

基于IF-GAM的大坝异常监测数据识别算法研究

李毅男¹,陈建营¹,魏立巍²,吕佳凝²

(1.黄河勘测规划设计研究院有限公司,河南 郑州,450003;

2.黄河水利水电开发集团有限公司,河南 郑州,450003)

摘 要:随着数字孪生水利工程建设不断推进,安全监测作为数据底板构建的重点,是后续专业模型和业务应用的基础。然而,采集和传输过程中产生的异常数据影响了数据可靠性和完整性,因此识别和清除异常数据是提高监测数据质量、确保大坝安全评估有效性的关键。本研究采用广义相加模型(GAM)结合孤立森林算法(IF)进行大坝监测异常数据识别。首先通过IF识别监测数据中的明显异常值,在此基础上利用GAM提取监测数据序列的趋势项,对剔除趋势项信息的残余分量采用3 δ 方法进行精细化异常值识别,实例数据分析表明该方法可以有效识别大坝监测数据中的异常值。

关键词:安全监测;数据治理;异常值识别;孤立森林;广义相加模型

Title: Study of identification algorithm for abnormal dam monitoring data based on IF-GAM// by LI Yinan, CHEN Jianying, WEI Liwei and LYU Jianing// Yellow River Engineering Consulting Co., Ltd.

Abstract: With the continuous construction of digital twin hydraulic engineering, safety monitoring, as the focus of data base construction, is the basis of subsequent professional models and business applications. As the abnormal data generated during the collection and transmission process would affect the reliability and integrity of data, identifying and clearing abnormal data is the key to improve the quality of monitoring data and ensure dam safety assessment. Therefore, the generalized additive model combined with isolated forest algorithm is applied in the identification of abnormal data of dam monitoring. Firstly, the obvious outliers in monitoring data are identified by isolated forest algorithm. Then, the trend items of monitoring data series are extracted by generalized addition model, and trend item information is eliminated. And the 3 δ method is used for refined identificative of outliers. It is proved that the method can effectively identify the outliers in dam monitoring data through the example data analysis.

Key words: safety monitoring; data governance; abnormal data identification; isolated forest; generalized additive model

中图分类号:TV698.1

文献标志码:B

文章编号:1671-1092(2026)02-0037-05

1 研究背景

大坝安全监测数据质量是实现数字孪生工程安全“四预”的核心和关键^[1-2],为了精确掌握大坝的实际运行状态和结构稳定性,大坝安全监测系统依赖于一系列嵌入大坝结构内部或安装在大坝表面的专用监测设备,连续不断地收集关于大坝运行状态的各项关键指标数据。但是,在采集、传输及储存大坝安全监测数据的过程中,由于外部环境、设备故障、传输问题或人为错误等因素,会产生一些偏离正常模式、趋势或期望值的数据点^[3]。因此,对

这些异常值进行识别和剔除以提高观测数据质量具有十分重要的意义。

目前,为有效识别大坝安全监测数据中的异常值,通常采用传统的统计回归模型方法^[4-5]、包络线法^[6]、对比法^[7]等。其中,统计回归模型的优势在于检测准确性高,不易错判正常数据。但是,建立并优化回归模型的过程往往需要深厚的数据分析专业背景,且不同操作者在建模过程中的标准一致性难以保证。特别是在面对大量监测点位和长时间序列数据时,该方法的实施成本较高,响应速度受限,同时对输入数据(如环境因素:水位、降雨、温

度)的质量有较高要求,低质量数据可能会影响模型的表现。包络线法的核心在于设定一个正常数据应落入的区间,其依赖于数据变动过程中异常值与正常值之间的跳变关系,在面对因外部环境急剧改变引发的异常数据时,由于过分关注数据连续变化的一致性而导致辨识此类异常的能力降低。对比法则主要依靠逻辑推理规则或行业专家积累的实践经验来区分正常数据与异常数据,该方法虽然直观实用,但在转化为自动化执行的计算机程序时,因其高度依赖于经验和主观判断,难以实现有效的算法化和标准化表达。

针对上述算法存在的问题和大坝异常值识别的重要性,提出基于孤立森林算法(IF)和广义相加模型(GAM)的组合异常值识别算法,分析该算法在异常值识别过程中的可行性,并探讨该方法在各类型大坝监测数据中的适用性。

2 算法原理

2.1 孤立森林算法

孤立森林算法^[8]作为一种高效的异常检测方法,基于两个关键点观察:异常数据往往数量较少,且在特征空间中与大多数数据点有明显区别。这意味着异常数据通常分布较为稀疏,且距离高密度区域较远,使得它们更容易被“孤立”。通过随机采样和构建隔离树(iTree)的集合,即孤立森林(iForest),来检测异常数据。

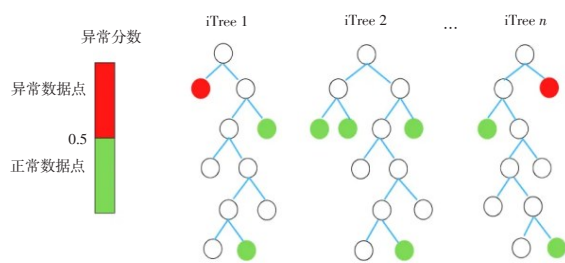


图1 孤立森林算法原理

Fig. 1 The principle of isolated forest algorithm

构建孤立森林的详细步骤如下:

(1)子采样集的选择:从包含连续性数据的训练集中随机选取 m 个样本数据点,形成子采样集 $D=\{d_1, d_2, \dots, d_m\}$,每个数据点具有 n 维特征,这些数据点作为构建树的根节点。

(2)随机分裂条件的选择:每一棵隔离树的构建开始于从原始数据集中随机抽取的一个较小子采样集 D ,对于这个子集,算法首先随机选取一个特征 A ,并在该特征的值域内随机确定一个切分点 p ,以此将数据分割成 D_{left} 和 D_{right} 两个子集,分别对应特征 A 值域内小于 p 和大于等于 p 的样本点。

(3)数据划分:根据数据点在维度 A 上的值 $d_i(A)$ 来划分数据,如果 $d_i(A) < p$,则该数据点被划分到左子树,否则划分到右子树。

(4)递归构建:重复步骤(2)和步骤(3),不断构造新的左、右子树,直到满足以下条件之一——子采样集 D 中只剩下一个数据点或所有数据点相同,无法进一步划分;隔离树的高度达到预定限制。

(5)森林构建:重复上述步骤,直至构建的隔离树数量达到预定数目 N ,这些隔离树共同组成孤立森林。

针对任何给定的查询数据 x ,通过计算其在孤立森林中每棵隔离树上的路径长度 $h(x)$,可以得到该数据点在孤立森林中的平均路径长度 $E(h(x))$ 。对于包含 m 个样本的子采样集 D ,隔离树的平均路径长度 $c(m)$ 可以视为等价于二叉搜索树中失败查询的平均路径长度,其计算公式如下:

$$c(m) = 2H(m-1) - \frac{2(m-1)}{m} \quad (1)$$

其中, $H(k)$ 表示调和数,可以近似为 $H(k) = \ln(k) + 0.577\,215\,664\,9$,式中 $0.577\,215\,664\,9$ 是欧拉常数。

查询数据 x 的异常分数 $s(x, m)$ 定义为:

$$s(x, m) = 2 \frac{-E(h(x))}{c(m)} \quad (2)$$

根据异常分数 s 的值,可以对查询数据 x 进行异常判断:当 $E(h(x))$ 接近 $c(m)$ 时, s 接近0.5,说明 x 的平均路径长度与标准平均路径长度相近,无法确定 x 是否为异常;当 $E(h(x))$ 接近0时, s 接近1, x 被识别为异常点;当 $E(h(x))$ 接近 $m-1$ 时, s 接近0, x 被识别为正常点。

2.2 广义相加模型

广义相加模型^[9]作为一种高度灵活的统计模型,扩展了广义线性模型(GLM)的概念,通过引入非线性平滑函数来允许响应变量与一个或多个预

测变量的关系以加和的形式呈现,并通过调整平滑参数来控制这些关系的弯曲程度。GAM特别适合解析和预测涉及复杂、非线性相互作用的数据场景,其核心在于使用平滑函数来捕捉变量之间的非线性关系。

GAM的一般形式可以表示为:

$$y_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_m(x_{im}) + \varepsilon_i \quad (3)$$

其中, y_i 是响应变量; $x_{i1}, x_{i2}, \cdots, x_{im}$ 是预测变量; β_0 是截距项; $f_1, f_2, \cdots, f_m$ 是预测变量的平滑函数,用于捕捉预测变量与响应变量之间的非线性关系; ε_i 是误差项。

在GAM中,异常点表现为与大多数数据点有显著不同的响应值,会在平滑函数的拟合中表现出异常残差。因此,通过评估数据点的残差可以识别异常点。

3 大坝异常监测数据识别模型建立

孤立森林算法通过隔离异常值或离群点进行识别,该方法可以快速有效识别序列数据中的明显异常值,但在精细化异常值识别的过程中,由于孤立森林算法为无监督训练模型,存在严重的误判情况。因此,首先利用IF识别监测数据中的明显异常值,主要包括明显突变异常值、超量程监测值,在此基础上,通过GAM算法拟合出整体数据的变化趋势,该拟合过程不受大量连续明显异常值的影响,并利用GAM算法再次校验IF算法识别的异常值,最终定位异常值的准确位置。

综合上述分析,组合IF+GAM大坝安全监测数据异常值识别算法的建立步骤如下(流程见图2)。

(1) 获取监测时序数据 $x(t) = \{x_1, x_2, \cdots, x_t\}$, 建立IF算法, 设定森林中隔离树数量、样本容量、最大深度, 计算每个数据点的异常分数来评估其异常程度。

(2) 将步骤(1)识别的异常值剔除, 得到初步异常值识别序列 $x_n(m) = \{x_1, x_2, \cdots, x_m\}$, 利用GAM算法建立响应变量 $x_n(m)$ 关于预测变量时间 t 的非线性模型 $x_n(m) = \beta_n + f_1(t_n) + \varepsilon_n$, 平滑函数选择高斯核平滑。

(3) 获取 $x_n(m)$ 的趋势项信息 $x'_n(m)$, 计算残差项 ε_n 。

(4) 评估每个数据点的残差项, 若残差不在设定的阈值区间内, 则认为该数据点为异常值, 阈值

计算公式为:

$$\delta = \pm 3\sigma = \pm 3 \sqrt{\frac{1}{m} \sum_j^m \left(\varepsilon_{nj} - \frac{\sum_i^m \varepsilon_{ni}}{m} \right)^2} \quad (4)$$

(5) 在步骤(1)~(4)处理结果的基础上, 再次建立GAM模型, 利用本次构建关于时间 t 的非线性模型, 计算所有被识别为异常值的数据点, 根据残差项验证该数据是否为异常值。

(6) 重复步骤(1)~(5)算法, 直到识别的异常值个数为0或本次累计识别的异常值占比与上次累计识别的异常值占比之差小于0.01。

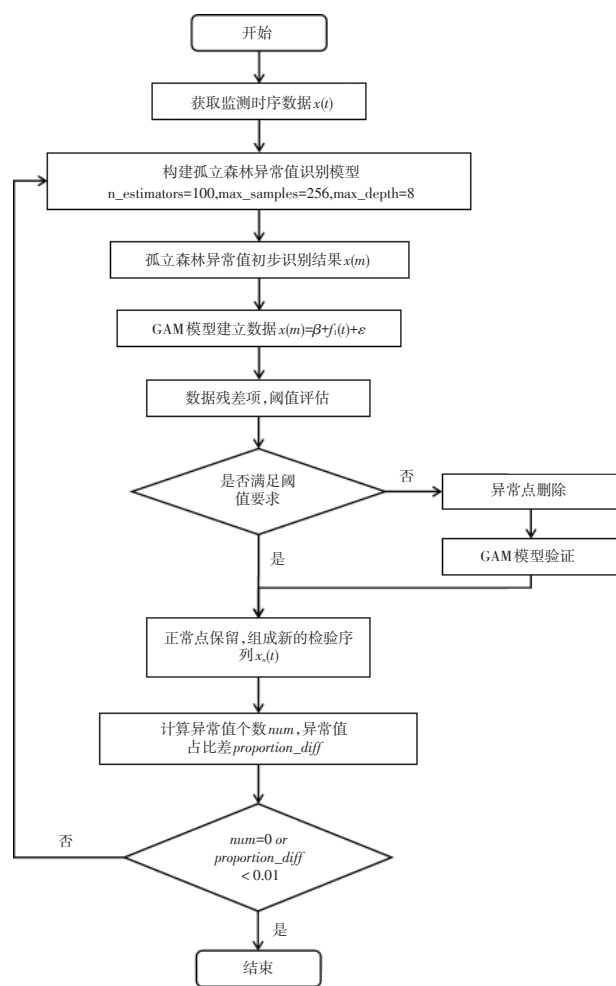


图2 算法流程

Fig. 2 Algorithm flow

4 实例分析

为验证本研究提出的大坝安全监测数据异

常值识别算法的有效性和泛化性,以某大型土石坝不同位置、不同类型的监测仪器数据为例,进行算法验证。该土石坝位于河南省境内,坝顶高程 281 m,正常高水位 275 m,库容 126.5 亿 m^3 ,2003 年开始大坝安全监测项目数据采集陆续实现自动化。

4.1 渗压监测数据识别效果

以防渗墙后重点渗压计的实测数据为例检验该算法异常值识别的有效性和精度,同时与单一 IF 算法和 GAM 算法进行对比分析。该渗压计从 1999 年 11 月起开始监测,截至目前已积累了 7 000 余条实测数据,由于监测周期长,数据量大,监测数据中存在大量异常值数据,共计有 1 306 个,严重影响了后期大坝质量安全评估和生产决策方案制定。图 3 为该渗压计的实测数据过程线。

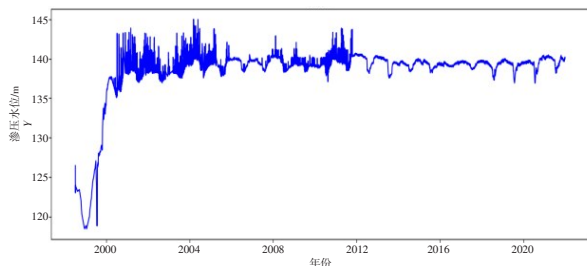


图3 渗压计实测数据过程线

Fig. 3 Measured data by osmometer

分别用 IF 算法、GAM 算法及 IF-GAM 组合算法对渗压计实测数据进行异常值识别处理。其中,IF 算法设置 $n_estimators=100$, $max_samples=256$, $max_depth=8$; GAM 算法采用高斯核平滑函数,平滑参数设置为 0.15,残差评估条件采用 3σ ; 组合算法采用相同的参数设置。各方法进行循环迭代异常值识别,直到本次循环异常值识别的数量为 0 或本次累计识别的异常值占比与上次累计识别的异常值占比之差小于 0.01,异常值识别结束。

各算法识别结果见表 1 和图 4~6。由图表可以看出,IF-GAM 组合算法异常值识别的准确率最高且误判率最低,可以精确定位识别监测数据序列中的异常值。部分正常值被误判为异常值,这是由于在循环迭代过程中,随着数据序列中的异常值被识别剔除,相邻数据点间的时间间隔被拉长,导致数据连续强度减弱,使拟合预测的数据趋势在一定区域出现不同程度的波动变化,从而造成正常值的偏离残差较大,被误判为异常值。

表 1 各算法异常值识别结果对比

Table 1 Comparison of outlier recognition results by different algorithms

识别算法	识别异常值个数	识别正确个数	误判个数	正确率	误判率
IF	980	705	275	53.98%	28.06%
GAM	1 536	951	585	72.82%	38.09%
IF-GAM	1 353	1 305	48	99.92%	3.55%

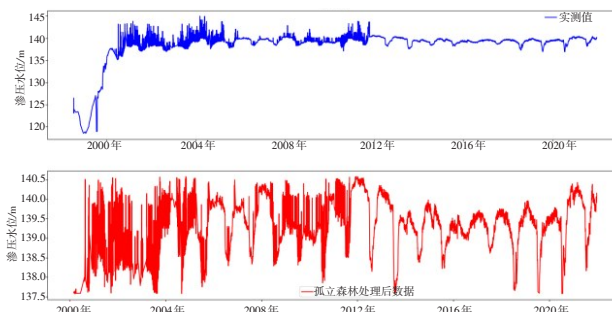


图4 IF异常值识别结果

Fig. 4 Outlier identification results by IF algorithm

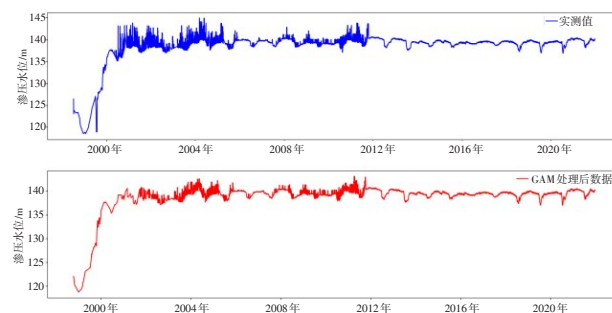


图5 GAM异常值识别结果

Fig. 5 Outlier identification results by GAM algorithm

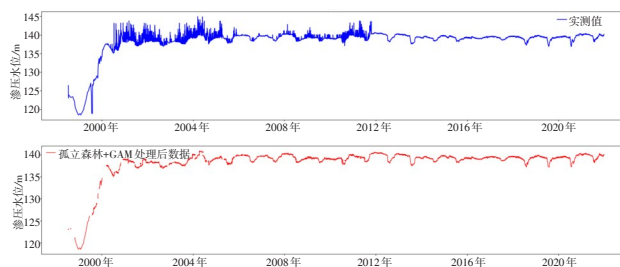


图6 IF-GAM异常值识别结果

Fig. 6 Outlier identification results by IF-GAM algorithm

4.2 变形监测数据识别效果

为进一步验证所提算法的泛化能力及其在不同数据类型上的适用性,对该大坝其他部位的静力水准、水平位移等实测数据进行异常值识别。结果表明,IF-GAM 组合算法能有效识别不同监测

序列数据中的异常值,还原监测数据的原始变化特征,为后续大坝运行安全分析提供有效的数据基础。

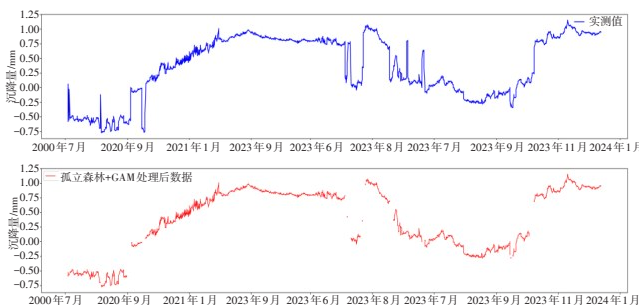


图7 静力水准IF-GAM异常值识别结果

Fig. 7 Outlier identification results of static leveling by IF-GAM

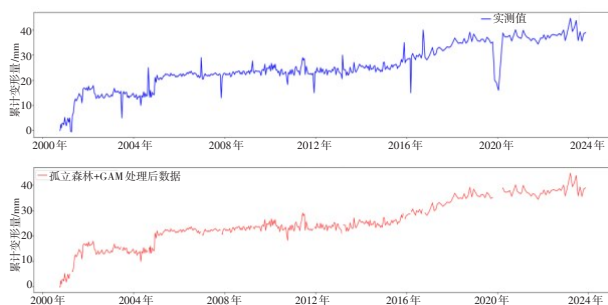


图8 水平位移IF-GAM异常值识别结果

Fig. 8 Outlier identification results of horizontal displacement by IF-GAM

5 结 语

依托具体工程项目的实际监测数据,深入研究了基于孤立森林(IF)与广义相加模型(GAM)组合算法的大坝安全监测数据异常值识别技术,并通过实证分析验证了该组合算法在大坝安全诊断领域的可行性和实用性。经过综合研究,得出以下结论:

(1) IF-GAM组合算法通过融合两种方法的优势,在大坝安全监测数据异常值识别中展现出高效性和广泛的适应性。IF-GAM模型在识别精度方面表现优异,误报率较低,尤其适用于处理具有复杂时空特征的大坝监测数据。

(2) 实例分析结果表明,IF-GAM组合算法在多次迭代识别和校验过程中,能够动态调整异常识别的阈值和残差标准,从而适应不同环境和监测条件下的数据信号波动特性。该算法在实际应用中表现出较强的泛化能力,不仅适用于单一类型的监测数据,还能广泛应用于多监测项目,为大坝安全管理和工程诊断提供可靠的数据支持。

参考文献:

- [1] 吴中如,顾冲时. 大坝安全综合评价专家系统[M]. 北京:科学技术出版社,1997.
- [2] 周波,唐桂彬. 大坝安全监测专家系统知识库的研究[J]. 计算机时代,2011(1):23-24,27.
- [3] 王译玲,丁勇,李登华. 耦合DWT-IFLOF的大坝监测数据异常检测算法[J]. 中国农村水利水电,2024(7):203-209.
- [4] 滕树强. 基于偏最小二乘回归的大坝安全监测数据异常识别模型研究[J]. 中国水运(下半月),2022,22(11):28-29,64.
- [5] 赵利锋,苏春荣,高彦国,等. 大黑汀水库主坝坝基21号坝段量水堰异常值原因分析[J]. 水利水电技术,2014,45(8):52-55,61.
- [6] 唐敏. 大坝安全监测异常值分析及计算机实现[D]. 武汉:武汉大学,2006.
- [7] 杨鸽,范振东,傅春江,等. 基于奇异谱分析的大坝安全监测数据异常值识别技术研究[J]. 水力发电,2021,47(8):125-129.
- [8] 赵新华,范振东,何宇,等. 基于数据重构与孤立森林法的大坝自动化监测数据异常检测方法[J]. 中国农村水利水电,2021(9):174-178.
- [9] Sullivan K J, Shadish W R, Steiner P M. An introduction to modeling longitudinal data with generalized additive models: applications to single-case designs[J]. Psychol Methods, 2015,20(1):26-42.

收稿日期:2024-10-17;修回日期:2024-11-27

作者简介:李毅男(1985—),男,河南郑州人,硕士,高级工程师,主要从事大坝安全监测设计和信息化系统开发工作。

作者邮箱:85304707@qq.com

欢迎投稿 欢迎订阅 欢迎刊登广告