引文格式：闫少霞. 根据SAE特征优选和集成学习的大坝变形预测方法[J]. 地理空间信息, 2026, 24(1): 110-113. doi:10. 3969/j.issn. 1672-4623.2026.01.024

根据SAE特征优选和集成学习的大坝变形预测方法

闫少霞
（广州南方测绘科技股份有限公司，广东 广州 510000）

摘　要：大坝变形是一种复杂的非线性、非平稳时间序列，传统方法存在预测精度低、泛化能力弱等问题。提出了一种基于稀疏自编码器（SAE）和Stacking集成学习的大坝变形预测模型。首先利用SAE对影响大坝变形的19种因素进行分析，自动提取第二阶水压因子，第二、三阶温度因子以及第一、二个时效因子5维特征作为主控因素；再以SAE提取的主控因素为输入，建立以多元线性回归、支撑向量回归和灰色理论模型为基学习器，以长短时记忆神经网络为元学习的Stacking集成学习模型，预测大坝变形趋势。以我国中东部地区某混凝土大坝变形监测实际数据开展验证试验，结果表明，相较于传统方法，该方法的预测精度更高、泛化能力更强，具有更广的应用前景。

关键词：SAE；主控因素；Stacking集成学习；大坝变形预测模型；预测结果分析
中图分类号：P258　　**文献标志码：**B　　**文章编号：**1672-4623（2026）01-0110-04

Dam Deformation Prediction Method Based on SAE Feature Optimization and Ensemble Learning

YAN Shaoxia
(Guangzhou Southern Surveying and Mapping Technology Co., Ltd., Guangzhou 510000, China)

Abstract: The dam deformation is a complex nonlinear and non-stationary time series. The traditional methods have the problems of low prediction accuracy and weak generalization ability. We proposed a dam deformation prediction model based on sparse auto-encoder (SAE) and Stacking ensemble learning. Firstly, we used SAE to analyze nineteen factors affecting dam deformation, and automatically extracted the five-dimensional characteristics of the second-order water pressure factor, the second-order and third-order temperature factors, and the first and second aging factors as the main control factors. Then, taking the main control factors obtained by SAE as inputs, we established a Stacking ensemble learning model based on multiple linear regression, support vector regression, grey model and long short term memory neural network, to predict the dam deformation trend. Based on the actual deformation monitoring data of a concrete dam in central and eastern China, we carried out the verification test. The result shows that compared with the traditional methods, the proposed method has higher prediction accuracy, stronger generalization ability and broader application prospect.

Key words: SAE, main control factor, Stacking ensemble learning, dam deformation prediction model, prediction result analysis

随着水利资源的深入开发，我国在册大坝数量已超过10万座，位居世界首位。各类大坝在我国发电、防洪、供水等方面发挥巨大作用的同时，其安全问题也日益凸显。实时、准确获取大坝运维状态并预测未来变形趋势，对于保证大坝长效服役和除险加固具有重要意义[1-2]。目前国内外学者对大坝变形预测的研究可归纳为统计类方法和人工智能类方法两大类。统计类方法的核心是采用准确的数学模型对大坝变形过程中涉及的因变量与自变量关系进行建模，从而定量分析未来发展趋势[3]，以求和自回归滑动平均[4]，多元线性回归（MLR）[5]和灰色理论模型（GM）[6]等为代表，具有原理简单、可解释性强和实时性高的优点。该类方法大多采用线性模型对大坝变形过程进行分析，而大坝变形过程是一种典型的非线性、非平稳随机过程，因此预测精度较低，且存在误差累积问题[7-8]。人工智能类方法的核心在于数据驱动，即利用人工智能算法的自适应和自学习能力从历史数据中抽象出数据的变化规律，从而预测未来发展趋势[9]，以支撑向量回归[10]（SVR）、长短时记忆神经网络[11]（LSTM）和卷积神经网络[12]（CNN）等为代表。该类方法虽能获得优于统计类方法的预测精度，且对非线性、非平稳数据的适应能力更强，但需要大量统计特性相近的训练数据，且存在模型的可解释性差等问题，限制了其在实际中的推广应用[13-14]。上述方法均采用单一模型进行大坝变形预测。集成学习理论的发展表明，由于随机性的存在，单一模型的预测性能始

作者简介：闫少霞（1987—），本科，高级工程师，研究方向为高精度GNSS，E-mail：Sdibnatjll@163.com。

终存在局限，且泛化能力较弱，而集成不同模型能有效提升预测精度以及对复杂数据的适应性和泛化能力。基于此，本文提出了一种基于稀疏自编码器（SAE）和Stacking集成学习的混凝土大坝变形预测模型；并采用我国中东部地区某混泥土大坝实际数据开展验证实验，将预测结果与传统方法进行对比分析。

1　混凝土大坝变形预测方法

本文方法首先利用SAE自动选择大坝变形的主控因素，去除冗余信息降低后续预测模型的复杂度；再以主控因素为输入，建立以MLR、SVR和GM(1,1)为基学习器，以LSTM为元学习器的Stacking集成学习模型进行变形预测。本文采用平均绝对误差（MAE）、均方根误差（RMSE）和决定系数（R^2）3项指标评估方法的预测性能。

1.1　基于SAE的大坝变形主控因素选择

大坝变形过程受水压、温度和时效等因素影响，因此其变形监测数据是一种典型的非线性、非平稳时间序列。若直接对原始数据进行建模分析，将出现预测模型复杂和运算量大的问题，且原始数据中存在的冗余信息也会引起过拟合和维数灾难。

1）影响因素分析。水压变化会对大坝坝体和地基土体产生不同力度的荷载，从而导致大坝形变。水压因子的表达式为：

$$\begin{cases} H_1 = H \\ H_2 = H^2 \\ H_3 = H^3 \end{cases} \tag{1}$$

式中，H 为大坝上游水位；H_1、H_2 和 H_3 分别为第一、二和三阶水压因子。

温度变化会引起大坝坝体和地基土体的无规律膨胀或收缩，导致大坝形变。温度因子的表达式为：

$$\begin{cases} T_1 = \sin\left(\dfrac{2\pi t}{365}\right) \\ T_2 = \cos\left(\dfrac{2\pi t}{365}\right) \\ T_3 = \sin\left(\dfrac{4\pi t}{365}\right) \\ T_4 = \cos\left(\dfrac{4\pi t}{365}\right) \end{cases} \tag{2}$$

式中，T_i 为第 i 阶温度因子；t 为累积观测时间。

时效因子是指大坝坝体和地基土体在持续荷载作用下产生的蠕变和非弹性变化，表达式为：

$$\begin{cases} I_1 = \log_{10}\theta \\ I_2 = \theta^{-0.5} \\ I_3 = \theta^{0.5} \\ I_4 = \theta \\ I_5 = \theta^2 \end{cases} \quad \begin{cases} I_6 = \dfrac{1}{1-e^{-\theta}} \\ I_7 = \dfrac{1}{1-e^{-\theta}} \\ I_8 = \dfrac{\theta}{\theta+1} \end{cases} \quad \begin{cases} I_9 = \dfrac{\log_{10}\theta}{\log_{10}(\theta+1)} \\ I_{10} = e^{-\theta} \\ I_{11} = e^{-0.1\theta} \\ I_{12} = e^{-0.01\theta} \end{cases} \tag{3}$$

式中，I_j 为第 j 个时效因子；$\theta = t/100$。

将水压因子、温度因子和时效因子等19项影响因子写成向量形式，则有：

$$\boldsymbol{f} = \begin{bmatrix} H_1\ H_2\ H_3\ T_1\ T_2\ T_3\ T_4\ I_1\ I_2 \\ I_3\ I_4\ I_5\ I_6\ I_7\ I_8\ I_9\ I_{10}\ I_{11}\ I_{12} \end{bmatrix}^{\mathrm{T}} \tag{4}$$

2）主控因素选择。本文采用SAE对19项影响因子进行分析。SAE是在自编码器（AE）的基础上发展起来的一种无监督机器学习方法，通过在AE编码过程中引入稀疏性惩罚因子抑制隐层神经元的活跃程度，从而在压缩降维输入高维数据的同时避免过拟合问题发生[15]。利用SAE对特征向量 $\boldsymbol{f}$ 进行特征选择的过程主要分为编码和解码两个阶段。在编码阶段SAE通过tanh激活函数 h 将 $\boldsymbol{f}$ 映射至隐藏层，得到隐层神经元 s，即

$$s(\boldsymbol{f}, \boldsymbol{w}, b) = h\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{f} + b\right) \tag{5}$$

式中，$\boldsymbol{w}$ 为编码阶段的权值向量；b 为编码阶段的偏置参数。

在解码阶段SAE采用Relu激活函数 g 对 s 进行重构，并将其映射至输出层，从而得到降维后的特征向量 $\bar{\boldsymbol{f}}$，即

$$\bar{\boldsymbol{f}} = g\left[\boldsymbol{\omega} \times s(\boldsymbol{f}, \boldsymbol{w}, b) + \bar{b}\right] \tag{6}$$

式中，$\boldsymbol{\omega}$ 为解码阶段的权值向量；$\bar{b}$ 为解码阶段的偏置参数。

为在降维过程中尽可能地避免信息丢失，SAE在上述过程中增加了包含重构误差和稀疏性惩罚因子的代价函数，即

$$J\left(\boldsymbol{w}, \boldsymbol{\omega}, b, \bar{b}\right) = \left(\boldsymbol{f} - \bar{\boldsymbol{f}}\right)^{\mathrm{T}}\left(\boldsymbol{f} - \bar{\boldsymbol{f}}\right) + \frac{\beta}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\omega} + \gamma\left[\sum_{j=1}^{L}\rho\lg\frac{\rho}{\bar{\rho}_j} + (1-\rho)\lg\frac{1-\rho}{1-\bar{\rho}_j}\right] \tag{7}$$

式中，$\bar{\rho} = \left\langle s(\boldsymbol{f}, \boldsymbol{w}, b)\right\rangle$ 为隐层神经元激活函数的期望值；ρ 为稀疏系数；γ 为稀疏性惩罚因子。

在代价函数的约束下，通过编码和解码过程的反复迭代，$\bar{\boldsymbol{f}}$ 中权值较大的特征即为SAE选择的主控因子。通过SAE计算实验分析，选取第二阶水压因子，第二、三阶温度因子以及第一、二个时效因子5维特征作为主控因素，不仅能减少后续预测模型的复杂度和运算量，而且可提升预测准确度。

1.2 基于Stacking集成学习的预测模型构建

集成学习理论的研究表明，单一模型只能从某些特征角度分析未来变形趋势，且由于随机性的存在，单一模型的泛化能力和数据适应性存在局限；而通过集成多个异质模型能有效提升对复杂非线性过程的表征能力和适应性。Stacking是目前应用较广的一种集成学习方法，通常包含基学习器和元学习器两级架构。基学习器既要选择预测性能优异的模型，以保证预测精度；也要考虑不同模型之间的差异性，以保证集成学习的泛化能力和鲁棒性。MLR、GM(1,1)和SVR模型各有侧重，其中MLR采用多元线性方程描述多个自变量与因变量之间的关系；SVR是贝叶斯理论下基于结构风险最小化准则得到的最优算法；GM(1,1)采用灰色微分方程对时间序列中隐藏的趋势信息进行挖掘和建模。元学习器需要泛化能力强和鲁棒性高，避免因个别基学习器的错误而影响整体性能，LSTM模型能挖掘时间序列中隐藏的长短期依赖信息，且对非线性非平稳随机过程具有较强的描述能力，因此本文选择LSTM作为Stacking集成学习的元学习器。

综合考虑，本文构建的Stacking集成学习模型以MLR、SVR和GM(1,1)为第一层基学习器，以LSTM为第二层元学习器，首先将主控因素输入基学习器中进行训练，得到对应的预测结果；再将预测结果输入元学习器进一步集成学习，获得最终预测结果。

2 实验结果与分析

2.1 实验数据

为验证本文方法的变形预测性能，采用我国中东部某混凝土大坝的变形监测数据开展验证实验。该大坝以发电为主，兼容下游防洪，蓄水总容量为6 430万m^3，年均径流量为43.24亿m^3，正常蓄水位为1 548 m。为监测大坝的水平位移，在坝体布设了2条正垂线和3条倒垂线，并设置5个监测点（J01~J05）监测各部位的水平位移。采集2021年1月—2022年12月5个监测点的数据，采样间隔为2 d，共获得300期有效数据。监测点J01记录的大坝水平位移变化曲线见图1，可以看出，变形数据曲线波动性较复杂，给变形预测带来了较大困难。实验将1~200期数据作为训练样本，201~300期数据作为测试数据，用于验证模型的预测性能。实验数据分析与处理采用Matlab R2022A软件平台，联想ThinkPad X1 Carbon，Intel 酷睿I9处理器，3.2 G主频，32 G内存，1 TB固态硬盘。

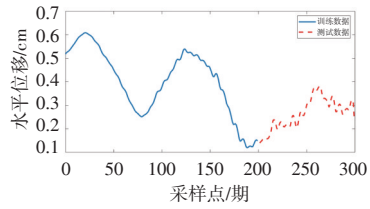


图1 监测点J01水平位移变化曲线

2.2 SAE特征选择结果

实验中将权值阈值设置为0.1，即选择权值大于0.1的特征作为主控因素。SAE特征选择结果见图2，可以看出，提取的19维特征中绝大部分对应的权值均小于0.1，即与大坝变形的关联性不强，属于冗余特征；因此H_2、T_2、T_3、I_1和I_2被SAE自动选择为主控因子。

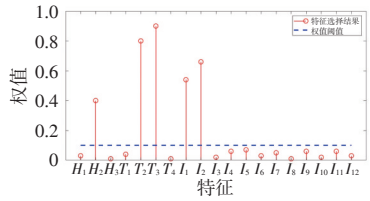
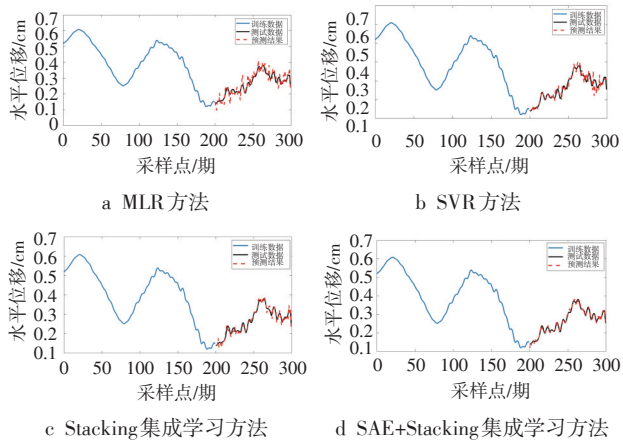


图2 SAE主控因子选择结果

2.3 预测结果与对比分析

本文对比分析了MLR、SVR、Stacking集成学习和SAE+Stacking集成学习等方法的预测结果。4种方法的预测结果以及与实际数据之间的误差见图3、4，可以看出，MLR方法的预测性能最差，其原因在于，实际变形数据是一种典型的非线性非平稳时间序列，而MLR是线性模型难以有效描述非线性变化趋势；SVR方法的预测性能优于MLR，且在预测初期精度较高，但随着预测时间的增加，预测误差出现了累积效应，预测结果发散；Stacking集成学习方法的预测性能优于SVR，这是由于Stacking集成学习通过融合多个预测模型实现了预测精度的提升，但输入数据中的冗余信息会对基学习器和元学习器的预测精度产生影响，因此其预测误差大于SAE+Stacking集成学习方法。



a MLR方法 b SVR方法

c Stacking集成学习方法 d SAE+Stacking集成学习方法

图3 4种方法的预测结果对比分析

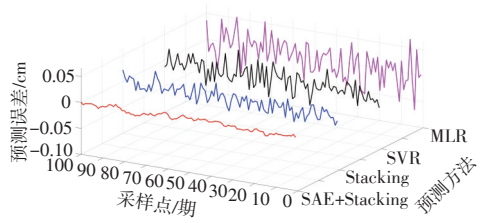
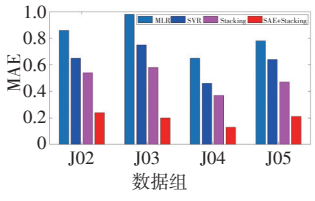


图4　4种方法的预测误差曲线

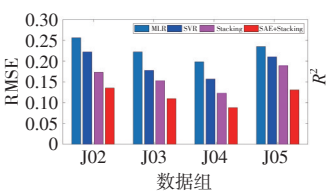
MLR、SVR、Stacking集成学习和SAE+Stacking集成学习方法的MAE分别为0.46、0.32、0.28、0.16 cm，RMSE分别为0.73、0.68、0.55、0.42 cm，R^2分别为0.866、0.925、0.946、0.973，可以看出，SAE+Stacking集成学习方法的3项指标均为最优。相对于MLR、SVR和Stacking集成学习方法，SAE+Stacking集成学习方法的MAE分别提升187.5%、100%和75%，RMSE分别提升73.8%、61.9%和30.9%，R^2分别提升10.7%、4.8%和2.7%，体现了其有效性与优越性。

2.4　泛化能力对比分析

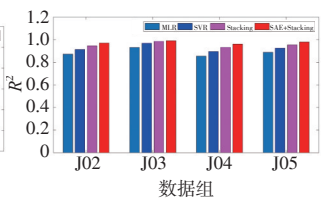
为进一步评估本文方法对不同数据的泛化能力和适应性，将上述4种方法应用于J02~J05四组数据，每组数据同样以1~200期作为训练样本，200~300期作为测试样本。4种方法每组数据的预测结果见图5，可以看出，本文方法对于J02~J05四组数据均可获得最优的预测性能。MLR、SVR、Stacking集成学习和SAE+Stacking集成学习方法的4组数据预测结果的RMSE均值分别为0.53、0.35、0.30、0.17 cm，MAE均值分别为0.78、0.64、0.46、0.38 cm，R^2分别为0.861、0.920、0.938、0.977，说明本文方法对不同数据预测结果的一致性最高，对不同数据的适应性和泛化能力最强。



a　MAE



b　RMSE



c　R^2

图5　不同方法对不同数据的预测结果

3　结　语

大坝变形监测数据是一种典型的非线性非平稳时间序列，基于单一模型的预测方法精度较低且泛化能力较弱，难以满足实际应用需求。本文将集成学习理论引入大坝变形预测领域，提出了SAE+Stacking集成学习方法。首先利用SAE自动筛选与大坝变形关联性较强的主控因子，剔除冗余信息以降低后续预测模型的复杂度和运算量；再以主控因子为输入，建立Stacking集成模型，通过集成多个异质模型的方式提升预测性能。实验结果表明，本文方法的预测精度更高、泛化能力更强，具有较广的应用前景。

参考文献

[1]　王瑞婕，包腾飞，李扬涛，等. 基于多因子融合和Stacking集成学习的大坝变形组合预测模型[J]. 水利学报，2024，54(4):497-506

[2]　原佳帆，李丹扬，李佳霖，等. 基于TimeGAN和CNN-BiLSTM-Attention的大坝变形预测混合模型[J]. 人民黄河，2024，46(12):127-130

[3]　冯子强，李登华，丁勇. 基于Blending-Clustering集成学习的大坝变形预测模型[J]. 水利水电技术(中英文)，2024，55(4):59-70

[4]　冯杭华. 基于ARIMA-TGM组合模型的边坡变形预测[J]. 地理空间信息，2022，20(10):98-102

[5]　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

[6]　潘国荣，乔立洋. 非等间距改进灰色模型在基坑变形预测中的应用[J]. 测绘地理信息，2019，44(1):9-13

[7]　邓思源，周兰庭，王飞，等. 大坝变形的XGBoost-LSTM变权组合预测模型及应用[J]. 长江科学院院报，2022，39(10):72-79

[8]　黄耀英，何一洋，沈振中，等. 大坝监测量最佳统计模型优选方法[J]. 水利学报，2022，53(2):154-164

[9]　胡毅，瞿博阳，梁静，等. 进化集成学习算法综述[J]. 智能科学与技术学报，2021，3(1):18-33

[10]　朱小韦，袁占良，李宏超. 基于BP-PCA-WCA-SVM的混凝土大坝变形预测方法[J]. 长江科学院院报，2024，41(9):138-145

[11]　刘丹，吕倩，胡少华，等. 大坝变形GA-LSTM组合预测模型研究[J]. 安全与环境学报，2023，23(7):2 246-2 253

[12]　毕自航，李素敏，张龙宇，等. 结合Prophet-CNN模型的多轨时序InSAR矿区三维形变监测预警[J]. 测绘通报，2024(5):53-59

[13]　张旭辉，郭际明，徐士月. 基于高斯过程的变形监测组合建模与预测[J]. 测绘地理信息，2023，48(4):70-73

[14]　Chen S Y，Gu C S，Lin C N，et al. Prediction of Arch Dam Deformation Via Correlated Multi-target Stacking[J]. Applied Mathematical Modelling，2021，91:1 175-1 193

[15]　Sircar A，Yadav K，Rayavarapu K，et al. Application of Machine Learning and Artificial Intelligence in Oil and Gas Industry[J]. Petroleum Research，2021，6(4):379-391