研究［J］. 微型电脑应用，2020，36（3）：1－3.
[24] 柏语蔓，于莲芝. 基于象群-蚁群算法改进的小车路径规划［J］. 智能计算机与应用，2021，11（12）：179－183，189.
[25] 吕　响，张书玉，宋英楠，等. 基于深度学习下的卷积神经网络参数学习［J］. 渤海大学学报（自然科学版），2021，42（4）：369－375.
[26] 刘艳琪，刘一杰. 基于病毒侵染和逆转操作的改进遗传算法［J］. 湖南文理学院学报（自然科学版），2022，34（3）：23－29.
[27] 夏　奎，李　炜，邱意敏，等. 基于改进型象群优化算法的 BSS 方法［J］. 电子测量与仪器学报，2021，35（10）：153－160.
[28] 张子建，王宏伟，周怀芳，等. 基于多机制混合象群算法的混沌系统参数估计［J］. 微电子学与计算机，2020，37（6）：40－45.
[29] 郑钦元，赵乃东. 四重素数 RSA 非对称加密算法的研究与实现［J］. 网络安全技术与应用，2022（5）：38－40.
[30] 邓一新. 基于全同态加密算法的医院财务数据安全存储系统［J］. 自动化技术与应用，2022，41（7）：44－47.
[31] 魏　伟，陈佳哲，李　丹，等. 椭圆曲线 Diffie-Hellman 密钥交换协议的比特安全性研究［J］. 电子与信息学报，2020，42（8）：1820－1827.
[32] 罗予东，陆　璐. 基于人工神经网络和遗传算法的网络攻击检测［J］. 计算机工程与设计，2021，42（9）：2446－2454.
[33] 李　欣，俞卫琴. 基于统计信息聚类边界的不平衡数据分类方法［J］. 计算机工程与设计，2021，42（8）：2218－2223.