DOI:10.20040/j.cnki.1000-7709.2023.20231076

# 基于多层次数据处理的 NGO-XGBoost 大坝变形预测模型及其应用

李晨阳[a,b],郑东健[a,b]

(河海大学 a. 水利水电学院; b. 水文水资源与水利工程科学国家重点实验室, 江苏 南京 210098)

**摘要**: 混凝土坝变形序列中存在的噪声和非线性特征严重影响了大坝变形预测的精度。为此,采用集合经验模态分解(EEMD)对坝体水平位移信号进行分解,挖掘其中有效变形信息;并利用奇异谱分析(SSA)对分解所得高频本征模态分量(IMF)进行特征提取,以减少有效信息的丢失。考虑到效应量与环境量之间复杂的随机性和非线性映射关系,采用极限梯度提升(XGBoost)对降噪后的数据建模预测;考虑到 XGBoost 超参数对模型预测性能的显著影响,引入全局搜索能力较好的北方苍鹰算法(NGO)对其参数寻优,构建了基于 NGO-XGBoost 的大坝位移预测模型。计算结果表明,EEMD-SSA 能有效地去除大坝位移监测信息中的噪声,NGO-XGBoost 模型显著提高了大坝位移预测模型的精度。

**关键词**: 大坝变形预测;集合经验模态分解;奇异谱分析;北方苍鹰算法;极限梯度提升

**中图分类号**: TV698.1 **文献标志码**: A **文章编号**: 1000-7709(2023)11-0077-05

## 1 引言

由于大坝变形结构的特殊性及作用荷载的随机性,大坝变形数据中常常表现出非线性特征[1-2],这严重限制了统计回归模型的精度。随着机器学习算法的不断发展,各种新兴算法被运用到大坝变形监控领域,改善了统计模型的预测效果。徐韧等[3]采用基于高斯过程的贝叶斯优化方法对 XGBoost 算法进行参数优化,构建了基于 GP-XGBoost 的大坝变形预测模型,有效地提高了模型的学习效率与预测精度;董泳等[4]采用了 EMD 和 EEMD 对大坝位移序列进行降噪,并建立了基于长短时记忆网络的大坝位移预测模型。然而,大坝实测的位移监测数据中存在许多环境因素导致的干扰,与真实位移混杂在一起,从而导致了监测数据质量较差和模型精度不高的问题。为此,本文采用 EEMD 对坝体水平位移信号分解,并利用奇异谱分析(SSA)对分解所得高频 IMF 进一步提取特征,引入具有良好全局搜索能力的 NGO 对 XGBoost 参数寻优,构建了基于 NGO-XGBoost 的大坝变形预测模型,工程实例应用结果表明,本文模型能有效地去除监测数据中的噪声,显著提高了大坝位移预测模型的精度。

## 2 基于 EEMD-SSA 的多层次数据处理方法

针对 EMD 分解模态混叠的问题,EEMD[5]通过逐次加入同等幅值的白噪声来改变序列的极值点特性,对多次 EMD 分解得到的相应模态的 IMF 平均处理来抵消加入的白噪声,从而有效抑制模态混叠的产生。

SSA 算法能根据时间序列数据构造出轨迹矩阵,通过对该矩阵进行分解、重构,提取出原始序列中的趋势、振荡分量和噪声等,从而分析时间序列的结构。

直接采用 EEMD 对时序数据降噪时会将所有高频 IMF 当作噪声去除,从而在保留原始信号趋势特征的同时滤除其中大部分毛刺、尖锐部分以达到降噪的目的。但高频 IMF 中通常包含大量原始序列的细节特征信息,直接去除全部高频 IMF 会损失很多有效信息,使降噪后的序列失去原序列的特殊性。为充分挖掘高频分量中的有效

**收稿日期**: 2023-06-29,**修回日期**: 2023-07-30

**基金项目**: 国家自然科学基金项目(52179128)

**作者简介**: 李晨阳(1999-),男,硕士研究生,研究方向为水工结构安全监测,E-mail:353831849@qq.com

**通讯作者**: 郑东健(1965-),男,博士、教授、博导,研究方向为水工结构安全与健康诊断,E-mail:zhengdj@hhu.edu.cn

信息，本文使用 SSA 对 EEMD 分解后的高频 IMF 进行数据特征再处理，并重构得到有效去噪分量，从而在去除噪声的同时保留高频特征信息，提高降噪序列的准确性。

在将大坝变形监测信号进行 EEMD 分解后采用连续均方差（CMSE）方法[6]对分解得到的 IMF 信号属性进行判别，以对高频 IMF 做进一步挖掘。CMSE 是相应 IMF 分量在时域中各时间点上振幅的平方和，相当于 IMF 分量的能量密度。该算法通过计算相邻 IMF 分量误差，当误差在第 $k$ 个信号上出现突变时，就作为信号和噪声的分离点。

CMSE 的计算方法为：

$$\delta_{\mathrm{CMSE}}(\widetilde{x}_k,\widetilde{x}_{k+1})=\frac{1}{n}\sum_{i=1}^{n}[\widetilde{x}_k(t_i)-\widetilde{x}_{k+1}(t_i)]^2=\frac{1}{n}\sum_{i=1}^{n}[c_k(t_i)]^2 \tag{1}$$

$$\widetilde{x}_k(t_i)=\sum_{j=k}^{n}c_j(t)+r_n(t)\quad k=2,3,\cdots,N-1 \tag{2}$$

式中，$\delta_{\mathrm{CMSE}}(\cdot)$为连续均方差值；$\widetilde{x}_k(t)$ 为部分重构后的信号；$c_k(\cdot)$为 IMF 分量；$r_n(\cdot)$为残差项；$n$ 为信号长度；$N$ 为 IMF 的个数。

当利用式(1)、(2)求得所有连续重构信号之间的均方误差后绘制连续均方差图，曲线斜率突变点即为高频和低频 IMF 的分界点。

在区分 IMF 的频率属性后将所有高频 IMF 叠加得到总体高频分量作为 SSA 的原始时间序列输入，从而进一步挖掘其中残留的原始信号的趋势特征信息并重构得到有效分量。之后再将经 SSA 挖掘去噪后的总体高频分量与 EEMD 分解得到的所有低频 IMF 重构，最终得到数据特征深度再挖掘后重构的去噪数据。

# 3 NGO 优化 XGBoost 大坝变形预测模型

## 3.1 XGBoost

XGBoost 通过对损失函数二阶泰勒展开并引入正则项，有效地缓解了过拟合和低效率等问题。XGBoost 的原理可用下式来表示：

$$\hat{y}_i=\phi(x_i)=\sum_{k=1}^{K}f_k(x_i) \tag{3}$$

式中，$\hat{y}_i$ 为预测值；$\phi(\cdot)$为 $y$ 与 $x$ 之间的映射函数；$f_k$ 为第 $k$ 棵树的权重函数。

其目标函数 $L(\theta)$由损失函数 $l(y_i,\hat{y}_i)$和正则项 $\Omega(f_k)$组成：

$$L(\theta)=\sum_{i=1}^{N}l(y_i,\hat{y}_i)+\sum_{k=1}^{K}\Omega(f_k) \tag{4}$$

$$\Omega(f_k)=\gamma T+\lambda\|w\|^2/2 \tag{5}$$

式中，$\gamma$ 为决策树叶节点复杂性；$T$ 为决策树的叶节点数量；$\lambda$ 为惩罚系数；$w$ 为其权重。

XGBoost 模型采用梯度提升方法对目标函数进行优化。目标函数 $L$ 在第 $t$ 次迭代下的二阶泰勒展开式可表示为：

$$L^{(t)}\simeq\sum_{i=1}^{N}[l(y_i,\hat{y}_i{}^{(t-1)})+g_if_t(x_i)+h_if_t{}^2(x_i)/2]+\Omega(f_t)+\mathrm{constant} \tag{6}$$

其中

$$g_i=\partial_{\hat{y}^{(t-1)}}l(y_i,\hat{y}_i{}^{(t-1)});h_i=\partial^2_{\hat{y}^{(t-1)}}l(y_i,\hat{y}_i{}^{(t-1)})$$

式中，$g_i$ 为一阶导数；$h_i$ 为二阶导数；constant 为常数项。

将式(6)中的常数项全部移除后可得到 XGBoost 的目标函数：

$$L^{(t)}\simeq\sum_{i=1}^{N}\left[g_if_t(x_i)+\frac{1}{2}h_if_t{}^2(x_i)\right]+\Omega(f_t) \tag{7}$$

## 3.2 北方苍鹰算法 NGO

NGO 算法模拟了北方苍鹰捕猎过程，其过程主要分为猎物识别和攻击及追击和逃亡两大阶段。

### 3.2.1 猎物识别和攻击阶段

在最初的狩猎阶段，苍鹰随机发现目标然后迅速攻击。由于在搜索空间中随机选择猎物，这一阶段提高了算法的搜索能力。该阶段包括搜索空间的全局搜索以发现最佳区域，可表示为：

$$P_i=X_k\quad i=1,2,\cdots,N;k=1,2,\cdots,i-1,i+1,\cdots,N \tag{8}$$

$$x_{i,j}^{\mathrm{new},P1}=\begin{cases}x_{i,j}+r(P_{i,j}-Ix_{i,j}) & F_{P_i}<F_i\\ x_{i,j}+r(x_{i,j}-P_{i,j}) & F_{P_i}\geqslant F_i\end{cases} \tag{9}$$

$$X_i=\begin{cases}X_i^{\mathrm{new},P1} & F_i^{\mathrm{new},P1}<F_i\\ X_i & F_i^{\mathrm{new},P1}\geqslant F_i\end{cases} \tag{10}$$

式中，$P_i$ 为第 $i$ 只苍鹰对应的猎物位置；$F_{P_i}$ 为其目标函数值；$k$ 为从 $[1,N]$ 中随机选取的自然数；$X_i^{\mathrm{new},P1}$ 为所提出解决方案的第 $i$ 个更新状态，其第 $j$ 个维度为 $x_{i,j}^{\mathrm{new},P1}$；$F_i^{\mathrm{new},P1}$ 为算法初始阶段的目标函数值；$r$、$I$ 分别为搜索、更新时的随机数，且 $r\in[0,1]$，$I=1$ 或 2。

### 3.2.2 追捕和逃亡阶段

苍鹰袭击猎物后，猎物会试图逃跑，而苍鹰会继续沿猎物踪迹追击猎物。该阶段算法的表达式为：

$$x_{i,j}^{\text{new},P2}=x_{i,j}+R(2r-1)x_{i,j} \tag{11}$$

$$R=0.02(1-t/T) \tag{12}$$

$$X_i=\begin{cases}X_i^{\text{new},P2} & F_i^{\text{new},P2}<F_i\\ X_i & F_i^{\text{new},P2}\geqslant F_i\end{cases} \tag{13}$$

式中，$t$ 为当前迭代次数；$T$ 为最大迭代次数；$x_{i,j}^{\text{new},P2}$ 为第 $i$ 只苍鹰在新位置的解在第 $j$ 维度的值；$R$ 为攻击范围的半径；$F_i^{\text{new},P2}$ 为该阶段的目标函数值。

种群成员根据式(7)～(12)不断更新，直到完成最后一次迭代。

### 3.3　模型构建流程

综上所述，构建基于 EEMD-SSA 的数据降噪方法的 NGO-XGBoost 混凝土坝位移预测模型，提高混凝土坝的变形预测精度，具体建模流程如下。

**步骤 1**　通过 EEMD 将混凝土坝位移监测序列分解为不同频率的 IMF 信号，并利用 CMSE 确定高低频 IMF 的分界点。

**步骤 2**　采用 SSA 对 EEMD 分解后的高频 IMF 进行数据特征深度再挖掘，并重构得到去噪分量，再将各分量进行叠加得到处理后的混凝土坝变形序列 $y^*$。

**步骤 3**　定义模型输入与输出。输入为各环境量因子，输出为位移序列，即 $x=[H^1,H^2,H^3,\sin\frac{2\pi t}{365}-\sin\frac{2\pi t_0}{365},\cos\frac{2\pi t}{365}-\cos\frac{2\pi t_0}{365},\sin\frac{4\pi t}{365}-\sin\frac{4\pi t_0}{365},\cos\frac{4\pi t}{365}-\cos\frac{4\pi t_0}{365},\theta-\theta_0,\ln\theta-\ln\theta_0]$，其中，$H^1$、$H^2$、$H^3$ 均为上游水位表征的水压因子；$\sin\frac{2\pi t}{365}-\sin\frac{2\pi t_0}{365}$、$\cos\frac{2\pi t}{365}-\cos\frac{2\pi t_0}{365}$、$\sin\frac{4\pi t}{365}-\sin\frac{4\pi t_0}{365}$、$\cos\frac{4\pi t}{365}-\cos\frac{4\pi t_0}{365}$ 为表征温度效应的谐波因子；$\theta-\theta_0$、$\ln\theta-\ln\theta_0$ 分别为时效因子、对数时效因子。

**步骤 4**　通过 NGO 算法优化 XGBoost 的超参数，采用最优超参数建模分析。

**步骤 5**　通过与常用大坝位移预测模型对比分析，验证本文所提模型的优越性。

## 4　工程实例应用

### 4.1　数据获取与处理

以某混凝土重力坝 EX53 测点监测数据为例，来验证本文提出的大坝变形分析模型的有效性和准确性，解释该模型在大坝变形建模方面的优越性。

该混凝土坝位于福建省闽江干流上，其最大坝高 101 m，坝顶高程 75 m，正常蓄水位 65 m。为了便于建模，需对连续监测序列进行分析。故选取 2017 年 10 月～2021 年 12 月的监测资料进行建模分析，取 2017 年 10 月～2020 年 9 月期间的 1 082 组数据为训练集，2020 年 10 月～2021 年 6 月期间的 273 组数据为测试集。EX53 测点位移曲线和水位变化曲线见图 1。

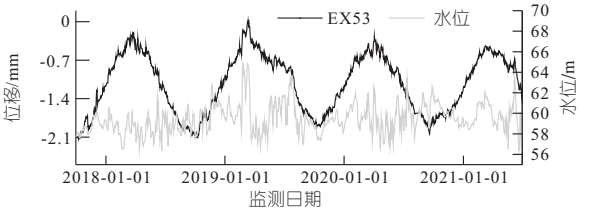


**图 1　EX53 测点水位与位移曲线**

**Fig. 1　Water level and displacement curves of EX53 point**

### 4.2　基于 EEMD-SSA 的多层次数据降噪处理

通过 EEMD 方法将大坝原始变形监测数据分解为 6 个不同频率的 IMF，分解结果见图 2。CMSE 变化曲线呈“手肘”形状，肘部对应的 IMF 即为高低频 IMF 分量的分界点。肘部位于 $IMF_4$ 处，即 $IMF_1$～$IMF_4$ 为高频模态分量，$IMF_5$～$IMF_6$ 为低频模态分量。

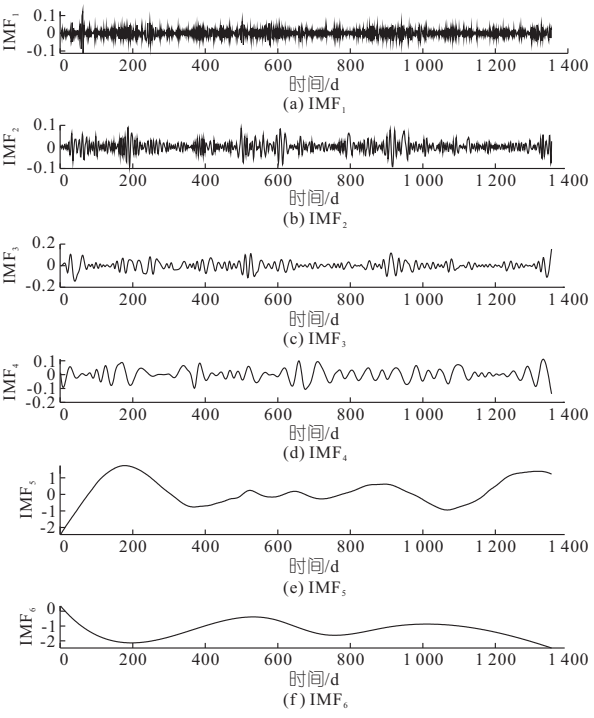


**图 2　EEMD 分解结果**

**Fig. 2　EEMD decomposition results**

利用 SSA 对高频分量 $IMF_1$～$IMF_4$ 进行数据再挖掘以提取高频分量中的有效信息，从而在

尽可能去除噪声的同时保留数据有效特征信息，避免失去序列中的部分特征信息。通过 SSA 将高频 IMF 分解为不同子序列，取贡献率阈值 $k_e$ 为 90%，当前 $n$ 个子序列特征值的累计占比超过 $k_e$ 时，其余子序列可视为噪声部分，对前 $n$ 个子序列进行重构，进而得到经 SSA 挖掘后的有效特征信息。重构 EEMD 分解得到的低频分量和 SSA 挖掘得到的有效特征信息，从而获得经 EEMD-SSA 的多层次降噪处理后的数据。图 3 为 EEMD-SSA 多层次降噪与 EEMD 降噪(即直接去除分解后得到的高频 IMF)的对比。为了更加清晰直观地对比两种降噪方法，分别计算两种方法与原始序列之间的均方根误差($R_{RMSE}$)和皮尔逊相关系数($r$)，结果见表 1。

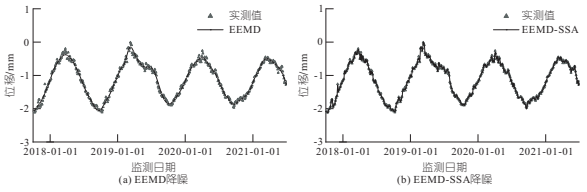


**图 3 降噪效果对比**

**Fig. 3 Comparison of noise reduction effect**

**表 1 EEMD-SSA 与 EEMD 降噪指标**

**Tab. 1 EEMD-SSA and EEMD noise reduction indexes**

| 降噪方法 | $R_{RMSE}$ | $r$ |
|---|---|---|
| EEMD | 0.063 | 0.993 |
| EEMD-SSA | 0.019 | 0.999 |

由表 1 可看出，直接采用 EEMD 降噪后的位移序列过程线较光滑，但其很大程度上丢失了原始序列的特征信息，特别是水位的突变情况导致的位移变化，这使序列的真实性大大降低。而 EEMD-SSA 多层次挖掘降噪保留了大部分原始序列的特征信息，同时也对原始序列中的毛刺、尖锐问题做了平滑处理，不仅有效降低了噪声干扰，也尽可能减少了细节的流失，能较好地反映真实的大坝位移变化情况。此外，EEMD-SSA 降噪方法的 $R_{RMSE}$ 显著小于 EEMD 降噪，且相关系数大于 EEMD 降噪。因此，相比于 EEMD 降噪，EEMD-SSA 降噪能更有效地去除大坝位移序列中的噪声影响，并保留其中的特征信息，其降噪结果能更准确地反映大坝位移真实变化情况。

### 4.3 NGO-XGBoost 模型构建

采用环境因子序列作为输入，将 EX53 测点降噪后的位移数据作为输出，利用 NGO 算法对 XGBoost 的超参数分别进行优化，各参数的选取范围和优化结果见表2。其中NGO算法的最大迭代次数设置为 300 次，种群大小设置为 100。

**表 2 XGBoost 参数优化结果**

**Tab. 2 XGBoost parameter optimization results**

| 优化参数 | 寻优范围 | 寻优结果 |
|---|---|---|
| 最大深度 | [3,30] | 15 |
| 学习率 | [0.01,1] | 0.782 |
| 子节点权重的最小总和 | [0,1] | 0.620 |
| 子样本占集合的比例 | [0,1] | 0.681 |
| 每棵随机采样的列数的占比 | [0,1] | 0.723 |

将优化后参数代入 XGBoost 模型后对位移序列进行建模分析。为了证明 NGO 算法的有效性和 NGO-XGBoost 模型在大坝位移预测的准确性，采用未优化的 XGBoost、LSTM 和统计回归模型对相同输入输出数据进行建模分析，其位移预测结果对比和残差箱型图分别见图 4、图 5。

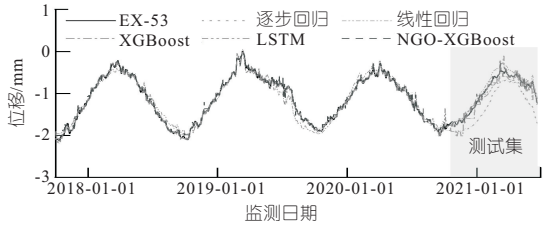


**图 4 模型预测结果对比**

**Fig. 4 Comparison of model prediction results**

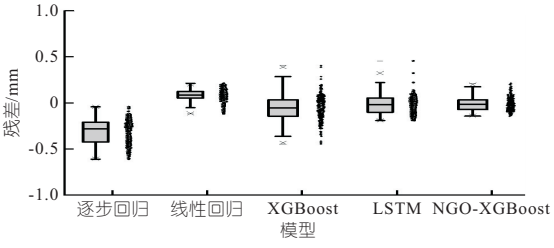


**图 5 预测阶段模型残差**

**Fig. 5 Residuals of models in test set**

为了更好地定量对比分析不同模型的准确性，引入 $R_{RMSE}$、平均绝对误差($M_{MAE}$)和决定系数($R^2$)对预测结果进行量化分析。

不同模型的误差指标结果见表 3。

**表 3 不同模型误差指标结果**

**Tab. 3 Different model error index results**

| 模型 | 训练阶段 | | | 预测阶段 | | |
|---|---|---|---|---|---|---|
| | $R_{RMSE}$ | $M_{MAE}$ | $R^2$ | $R_{RMSE}$ | $M_{MAE}$ | $R^2$ |
| 逐步回归 | 0.052 | 0.040 | 0.991 | 0.340 | 0.312 | 0.408 |
| 线性回归 | 0.111 | 0.090 | 0.958 | 0.105 | 0.903 | 0.944 |
| XGBoost | 0.031 | 0.022 | 0.997 | 0.146 | 0.113 | 0.901 |
| LSTM | 0.026 | 0.190 | 0.998 | 0.097 | 0.080 | 0.952 |
| NGO-XGBoost | 0.054 | 0.039 | 0.991 | 0.074 | 0.060 | 0.972 |

由表 3 可知，在训练阶段各模型的 $R^2$ 均大于 0.9，说明这些模型均能较好地对大坝位移进行拟合。然而在预测阶段逐步回归模型的 $R^2$ 仅有 0.408，其无法学习大坝位移数据中的非线性特征，只能预测出大坝位移变化趋势，难以精准预测大坝实际变形。而线性回归模型虽然在训练阶

段 $R^2$ 只有 0.958,但在预测阶段精度较高,一定程度上能预测大坝位移变化趋势。相比于逐步回归模型,线性回归模型、LSTM 和 XGBoost 模型在预测阶段的效果有显著提升,其 $R^2$ 均大于 0.9,能学习到大坝位移与环境因子之间的非线性关系,较好地预测大坝位移的变化趋势,但在预测位移的具体大小时存在一定的不足。特别是当大坝变形达到局部最大值和最小值时,其预测结果明显偏离实际位移,在实际应用过程中会产生诸多问题。

NGO-XGBoost 模型针对 XGBoost 中参数进行优化后 $R_{RMSE}$、$M_{MAE}$ 分别降低了 49.3%、46.9%,$R^2$ 提高了 8%,各指标也显著优于其他模型,其预测效果得到了显著提高。由图 4 可知,NGO-XGBoost 模型的残差中位数最接近 0,且分布较为集中,绝大部分位于 1.5 倍四分位距内,表明本模型较对照模型具有更好的预测精度和稳定性。NGO-XGBoost 模型不仅能准确预测大坝位移的变化趋势,同时能有效地预测大坝位移变化的极值,大大提高了大坝位移预测的准确性。

## 5　结论

**a.** 基于 EEMD-SSA 的多层次数据降噪处理方法能较好地区分噪声与特征信息,从而在保留特征信息的基础上,去除大坝位移监测信息中的噪声,其结果能较好地反映出大坝位移的真实变化情况。

**b.** 通过 NGO 算法对 XGBoost 模型参数进行优化能显著提高 XGBoost 模型预测的准确性,在大坝变形预测上具有显著优势。

**参考文献:**


[1] 曹梦茜, 郑东健. 基于 FCM-WOA-LSTM 的大坝变形预测模型及其应用[J]. 水电能源科学,2023,41(5):71-75.

[2] 曹恩华, 包腾飞, 刘永涛, 等. 基于 VMD 的多尺度变量提取法在混凝土坝变形预测中的应用[J]. 水电能源科学,2022,40(2):114-118.

[3] 徐韧, 苏怀智, 杨立夫. 基于 GP-XGBoost 的大坝变形预测模型[J]. 水利水电科技进展,2021,41(5):41-46,70.

[4] 董泳, 刘肖峰, 李云波, 等. 基于 EMD-EEMD-LSTM 的大坝变形预测模型[J]. 水力发电,2022,48(10):68-71,112.

[5] WU Z, HUANG N E. Ensemble empirical mode decomposition: a noise-assisted data analysis method[J]. Advances in adaptive data analysis,2009,1(1):1-41.

[6] BOUDRAA A O, CEXUS J C. EMD-based signal filtering[J]. IEEE transactions on instrumentation and measurement,2007,56(6):2196-2202.


# Multi-Level Data Processing-Based NGO-XGBoost Model for Dam Deformation Prediction


LI Chen-yang[a,b], ZHENG Dong-jian[a,b]

(a. College of Water Conservancy and Hydropower Engineering; b. State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing 210098, China)



**Abstract**: The noise and nonlinear characteristics in the deformation sequence of concrete dam seriously affect the accuracy of dam deformation prediction. In this paper, ensemble empirical modal decomposition (EEMD) was used to decompose the horizontal displacement signal of the dam to mine the effective deformation information. The singular spectrum analysis (SSA) was used to extract features from the high-frequency eigenmodal components (IMF) obtained from the decomposition to reduce the loss of effective information. Considering the complex stochastic and non-linear mapping relationship between effector and environmental variables, extreme gradient boosting (XGBoost) was used to model the prediction of the noise-reduced data. Considering the significant influence of XGBoost hyperparameters on the prediction performance of the model, the Northern Goshawk algorithm (NGO) with better global search capability was introduced to perform parameter search, and an NGO- XGBoost-based dam displacement prediction model was constructed. The calculation results show that the EEMD-SSA can effectively remove the noise from the dam displacement monitoring information, and the dam deformation prediction model based on NGO-XGBoost can significantly improve the prediction accuracy.



**Key words**: dam deformation prediction; EEMD; SSA; NGO; XGBoost