

中文引用格式:尹孝君,丁勇,李登华. 多算法融合因子筛选的 LSTM 模型在大坝变形预测中的应用[J]. 中国安全科学学报, 2025, 35(12):129-138.

英文引用格式:YIN Xiaojun, DING Yong, LI Denghua. Application of LSTM model with multi-algorithm fusion factor screening in dam deformation prediction[J]. China Safety Science Journal, 2025, 35(12):129-138.

多算法融合因子筛选的 LSTM 模型在大坝变形预测中的应用*

尹孝君¹, 丁勇^{**1}副教授, 李登华^{2,3}高级工程师

(1 南京理工大学 安全科学与工程学院, 江苏 南京 210094; 2 南京水利科学研究院, 江苏 南京 210029; 3 水利部水库大坝安全重点实验室, 江苏 南京 210024)

中图分类号:X928.03

文献标志码:A

DOI: 10.16265/j.cnki.issn1003-3033.2025.12.0198

基金项目:国家重点研发计划资助项目(2024YFC3210703);国家自然科学基金资助(U2240221);中央级公益性科研院所基本科研业务费专项资金资助项目(Yk325002, Y724011)

【摘要】 针对传统大坝变形预测中影响因子选择方法单一、难以全面捕捉高维数据复杂因子关联性的问题,提出一种基于多算法融合的因子筛选方法,并应用于构建最优因子的长短期记忆(LSTM)网络模型。使用相关系数法、邻域成分分析(NCA)算法、最小绝对值收缩和选择算子法(LASSO)分别筛选影响因子,并合并筛选结果;由于因子间会存在相关性,再根据对称不确定性(SU)剔除高度相关的因子,实现因子优选,利用 LSTM 网络建立大坝变形预测模型;以新疆某混凝土面板堆石坝为例,以均方根误差(RMSE)、平均绝对误差(MAE)、均方误差(MSE)及决定系数(R^2)作为评价指标评估模型性能。结果表明:相较于传统因子筛选方法,本文方法融合多种算法优势,能够全面、准确地筛选出对大坝变形影响显著的因子,其中 MSE 较其他方法降低 20.11%~59.09%,RMSE 降低 10.61%~36.05%,MAE 降低 9.95%~37.86%,构建出更优的预测模型;部分测点的 MSE、RMSE 和 MAE 较其他对比模型最大降幅分别达 53.5%、31.9%和 34.7%,且决定系数最高达 0.986 0。

【关键词】 多算法融合; 因子筛选; 最优因子长短期记忆网络模型(LSTM); 大坝变形; 预测模型

Application of LSTM model with multi-algorithm fusion factor screening in dam deformation prediction

YIN Xiaojun¹, DING Yong¹, LI Denghua^{2,3}

(1 School of Safety Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; 2 Nanjing Hydraulic Research Institute, Nanjing 210029, China;

3 Key Laboratory of Reservoir Dam Safety, Nanjing 210024, China)

Abstract: To address the reliance on a single factor selection approach in traditional dam deformation prediction and the difficulty of comprehensively capturing complex inter-factor relationships among high-

* 文章编号:1003-3033(2025)12-0129-10; 收稿日期:2025-06-20; 修稿日期:2025-09-18

** 通信作者:丁勇(1977—),男,江苏南京人,博士,副教授,主要从事结构健康检测、智能检测的研究。E-mail:12005048@njust.edu.cn。

dimensional influencing factors, a factor screening method based on the fusion of multiple algorithms was proposed and applied to construct an LSTM model using optimally factors. Specifically, influencing factors were selected separately using the correlation coefficient method, neighborhood component analysis (NCA), and least absolute shrinkage and selection operator (LASSO) technique. The results from these individual methods were subsequently integrated. Since correlations exist among the factors, highly correlated ones were further eliminated using symmetrical uncertainty (SU), thereby an optimized factors set was obtained. The selected factor set was then used to develop a dam deformation prediction model via an LSTM network. A concrete-faced rockfill dam in Xinjiang was used as the case study. The model performance was evaluated using root mean square error (RMSE), mean absolute error (MAE), mean square error (MSE), and the coefficient of determination (R^2). The results demonstrate that, compared with traditional factor selection methods, the proposed multi-algorithms are integrated and factors with significant influence on dam deformation are comprehensively and accurately identified. MSE is reduced by 20.11%–59.09%, RMSE by 10.61%–36.05%, and MAE by 9.95%–37.86%, and a superior predictive model was obtained, compared with models using conventional factor selection methods. For specific monitoring points, the maximum reductions in MSE, RMSE, and MAE reach 53.5%, 31.9%, and 34.7%, respectively, while the highest R^2 value attains 0.986 0.

Keywords: multi-algorithm fusion; feature selection; long short-term memory (LSTM); dam deformation; prediction model

0 引言

20 世纪 50 年代以来,共修建约 83 000 座堤坝,其中 15 m 以上的大坝约 18 000 座,30 m 以上大坝 3 000 多座。然而,受水文、地质、施工质量及老化等原因影响,部分大坝存在安全问题^[1]。大坝变形能直观、可靠反映大坝服役性态^[2]。因此,精确预测大坝变形具有重要现实意义。

当大坝内温度监测资料不足时,可采用静水-季节-时间 (Hydrostatic-Seasonal-Time, HST) 模型^[3-4]分析和预测大坝变形。对大坝影响因子筛选能剔除无关和冗余因子,提升模型预测精度与效率,确保大坝安全评估更为准确可靠。传统的影响因子筛选方法已无法满足日益复杂的筛选需求,如何选择合适的筛选方法成为当前研究的重点,通过降维改造影响因子集是降低模型复杂程度的主要方法。特征降维分为特征提取与特征选择^[5],利用相关系数法进行特征选择在多个领域具有广泛应用。赵佳蕊等^[6]使用皮尔逊相关系数计算出各气象因素与光伏发电功率间的相关性,以确定模型的输入特征。罗璐等^[7]提出了最小绝对值收缩和选择算子法 (Least Absolute Shrinkage and Selection Operation, LASSO),剔除了大坝变形影响因子集中的不显著变量,选取最优因子集作为模型输入变量。郑好等^[8]使用集成因子优选算法获得关键因子。目前,在土

木水利领域选择变量多采用随机森林法^[9]和岭回归^[10]。在预测模型方面,机器学习模型正逐步取代传统统计模型。当前广泛应用的如支持向量回归机 (Support Vector Regression, SVR)^[11]、极限梯度提升 (eXtreme Gradient Boosting, XGBoost)^[12]、轻量级梯度提升机 (Light Gradient Boosting Machine, lightGBM)^[13]等的机器学习模型。虽提升了预测精度,但仍存在局限性。相比之下,LSTM 在表征动态系统方面更具优势^[14],但其性能仍高度依赖于输入因子的质量。因此,如何构建科学、高效的因子筛选模型,提升 LSTM 在大坝变形预测中的可靠性,成为当前的关键问题。

在传统研究中,影响因子选择大多依赖于单一算法或统计模型,其在一定程度上能够提高模型性能并减少计算复杂度,而单一算法通常难以全面捕捉高维数据中复杂特征间的关联性^[15]。为此,笔者拟提出一种基于最优因子长短期记忆 (Long Short-Term Memory, LSTM) 网络模型的多算法融合的因子选择方法。通过相关系数法、邻域成分分析 (Neighborhood Component Analysis, NCA)^[16]算法、LASSO 回归 3 种算法,初步筛选出对模型性能最有益的因子;引入对称不确定性 (Symmetrical Uncertainty, SU)^[17]指标评估因子间的相关性,剔除冗余因子,结合 LSTM 模型预测大坝变形;与多种因子筛选算法和多种预测模型对比,以期提高大坝变

形预测模型精度。

1 多算法融合因子筛选方法

1.1 相关系数法

相关系数法包括皮尔逊相关系数和距离相关系数,前者适用于衡量线性关系,后者能同时捕捉线性和非线性依赖关系^[18],因此,引入距离相关系数,克服皮尔逊相关系数仅能衡量线性关系的局限,从而更全面准确地捕捉大坝变形与影响因子之间可能存在的复杂非线性依赖关系。计算公式为:

$$r = \frac{\sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^n (x_k - \bar{x})^2 \cdot \sum_{k=1}^n (y_k - \bar{y})^2}} \quad (1)$$

式中: r 为皮尔逊相关系数; n 为样本数量; k 为样本序号,用于标识每一组具体数据; x_k 为大坝影响因子; y_k 为大坝位移; $\bar{x}$ 和 $\bar{y}$ 分别为 2 个变量的均值。

$$\text{dCor}(X, Y) = \frac{\text{dCov}(X, Y)}{\sqrt{\text{dCov}(X, X) \cdot \text{dCov}(Y, Y)}} \quad (2)$$

式中: $\text{dCor}(X, Y)$ 为 X 和 Y 的距离相关系数; $\text{dCov}(X, Y)$ 为距离协方差; $\text{dCov}(X, X)$ 和 $\text{dCov}(Y, Y)$ 分别为 X 和 Y 的自身的距离协方差。

1.2 LASSO 回归法

LASSO 回归是通过引入 L_1 范数正则化项实现特征选择的线性回归方法。能在最小化残差的同时压缩回归系数,通过拟合广义线性模型,正则化参数控制模型复杂度,促使模型选择更少的特征^[19]。表达式为:

$$\min_{\omega} \frac{1}{2N} \mathbf{X}_{\omega} - \mathbf{y}_2^2 + \alpha \|\omega\|_1 \quad (3)$$

$$T_L = \{F_{L,j} | j = 1, 2, \dots, p, \omega_j \neq 0\} \quad (4)$$

式中: ω 为特征的权重系数向量; N 为数据集的样本总量; $\mathbf{X}_{\omega}$ 为预测值向量; y 为每个样本对应的真实值; $\mathbf{X}_{\omega} - \mathbf{y}_2^2$ 为误差项; α 为正则化参数; $\|\omega\|_1$ 为惩罚项,用来控制模型惩罚项大小; T_L 为筛选后的特征集合; $F_{L,j}$ 为第 j 个被 LASSO 选中的特征; j 为特征索引; p 为选择出的特征数量; ω_j 为各特征的权值向量。

1.3 邻域成分分析法

邻域成分分析法 (Neighborhood Component Analysis, NCA) 是一种无监督学习算法,通过学习特征的线性变换提升样本分类或回归的准确性。其核心在于优化目标函数,使样本在投影后的特

征空间中,邻域结构更好地反映类别信息。通过优化留一法 (Leave-One-Out, LOO) 交叉验证,求得马氏距离中的变换矩阵,使近邻分类效果最优。公式为:

$$d(x_k, x_l) = \sqrt{(\mathbf{A}\mathbf{x}_k - \mathbf{A}\mathbf{x}_l)^T (\mathbf{A}\mathbf{x}_k - \mathbf{A}\mathbf{x}_l)} \quad (5)$$

式中: $d(x_k, x_l)$ 表示样本 k 和样本 l 在投影后的特征空间中的欧氏距离; $k, l = 1, 2, \dots, n$ 为样本序号。 $\mathbf{x}_k, \mathbf{x}_l$ 为原始特征向量。分别代表第 k 个和第 l 个样本的特征向量。 $\mathbf{A}$ 为线性变换矩阵。通过定义一种基于距离的概率分布来评估样本 k 被其邻近样本 l 正确分类的概率。

$$p_{kl} = \begin{cases} \frac{\exp(-\mathbf{x}'_k - \mathbf{x}'_l)^2}{\sum_{k \neq i} \exp(-\mathbf{x}'_k - \mathbf{x}'_s)^2}, & k \neq l, \\ 0, & k = l \end{cases} \quad (6)$$

式中 p_{kl} 为样本 k 选择样本 l 作为其参考邻居的概率;用于将距离转换为一个正的概率值,距离越近,指数项的值越大; $\mathbf{x}'_k, \mathbf{x}'_l$ 为投影后的特征向量; $\mathbf{x}'_k - \mathbf{x}'_l$ 为样本 k 和 l 在投影空间中的欧氏距离; s 为求和索引;

D 用于测量 Y 中样本的距离:

$$D = \sum_k \sum_l p_{kl} \cdot \delta(y_k, y_l) \quad (7)$$

式中: $\delta(y_k, y_l)$ 为指示函数,当样本 k 和 l 的类别相同时, $\delta = 1$, 否则 $\delta = 0$; y_k 和 y_l 分别为样本 k 和样本 l 的标签或目标值。

确定目标函数为:

$$f(w) = \frac{1}{N} \sum_{k=1}^N J_k + \lambda \sum_{r=1}^p \omega_r^2 \quad (8)$$

式中: $f(w)$ 为目标函数; J_k 为样本 k 的损失; λ 为正则化参数,用于控制模型复杂度与训练误差之间的平衡, $\lambda \in [0.1, 0.01, 0.001]$ 选取性能最优的 λ 值; r 为特征权重索引。

1.4 对称不确定性

对称不确定性是一种基于信息论的特征选择评价指标,用于衡量 2 个变量间的相关性。通过归一化互信息 (Mutual Information, MI) 判断变量之间的关联程度。计算公式如下:

$$\text{SU}(X, Y) = \frac{2I(X, Y)}{H(X) + H(Y)} \quad (9)$$

式中: $I(X, Y)$ 为变量 X 和 Y 的互信息; $H(X)$ 和 $H(Y)$ 分别为变量 X 和 Y 的信息熵。

$$I(X, Y) = H(X) - H(X|Y) \quad (10)$$

式中: $H(X)$ 为变量 X 的信息熵,表示 X 的不确定

性; $H(X|Y)$ 为条件熵,表示在已知 Y 的情况下 X 的不确定性。

$$H(X) = - \sum_{x \in X} P(x) \log_2 P(x) \tag{11}$$

式中 $P(x)$ 为随机变量 X 取值为 x 时的概率。

$$H(X|Y) = - \sum_{y \in Y} P(y) \sum_{x \in X} P(x|y) \log_2 P(x|y) \tag{12}$$

式中: $P(y)$ 为随机变量 Y 取值为 y 的概率; $P(x|y)$ 表示在 $Y=y$ 的条件下, X 取值为 x 的条件概率。

1.5 LSTM 网络

LSTM 是一种特殊的循环神经网络 (Recurrent Neural Network, RNN), 通过引入记忆单元和门控机制 (输入门、遗忘门、输出门及细胞状态) 解决传统 RNN 在长序列数据处理中出现的梯度消失和爆炸问题^[20], 有效捕捉时间序列中的长期依赖关系。相关公式如下:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{13}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{14}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{15}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{16}$$

$$h_t = o_t \odot \tanh(c_t) \tag{17}$$

式中: i_t, f_t, o_t 分别为 t 时刻的输入门、遗忘门和输出门的输出; σ 为 sigmoid 函数; t 为时间步; x_t, h_{t-1} 分别为 LSTM 模型在 $t, t-1$ 时刻的输入和输出; c_{t-1}, c_t 分别为 $t-1, t$ 时刻的细胞状态; $\odot$ 为按元素相乘; W 和 b 为模型参数。

LSTM 模型在 t 时刻的输出为 h_t , 同时 c_t 也会作为下一个时刻的单元状态输入到模型中。

2 多算法融合的因子选择方法

单一信息度量很难全面评价特征子集的性能, 为弥补单一信息度量的不足^[21], 采用并行筛选-互补融合策略建立多算法融合的因子选择方法。通过相关系数法、NCA 法和 LASSO 算法分别从不同维度筛选影响因素。相关系数法侧重捕获与目标变量的显性线性关联。NCA 方法专注于挖掘非线性判别模式, 弥补线性方法的感知盲区。LASSO 算法通过施加稀疏约束, 消除共线性冗余并优化高维数据的处理。将 3 种方法筛选出的所有因子合并作为初选因子集。初始因子集兼具显著性、判别性与稀疏性 3 个层面。由于大坝的影响因子之间可能存在相关性, 且因子通常具有时间依赖性, 可能存在滞后结构。通过对称不确定性判断因子间的相关性, 并设定阈值剔除高相关性的因子, 最终获得最优因子集。

3 多算法融合的因子筛选变形预测模型

综上, 多算法融合的因子筛选变形预测模型具体步骤如图 1 所示。

步骤 1: 数据集准备, 填补大坝的变形及水位数据缺失值, 确定大坝变形影响因素。

步骤 2: 影响因素筛选, 使用上述多算法融合的特征选择方法筛选影响因素, 得到输入因子集。

步骤 3: 划分数据集, 将样本数据按 8 : 2 比例划分训练集和测试集。

步骤 4: 模型预测, 将筛选得到的最优影响因素输入到 LSTM 模型进行预测, 并评价预测结果。

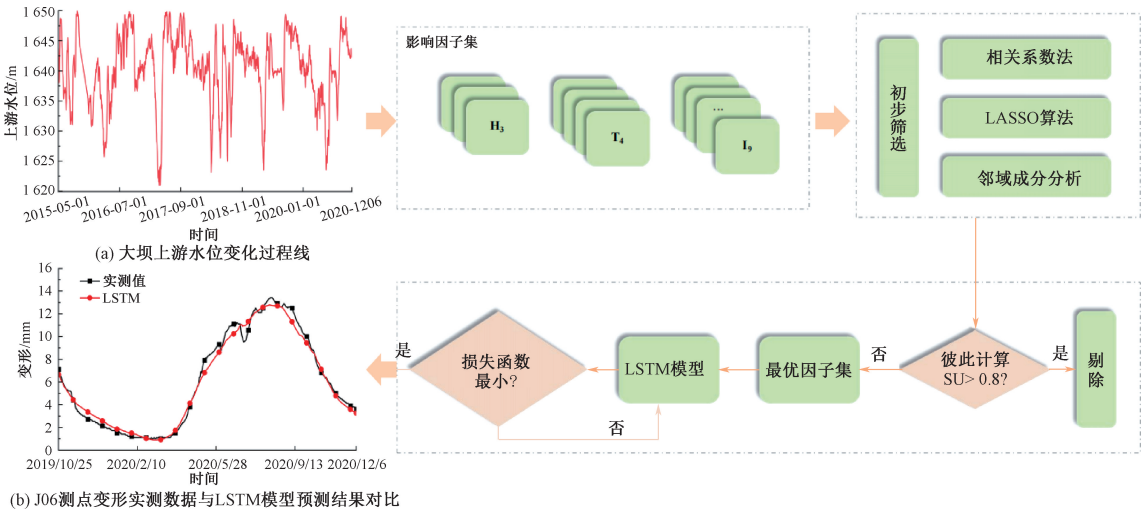


图 1 多算法融合因子筛选的 LSTM 变形预测模型构建流程

Fig. 1 Construction process of LSTM deformation prediction model with multi-algorithm fusion factor screening

4 大坝变形预测工程试验

4.1 大坝工程概况

以新疆某混凝土面板堆石坝为例,坝址的年平均径流量为 33.16 亿 m³,正常蓄水位为 1 649 m,水库总库容为 7 240 万 m³。上游水位 H 过程线如图 2 所示。

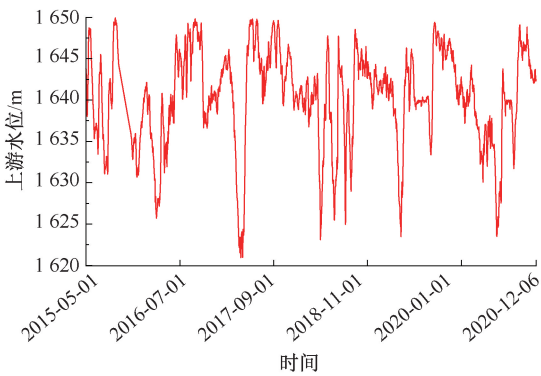


图 2 上游水位
Fig. 2 Upstream water level

由于该大坝所处区域的年降水量不足 180 mm,因此未考虑降水量对大坝的影响。文中选取坝左 0+160.000 和坝左 0+040.00 上的面板垂直缝(双向测缝计)测点 J-06 和 J2-02-1 进行对比分析。其中 J-06 位于坝左主坝中部、J2-02-1 位于坝肩区域,具有一定的区域差异性。测点监测频率为 1 次/d,使用简单实测值填补后的测点实测值过程线如图 3 所示。

为保证模型验证的科学性,采用严格的时间先后顺序划分数据集。将前期 80%时段的数据作为训练集,后期 20%时段的数据作为测试集。分析从 2015 年 5 月—2020 年 12 月共 2 048 组数据。监测数据按年为周期划分,训练集为 2015 年 5 月—2019 年 10 月共 1 638 组数据,测试集为 2019 年 10 月 25 日—2020 年 12 月 6 日共 409 组数据。

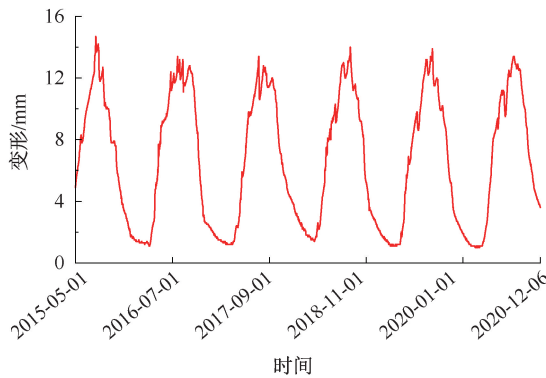
4.2 初始变量选取及模型参数

在构建大坝变形预测模型时,通常将水位因子、温度因子和时效因子作为关键影响变量。选取以下 16 种因子 $H_1 \sim H_3, T_1 \sim T_4, I_1 \sim I_9$ 作为初始变量输入。

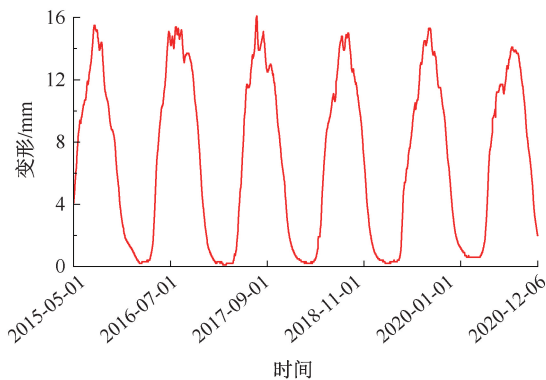
$$H = H_u - H_0 \tag{18}$$

$$H_k = \begin{cases} H, & k = 1 \\ H^2, & k = 2 \\ H^3, & k = 3 \end{cases} \tag{19}$$

式中: H_k 为水位因子; $K=1,2,3$; H_u 为当前上游水位, m; H_0 为初始上游水位, m。



(a) J06



(b) J2-02-1

图 3 实测过程线

Fig. 3 Process line of measured value

$$\begin{cases} T_1 = \sin \frac{2\pi\tau}{365} \\ T_2 = \cos \frac{2\pi\tau}{365} \\ T_3 = \sin \frac{4\pi\tau}{365} \\ T_4 = \cos \frac{4\pi\tau}{365} \end{cases} \tag{20}$$

式中: T_k 为温度因子; $K=1,2,3,4$; τ 为当前观测日到起始观测日累计天数。

$$\begin{cases} I_1 = \ln\tau_1 \\ I_2 = \ln(\tau_1 + 1) \\ I_3 = 1 - e^{-\tau_1} \\ I_4 = \frac{\tau_1}{1 + \tau_1} \\ I_5 = \tau_1 \\ I_6 = \tau_1^2 \\ I_7 = \tau_1^{0.5} \\ I_8 = \tau_1^{-0.5} \\ I_9 = \frac{1}{1 + \exp(-\tau_1)} \end{cases} \tag{21}$$

式中: I_k 为时效因子; $K=1\sim9$; τ_1 为当前时间距离起测日的天数。

在神经网络方面, LSTM 网络配置为 50 个隐藏单元、ReLU 激活函数、Adam(学习率为 0.001)、批次大小 32、训练轮次 50 并以均方误差 (Mean Squared Error, MSE) 作为损失函数; 门控循环单元 (Gated Recurrent Unit, GRU) 模型除采用默认激活函数, 其他参数与 LSTM 保持一致。在训练过程中, 使用测试集作为验证数据。关于机器学习模型采用网格搜索法进行超参数优化。所有模型均在统一的数据集划分和相同的输入因子集下训练与比较。

4.3 基于多算法融合的因子选择

采用第 2 节的影响因子筛选方法即使用相关系数法、LASSO 回归、NCA 法分别筛选因子, 识别与大坝变形最具相关性的影响因子。为避免信息冗余和模型过拟合, 限定每种筛选方法仅保留前 6 个最具相关性的因子。各步骤筛选得到的结果见表 1, 其中, $H_1\sim H_3$ 、 $T_1\sim T_4$ 、 $I_1\sim I_9$ 分别为 4.2 节介绍的水位、温度、时效因子。以测点 J-06 为例, 3 种方法均共同筛选出因子 H_3 与 T_3 , 说明其具有稳定的重要性。同时, H_1 、 T_1 、 T_2 和 T_4 等因子被多种方法重复识别, 体现了一定的一致性。NCA 法独立筛选出 I_2 和 I_7 , 则弥补了其他方法对非线性因子的忽略。LASSO 则筛选出 I_5 , 表明该方法所具有的稀疏性优势。在 J2-02-1 测点中, 3 种方法均筛选出 H_3 与 T_3 , 其他如 H_1 、 H_2 、 T_1 、 T_2 、 T_4 、 I_1 也在至少 2 种方法中被选中。 I_4 和 I_5 则分别由 NCA 和 LASSO 独立发现。综上所述, 多算法筛选策略在保障重要因子稳定性的同时, 也引入了互补性特征, 有助于提升最终特征子集的判别力与泛化能力, 进一步验证了并行筛选-互补融合策略的合理性与优势。将筛选后的因子合并为一个统一的因子集合。然后进一步采用 SU 方法再筛选合并后的因子。在选择阈值时, SU 值范围为 $[0, 1]$ 。SU 值越接近 1, 表示特征与目标变量的相关性越强。参考特征选择和降维的相关研究, 发现该值在 0.7~0.9 之间^[22]。为验证不同阈值对特征选择结果与预测性能的影响, 设置阈值为 0.7、0.75、0.8、0.85、0.9 共 5 组试验, 分别保留筛选后的特征子集, 并使用随机森林模型在相同训练集和验证集上预测。评估指标采用均方根误差 (Root Mean Square Error, RMSE)、平均绝对误差 (Mean Absolute Error, MAE)、决定系数 R^2 , 全面反映模型性能。SU 阈值对预测性能的影响见表 2。通过试验对比发现, 阈值设置为 0.8 能够在剔除冗余特征的同时保留足

够信息量。具体而言, 当 SU 阈值较低时, 尽管保留了较多特征, 但其中包含一定比例的信息冗余, 容易引入噪声, 导致模型预测精度下降; 而当阈值过高时, 虽然剔除了更多冗余特征, 但也可能误删了部分与目标变量存在非线性相关的重要信息, 造成模型欠拟合, 降低预测性能。在 SU 阈值为 0.8 时, 在 2 个测点内均达到较优指标。因此, 阈值 0.8 在保持特征集多样性基础上, 有效控制了特征间的冗余性, 故阈值设为 0.8。由筛选结果可知: 经筛选后的因子主要分量为水位分量与温度分量。这与大坝所在的地区水位波动较大、气候和降水季节性变化密切相关。由于建坝时间长, 时效性影响的累积效应相对较慢且逐渐趋于稳定, 使得时效分量占比最小。

表 1 各步骤筛选得到的影响因子

Table 1 Influencing factors obtained from screening at each step

筛选方法	相关系数法	LASSO 回归	NCA 法	合并的全部因子	再筛选后的最优因子集
J-06	$H_1\sim H_3$ $T_1\sim T_3$	H_3 $T_1\sim T_4$ I_5	H_1, H_3 T_3, T_4 I_2, I_7	$H_1\sim H_3$ $T_1\sim T_4$ I_2, I_5, I_7	H_1 $T_1\sim T_4$ I_2, I_5
	H_1, H_3 $T_1\sim T_3$ I_1	H_2, H_3 $T_1\sim T_4$ I_5	H_1, H_2 T_3, T_4 I_1, I_4	$H_1\sim H_3$ $T_1\sim T_4$ I_1, I_4, I_5	$H_1\sim H_3$ $T_1\sim T_3$ I_1, I_4

表 2 SU 阈值对预测性能的影响

Table 2 Impact of SU threshold on prediction performance

测点	SU 阈值	保留因子数	RMSE	MAE	R^2
J-06	0.7	9	0.721 4	0.592 1	0.964 0
	0.75	8	0.688 2	0.557 8	0.970 3
	0.8	7	0.665 0	0.532 4	0.974 8
	0.85	5	0.673 1	0.542 0	0.971 5
	0.9	4	0.702 8	0.562 3	0.965 9
J2-02-1	0.7	10	0.823 5	0.698 7	0.962 6
	0.75	9	0.793 2	0.670 1	0.967 1
	0.8	8	0.785 7	0.662 5	0.974 2
	0.85	6	0.791 4	0.668 9	0.970 4
	0.9	4	0.812 9	0.681 5	0.963 8

4.4 影响因子选择结果对比

为验证文中影响因子筛选方法的有效性, 采用 LSTM 作为预测模型并通过 MSE、RMSE、MAE 及 R^2 为评价指标分析模型性能。将文中方法与相关系数法、LASSO 回归、随机森林等传统因子选择法对比, 同时, 通过设置未使用对称不确定性再筛选的对比试验, 验证该方法在去除因子冗余性方面的作用。

从表 3 和图 4 看出,测点 J-06 和测点 J2-02-1 在未筛选时,数据中存在大量冗余特征。造成模型的预测效果与拟合精度较差。使用文中方法能够有效降低预测模型复杂度。在测试集中,模型能够较好地捕捉数据的整体趋势。在数据波动较大的区域,模型依然能够有效反映出数据的变化趋势。在评价指标方面,MSE、RMSE、MAE 均为最小。其中 MSE 较其他方法降低 20.11%~59.09%,RMSE 降低 10.61%~36.05%,MAE 降低 9.95%~37.86%。 R^2 为 0.986 0 与 0.981 1,表明拟合程度高且预测效果好,能够较准确地反映大坝变形的变化趋势。经过对称不确定性再筛选后,各项指标均有所提升。表明再筛选步骤在剔除因子间冗余信息以及降低模型复杂度方面具有重要意义。

表 3 筛选方法预测性能对比

Table 3 Comparison of the prediction performance of the screening methods					
测点	筛选方法	MSE	RMSE	MAE	R^2
J-06	未筛选	0.600 8	0.775 1	0.641 3	0.965 8
	相关系数法	0.507 2	0.712 2	0.505 9	0.971 1
	随机森林	0.442 3	0.665 0	0.532 4	0.974 8
	LASSO 回归	0.416 3	0.64 52	0.524 5	0.976 3
	未再筛选	0.312 0	0.558 5	0.442 5	0.982 2
	本文方法	0.245 7	0.495 7	0.398 5	0.986 0
J2-02-1	未筛选	0.866 1	0.930 6	0.721 8	0.963 8
	相关系数法	0.753 3	0.867 9	0.758 3	0.968 5
	随机森林	0.617 3	0.785 7	0.662 5	0.974 2
	LASSO 回归	0.612 1	0.782 4	0.635 5	0.974 4
	未再筛选	0.566 9	0.752 9	0.591 6	0.976 3
	文中方法	0.452 9	0.673 0	0.509 0	0.981 1

4.5 预测模型结果对比

将 LSTM 与大坝变形领域中常用的 XGBoost、随机森林、LightGBM 和 GRU 模型对比。2 个测点的预测精度评价指标见表 4 和如图 5 所示。

表 4 各模型预测精度评价指标

Table 4 Evaluation indexes of each model					
测点	预测模型	MSE	RMSE	MAE	R^2
J-06	XGBoost	0.529 4	0.727 6	0.610 3	0.969 8
	随机森林	0.467 1	0.683 4	0.496 9	0.973 4
	LightGBM	0.430 9	0.656 4	0.503 4	0.975 5
	GRU	0.329 3	0.573 8	0.445 0	0.981 2
	LSTM	0.245 7	0.495 7	0.398 5	0.986 0
J2-02-1	XGBoost	0.764 3	0.874 2	0.635 8	0.968 1
	随机森林	0.600 7	0.775 0	0.627 1	0.974 9
	LightGBM	0.565 7	0.752 1	0.594 2	0.976 4
	GRU	0.547 1	0.739 6	0.590 4	0.977 2
	LSTM	0.452 9	0.673 0	0.509 0	0.981 1

由表 4 和图 5 可知:LSTM 模型的 MSE 较对比模型降低 25.4%~53.5%,RMSE 较对比模型降低 13.6%~31.9%,MAE 较对比模型降低 10.5%~34.7%。在 J2-02-1 测点中,LSTM 模型的 MSE 较对比模型降低 17.2%~40.7%,RMSE 较对比模型降低 9%~23.1%,MAE 较对比模型降低 13.8%~19.9%,表明 LSTM 在误差控制上的优秀表现。LSTM 的 R^2 分别为 0.986 0 和 0.981 1,证明了 LSTM 能够更好地拟合数据的真实变化。

总体而言,LSTM 模型在文中表现显著优于其他预测模型,特别是在处理具有时序特征的复杂问题时,LSTM 能够充分利用其递归结构,捕捉数据中的时序依赖性,进而提升预测精度。在大坝变形中,LSTM 展现出更强的泛化能力和更高的预测准确性。

4.6 模型泛化能力验证

为验证模型在不同时间段下的泛化性能,采用时间序列五折交叉验证方法研究。基于 2015 年 5 月—2019 年 10 月共 1 638 组训练数据,按时间顺序将数据等分为 5 段,依次选取每段作为验证集,其前段数据作为训练集。每折中训练集占比约 80%,验证集占比约 20%,验证集时间始终晚于训练集,避免未来信息泄露,确保符合时序建模原则。2 个测点的验证结果见表 5、表 6。

表 5 测点 J-06 交叉验证与原文结果对比

Table 5 Comparison of cross-validation and original results for J06				
数据来源	MSE	RMSE	MAE	R^2
原文测试集	0.245 7	0.495 7	0.398 5	0.986 0
第 1 折	0.248 3	0.498 3	0.401 2	0.985 6
第 2 折	0.251 7	0.501 7	0.403 8	0.985 1
第 3 折	0.243 1	0.493 1	0.395 2	0.986 5
第 4 折	0.253 4	0.503 4	0.406 8	0.984 7
第 5 折	0.249 6	0.499 6	0.401 9	0.985 5
交叉验证平均	0.249 2±0.003 9	0.499 2±0.003 9	0.401 8±0.004 4	0.985 5±0.000 7

表 6 测点 J2-02-1 交叉验证与原文结果对比

Table 6 Comparison of cross-validation and original results for J2-02-1				
数据来源	MSE	RMSE	MAE	R^2
原文测试集	0.452 9	0.673 0	0.509 0	0.981 1
第 1 折	0.458 2	0.676 9	0.514 3	0.980 5
第 2 折	0.462 7	0.680 2	0.518 1	0.979 8
第 3 折	0.447 5	0.668 9	0.503 7	0.982 0
第 4 折	0.453 8	0.673 6	0.510 2	0.980 9
第 5 折	0.459 1	0.677 6	0.515 0	0.980 2
交叉验证平均	0.456 3±0.006 1	0.675 4±0.004 7	0.512 3±0.005 8	0.980 7±0.000 9

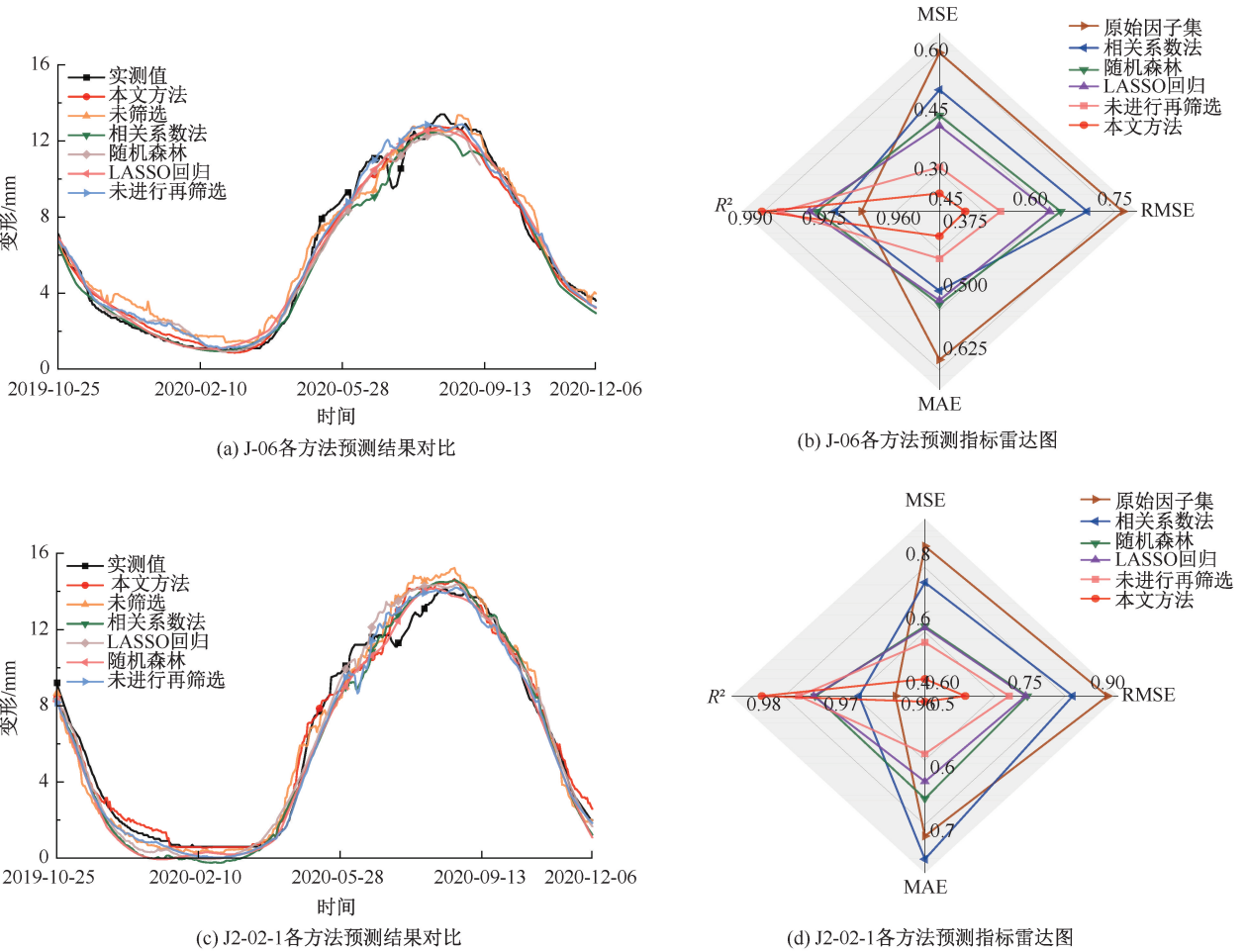


图 4 预测结果对比
Fig. 4 Comparative chart of prediction results

在不同时间段的验证集上,模型预测误差保持稳定。由表 5 可知:测点 J-06 的 MSE 波动范围仅为 0.243 1~0.253 4,相当于预测值的 $\pm 1.6\%$ 偏差。特别地,第 3 折与第 5 折分别表示典型水文年和特殊气候期,两者的 MSE 差异不足 4.2%,表明模型对季节性变化和异常气候事件均具有适应性。对比 2 个测点的表现,J-06 与 J2-02-1 的 R^2 平均值分别为 0.985 5 和 0.980 7,性能差异控制在 0.5% 以内。虽然 J2-02-1 位于地质条件更复杂的坝肩区域,其 MAE 标准差仍小于预测值的 1.2%,证明模型对不同空间位置的变形规律均有稳定捕捉能力。

为进一步验证模型在不同时间划分下预测性能的一致性,以 MSE 为统计指标,对比原始测试集结果与五折时间序列交叉验证的结果,采用单样本 Student 检验进行显著性分析。 p 值为衡量差异的显著性指标。通常,当 $p < 0.05$ 时,差异具有统计学显著性;当 $p \in [0.05, 0.10)$,差异边缘显著;当 $p \geq 0.10$,认为差异无统计学意义。以原始测试集的

MSE 为基准值,比较五折交叉验证中各折的 MSE 是否存在显著偏差。以测点 J-06 为例,原始测试集 MSE 为 0.245 7,五折交叉验证的 MSE 均值为 0.249 2,标准差为 0.003 9。T 检验结果表明($p = 0.12 > 0.05$):差异不具有统计显著性。J2-02-1 测点分析结果与此一致($p = 0.09 > 0.05$)。因此,模型在不同时间段下的预测误差无显著差异。进一步印证了模型在时序泛化性上的稳定性与可靠性。

5 结 论

1) 基于最优因子长短期记忆网络模型的多算法融合的因子选择方法,有效识别出对大坝变形具有显著影响的关键因子,其构成以水位分量 H_1 、 H_2 、 H_3 和温度分量 T_1 、 T_2 、 T_3 、 T_4 为主,并包含部分时效因子。显著降低了输入维度与模型复杂度。

2) 文中方法筛选出的最优因子集显著提升了 LSTM 模型的预测性能。在 J-06 和 J2-02-1 2 个测点上,MSE 最大降幅分别为 53.5% 和 40.7%,RMSE

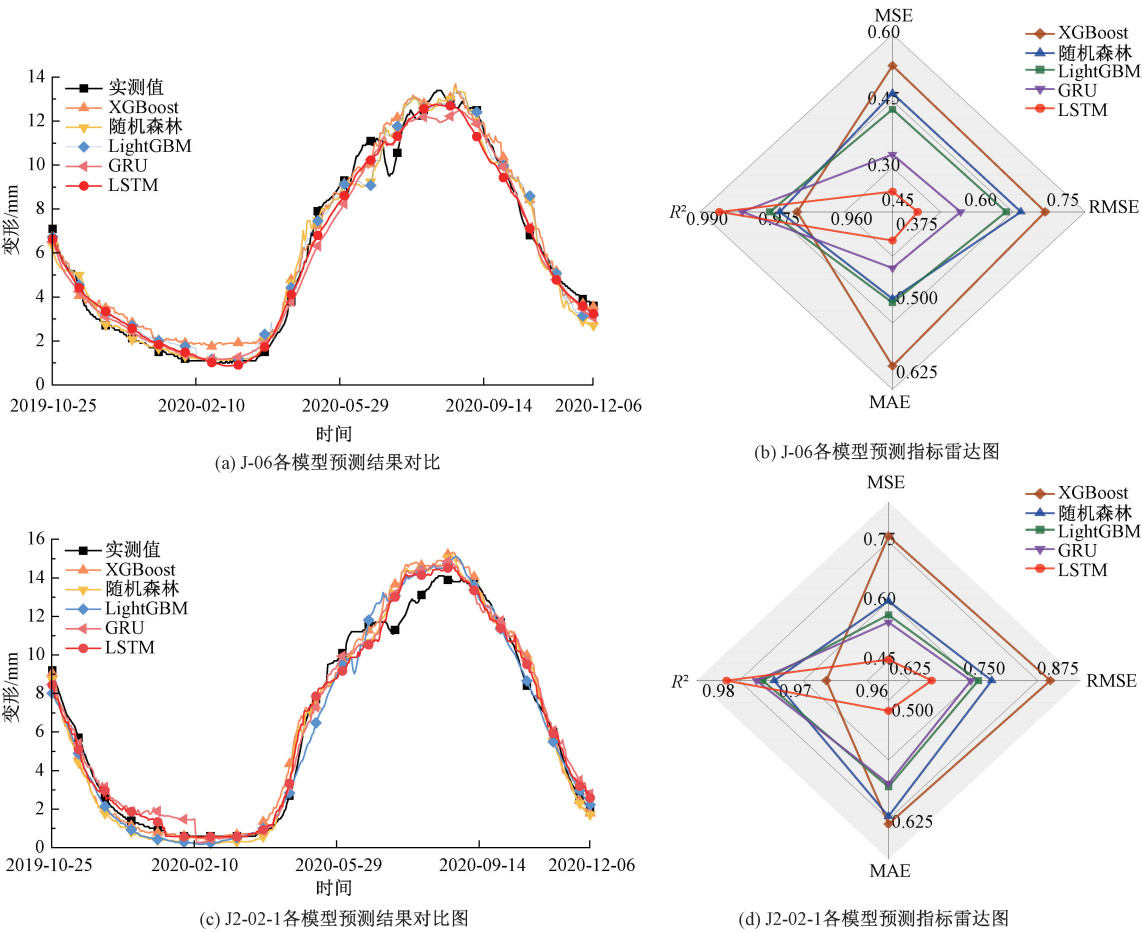


图 5 模型性能对比

Fig. 5 Model performance comparison chart

最大降幅分别为 31.9%和 23.1%,MAE 最大降幅分别为 34.7%和 19.9%,且决定系数最高达到 0.986 0,表明模型具有较高的拟合优度和预测准确性。

3)通过时间序列五折交叉验证,模型在不同测点上均表现出稳定的泛化能力,MSE 波动范围小,决定系数平均值保持在 0.980 7 以上,进一步证明了该方法在实际工程应用中的可靠性与鲁棒性。

参 考 文 献

[1] 吴中如. 水工建筑物安全监控理论及其应用[M]. 北京: 高等教育出版社, 2003:1.

[2] 周仁练,苏怀智,韩彰,等. 混凝土坝变形的长期预测模型与应用[J]. 水力发电学报,2021,40(9):122-131.

ZHOU Renlian, SU Huaizhi, HAN Zhang, et al. Long-term deformation prediction model of concrete dams and its application [J]. Journal of Hydroelectric Engineering, 2021, 40(9): 122-131.

[3] WILLM G, BEAUJOINT N. Les méthodes de surveillance des barrages au service de la production hydraulique d'Electricité de France, problèmes anciens et solutions nouvelles[C]. IXth International Congress on Large Dams,1967: 529-550.

[4] GAMSE S, OBERGUGGENBERGER M. Assessment of long-term coordinate time series using hydrostatic-season-time model for rock-fill embankment dam[J]. Structural Control and Health Monitoring, 2017, 24(1): DOI:10.1002/stc.1859.

[5] KARLUPIA N, ABROL P. Wrapper-based optimized feature selection using nature-inspired algorithms [J]. Neural Computing and Applications, 2023, 35(17): 12 675-12 689.

[6] 赵佳蕊,王玲芝,李晨阳. 基于 LEA-LSTM 的光伏发电功率短期预测方法[J]. 电力信息与通信技术,2024,22(11): 34-42.

ZHAO Jiarui, WANG Lingzhi, LI Chenyang. A short-term prediction method of photovoltaic power generation based on LEA-LSTM[J]. Electric Power Information and Communication Technology, 2024, 22(11): 34-42.

- [7] 罗璐,李志,张启灵. 大坝变形预测的最优因子长短期记忆网络模型[J]. 水力发电学报,2023,42(2):24–35.
LUO Lu,LI Zhi,ZHANG Qiling. Optimal factor set based long short-term memory network model for prediction of dam deformation[J]. Journal of Hydroelectric Engineering,2023,42(2):24–35.
- [8] 郑好,李登华,丁勇. 基于集成因子优选算法的面板堆石坝变形预测模型[J]. 排灌机械工程学报,2025,43(2):178–186.
ZHENG Hao, LI Denghua, DING Yong. Deformation prediction model of panel rockfill dam based on integrated factor optimization algorithm[J]. Journal of Drainage and Irrigation Machinery Engineering, 2025, 43(2): 178–186.
- [9] SU Yan, WENG Kailiang, LIN Chuan, et al. Dam deformation interpretation and prediction based on a long short-term memory model coupled with an attention mechanism [J]. Applied Sciences, 2021, 11 (14): DOI: 10.3390/app11146625.
- [10] 蒙伟,何川,陈子全,等. 岭回归在岩体初始地应力场反演中的应用[J]. 岩土力学,2021,42(4):1156–1169.
MENG Wei,HE Chuan,CHEN Ziquan, et al. Application of ridge regression in the inversion analysis of the initial geostress field of rock masses[J]. Rock and Soil Mechanics,2021,42(4):1156–1169.
- [11] 袁冬阳,顾冲时,顾昊. 严寒地区混凝土重力坝变形行为分析与预测模型[J]. 水利学报,2022,53(6):733–746.
YUAN Dongyang, GU Chongshi, GU Hao. Displacement behavior analysis and prediction model of concrete gravity dams in cold region[J]. Journal of Hydraulic Engineering,2022,53(6):733–746.
- [12] 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.
XU Ren,SU Huaizhi,YANG Lifu. Dam deformation prediction model based on GP-XGBoost[J]. Advances in Science and Technology of Water Resources, 2021, 41 (5): 41–46,70.
- [13] LIU Mingkai, WEN Zhiping, SU Huaizhi. Deformation prediction based on denoising techniques and ensemble learning algorithms for concrete dams[J]. Expert Systems with Applications, 2024, 238;DOI:10.1016/j.eswa.2023.122022.
- [14] 何裕坤,晁阳,李同春,等. 基于 SBO-LSTM 的大坝变形预测模型 [J]. 水利水电技术:中英文, 2024, 55 (增 1): 78–86.
HE Yukun,CHAO Yang,LI Tongchun,et al. Dam deformation prediction model based on SBO-LSTM[J]. Water Resources and Hydropower Engineering,2024, 55(S1): 78–86.
- [15] 吉丽萍,黄景涛,李一凡,等. 一种基于多信息融合的风电功率预测特征选择方法 [J]. 兰州理工大学学报, 2024, 50 (5): 94–100.
JI Liping,HUANG Jingtao,LI Yifan, et al. A wind power prediction feature selection method based on multi-information fusion [J]. Journal of Lanzhou University of Technology,2024, 50 (5): 94–100.
- [16] XIONG Rui, SUN Yue, WANG Chenxu, et al. A data-driven method for extracting aging features to accurately predict the battery health[J]. Energy Storage Materials, 2023, 57: 460–470.
- [17] 高浩田,李东喜,陈泽华,等. 基于交互式多特征融合算法的药物靶标预测 [J]. 太原理工大学学报, 2024, 55(4): 751–758.
GAO Haotian,LI Dongxi,CHEN Zehua, et al. Prediction of drug-target interactions based on interactive multi-feature fusion algorithm [J]. Journal of Taiyuan University of Technology, 2024, 55(4): 751–758.
- [18] 王晓楠,苏文浩,董灵波. 应用特征选择和机器学习方法建立兴安落叶松单木树龄预测模型 [J]. 东北林业大学学报, 2024, 52 (11): 64–71,82.
WANG Xiaonan, SU Wenhao, DONG Lingbo. Establishing a prediction model for the age of *Larix gmelinii* individual tree using feature selection and machine learning methods [J]. Journal of Northeast Forestry University, 2024, 52(11): 64–71, 82.
- [19] 毛杰,陈宓舟,许磊,等. 基于双重选择 LASSO 模型的我国股市定价因子边际有效性研究 [J]. 系统工程理论与实践, 2024, 44 (9): 2 993–3 008.
MAO Jie,CHEN Mizhou,XU Lei, et al. Research on the marginal effectiveness of Chinese stock markets'pricing factors: application of double-selection LASSO model[J]. Systems Engineering-Theory & Practice,2024, 44 (9): 2 993–3 008.
- [20] 刘天翼,艾星星,张九丹. 基于 MLR-DE-LSTM 的大坝变形串联组合预测模型 [J]. 中国农村水利水电,2025(2):207–212.
LIU Tianyi,AI Xingxing,ZHANG Jiudan. Dam deformation serial combined prediction model based on MLR-DE-LSTM [J]. China Rural Water and Hydropower,2025(2):207–212.
- [21] 史云峰,韩晓明,王利红,等. 基于特征选择和机器学习的锂电池寿命早期预测 [J]. 现代电子技术, 2024, 47 (22): 90–98.
SHI Yunfeng,HAN Xiaoming,WANG Lihong, et al. Lithium battery life early prediction based on feature selection and machine learning[J]. Modern Electronics Technique,2024, 47 (22): 90–98.
- [22] YU Lei, LIU Huan. Feature selection for high-dimensional data: a fast correlation-based filter solution[C]. Proceedings of the 20th International Conference on Machine Learning (ICML-03),2003: 856–863.

作者简介: 尹孝君 (2001—),男,山东青岛人,硕士研究生,研究方向为大坝结构健康监测。E-mail: 1526505150@qq.com。