

基于 Apriori 和 GP - XGBoost 的特高拱坝 变形缺失数据填补方法

吴诚姝^{1,2}, 陈波^{1,2}, 刘庭赫^{1,2,3}

(1. 河海大学 水利水电学院, 江苏 南京 210098; 2. 河海大学 水文水资源与水利工程科学国家重点
实验室, 江苏 南京 210098; 3. 中水东北勘测设计研究有限责任公司, 吉林 长春 130021)

摘要: 变形监测数据作为特高拱坝服役性态最直观的表征, 蕴藏着丰富的时空信息和演变规律, 对工程长治久安意义重大。然而, 多源多维的变形监测数据受仪器本身及外界因素影响, 往往存在数据缺失的现象, 会对接下来的数据分析工作造成干扰。针对大坝变形监测序列中的缺失数据, 基于 Apriori 关联规则算法挖掘测点变形在空间维度上的关联性, 得到目标测点的强关联测点, 随后以强关联测点的变形监测数据作为输入样本, 利用贝叶斯优化的 XGBoost 回归模型填补了目标测点的空缺变形监测序列。结合锦屏一级特高拱坝工程实例表明, 该填补方法实现了变形监测空缺信息的高效、精准填补, 可用于类似大坝工程的变形缺失数据填补。

关键词: 特高拱坝; 变形监测; 缺失数据填补; Apriori 关联规则; XGBoost 回归

中图分类号: TV698.1

文献标识码: A

文章编号: 1672-643X(2022)06-0151-08

Missing data filling method of ultra-high arch dam deformation based on Apriori and GP - XGBoost

WU Chengshu^{1,2}, CHEN Bo^{1,2}, LIU Tinghe^{1,2,3}

(1. College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing 210098, China; 2. State Key Laboratory of Hydrology - Water Resources and Hydraulic Engineering, Hohai University, Nanjing 210098, China;
3. Northeast Zhongshui Survey, Design and Research Corporation Limited, Changchun 130021, China)

Abstract: Deformation monitoring data can reflect the service behavior of ultra-high arch dams directly, they contain abundant spatio-temporal information and evolution law, which is of great significance to the long-term stability of the project. However, the multi-source and multi-dimensional deformation monitoring data are often affected by the monitoring instrument itself and other external factors, which will cause interference to the following data analysis. Regarding to the missing data in dam deformation monitoring sequence, the target monitoring sites with strong correlation were obtained based on the spatial correlation of monitoring site deformation calculated by Apriori association rule, and then the deformation monitoring data were adopted as the input samples to fill in the gaps in deformation monitoring sequence of the target monitoring sites using Bayesian optimized XGBoost regression model. The case study of Jinping ultra-high arch dam shows that this method can fill in the deformation monitoring gaps efficiently and accurately, so it is applicable to the filling of missing deformation data of similar dam projects.

Key words: ultra-high arch dam; deformation monitoring; missing data filling; Apriori association rule; XGBoost regression

1 研究背景

特高拱坝与近坝库岸山体构成一个巨型开放系统, 超高的坝高、超大的库容、复杂的地质条件及恶

劣的服役环境给工程安全保障带来极大挑战, 因而对大坝运行进行实时安全监测与评价尤为重要。随着监测技术水平的日益提高, 目前评定大坝运行状态的方法主要是采用埋设原位监测仪器获取能够表

收稿日期: 2022-04-14; 修回日期: 2022-07-22

基金项目: 国家自然科学基金项目(52079049、51739003)

作者简介: 吴诚姝(1999-), 女, 山西灵石人, 硕士研究生, 研究方向为水工结构安全监控。

通讯作者: 陈波(1986-), 男, 浙江绍兴人, 博士, 教授, 博士生导师, 研究方向为水工结构安全监控。

征大坝实时运行性态的监测数据,再通过数据处理和分析对大坝的运行状态作出评价。然而,由于监测数据量庞大,且受仪器本身及外界因素影响,监测信息中往往存在数据缺失的现象,会削弱监测数据的有效性,并对后期的数据分析工作带来困难。因此,以合理的方式从海量原始监测数据中集中挖掘潜在有意义的规律和模式,利用大数据技术完成对缺失数据的处理,能够为实际工程提供有价值的决策参考^[1-2]。

国内外学者针对缺失数据的填补方法做了相应研究,大致分为插值填补、回归填补、机器学习填补等几个方面。部分学者依据数据集在时间序列上的相关性取得了一些成果,如 Song 等^[3]分别采用插值填补、时间序列统计方法、RNN 和 LSTM 神经网络算法对缺失数据进行了填补,结果表明 LSTM 神经网络在填补长时间缺失数据方面有较大优势;李正欣等^[4]基于多元时间序列变量间的相关性进行了联合数据填补,该方法具有较高的填补精度和抗干扰能力;王一蓉等^[5]利用遗传优化算法获取最优的数据参数,如均值、方差等,在此基础上利用马尔科夫链蒙特卡罗法对缺失数据进行填补,有效提高了缺失数据填补的准确度。然而,上述缺失数据的填补方法仅从时间序列角度进行了相关分析,而未结合测点空间位置的邻近特性,尚存在一定缺陷。研究表明,通过空间数据挖掘技术可以寻找数据中隐藏的规律特征,如邱德俊等^[6]考虑大坝各测点的空间邻近关系,利用极限学习机拟合了缺失数值;Xu 等^[7]基于 FCM 聚类算法提出了实时更新隶属度矩阵的 OCS-FCM 水质监测缺失数据还原方法,结果表明该方法适用于缺失率大和多属性的数据集;郑霞忠等^[8]利用关联规则识别出大坝变形监测数据

中的强关联序列并结合 DBSCAN 聚类算法识别异常点,通过对比强关联序列来甄别由环境突变引起的异常值。有学者将空间数据挖掘技术引入到大坝安全监测领域内,发现了许多能够对工程安全运行决策起关键作用的因素,如利用关联规则研究工程运行影响因子的贡献程度^[9-11],利用聚类算法研究大坝运行的时空特性等^[12-14]。

基于上述缺失数据填补的研究现状,传统的依靠人工方法或统计学方法对大坝监测信息预处理在效率和精度方面均难以满足要求,而采用人工智能算法进行缺失数据填补仅考虑时间序列的相关性,在精度上具有一定的局限。据此,本文综合考虑特高拱坝监测信息的时空特性,首先利用 Apriori 空间关联规则挖掘各测点间的同步性,对测点数据的相关性进行分析,然后根据大坝各测点间监测数据的空间强关联信息,采用优化的 XGBoost (extreme gradient boosting) 回归模型进行目标测点的缺失测值填补。提出了从时间与空间维度融合关联规则挖掘和人工智能算法方面研究监测信息缺失的处理方法,有效提高了监测数据的可靠性和完整性,以期为特高拱坝服役性态分析做好基础性工作。

2 基于关联分析的监测信息空间挖掘

2.1 关联规则定义

假设存在数据集 $D = \{t_1, t_2, \dots, t_k, \dots, t_n\}$, $t_k = \{i_1, i_2, \dots, i_m, \dots, i_p\}$, $t_k (k = 1, 2, \dots, n)$ 称为事务, $i_m (m = 1, 2, \dots, p)$ 称为项。设 $I = \{i_1, i_2, \dots, i_m\}$ 为 D 中全体项组成的集合,则 I 的任何子集称为 D 中的项集。若 X, Y 为 D 中的两个项集,且 $X \cap Y = \emptyset$, 蕴涵式 $X \Rightarrow Y$ 成为关联规则, X, Y 分别称为 $X \Rightarrow Y$ 的前项和后项^[15]。关联规则常用的基本概念如表 1 所示。

表 1 关联规则中常见的基本概念

名称	说明
支持度	同时包含 X 和 Y 的事务数占事务集 D 中总事务数的比例: $\text{Sup}(X \Rightarrow Y) = P(X \cap Y) = \frac{n_{\text{count}}(X \cap Y)}{n_{\text{count}}(D)}$
置信度	同时包含 X 和 Y 的事务数占只包含 X 的事务数的比例: $\text{Con}(X \Rightarrow Y) = P(Y X) = \frac{P(X \cap Y)}{P(X)} = \frac{n_{\text{count}}(X \cap Y)}{n_{\text{count}}(X)}$
频繁项集	支持度大于或等于最小支持度 min_Sup 的项集。
强关联规则	若 $\text{Sup}(X \Rightarrow Y) \geq \text{min_Sup}$ 且 $\text{Con}(X \Rightarrow Y) \geq \text{min_Con}$ 时,称 $X \Rightarrow Y$ 为强关联规则。

2.2 Apriori 算法原理

Apriori 算法是使用频繁项集的先验知识从而生成关联规则的经典算法,该算法采用自底向上的遍历思想,主要使用逐层搜索的迭代方法挖掘大数据中潜在的事物联系,以确保关联规则的准确性^[16]。为提高算法效率,Apriori 算法利用了一个重要性质,即频繁项集的任一子集均为频繁项集。利用该性质可有效缩小频繁项集的搜索空间^[17]。

该算法建立关联规则主要包含两个步骤,一是通过迭代搜索事务集 D 中全部的频繁项集,非频繁项集在迭代过程中被剪枝剔除,直至不能搜索到任何频繁 k 项集为止;二是比较频繁项集间的最小置信度以建立强关联规则。该算法的主要流程如图 1 所示。

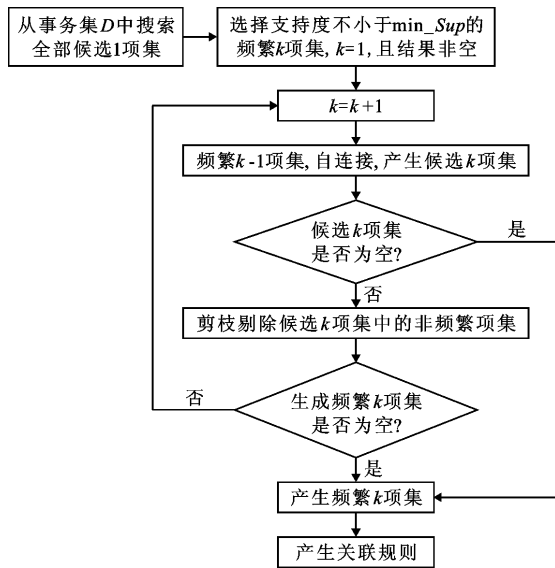


图1 基于 Apriori 算法的关联挖掘流程图

2.3 基于 Apriori 算法的大坝监测信息空间关联挖掘

为了评定大坝各测点监测信息的相关性,采用 Apriori 算法开展多测点空间关联挖掘,其流程如图 2 所示。

结合图 2 将大坝监测信息空间关联挖掘的主要步骤介绍如下:

(1) 设定前、后项。关联挖掘的对象主要是空间上位于不同位置的两个测点,同一时间截面内两个测点具有不同的监测效应量,因此将两个测点的监测效应量分别作为前项和后项,前项 X 表征某测点在 t 时刻的监测效应量信息,后项 Y 表征另一个测点在 t 时刻的监测效应量信息。

(2) 划分数据属性。由于大坝监测数据是连续的测值,难以反映定性的信息,因此在关联分析前,需要根据测值的特征将其归属于不同的属性区间,

这里结合监测数据的相似性特征,以测值的大小和增量作为衡量测值相似性的指标,按照测值的大小和增量划分属性区间如下:

假设大坝监测数据存在 T 个时间截面的测值,将测值特征数据集记为 D_T ,按照等距分箱的思想,结合特征数据集中的最大值 $D_{T_{\max}}$ 和最小值 $D_{T_{\min}}$ 将数据集 D_T 等距分为 3 个属性区间,即:

$$d = (D_{T_{\max}} - D_{T_{\min}}) / 3 \quad (1)$$

$$\begin{cases} a_1 = D_{T_{\min}} \\ a_2 = D_{T_{\min}} + d \\ a_3 = D_{T_{\max}} - d \\ a_4 = D_{T_{\max}} \end{cases} \quad (2)$$

由此生成测值特征的属性区间包括 $[a_1, a_2)$ 、 $[a_2, a_3)$ 、 $[a_3, a_4]$ 。

(3) 设定阈值。为了评定测点间的整体关联性,在支持度的基础上提出测点同步支持度的概念,定义如下:

假设存在测点 $\delta_1, \delta_2, \dots, \delta_n$, 共 T 个时间截面的监测数据,经步骤(2) 转换某一时间截面 t 的项集集合记为 $x_i (i = 1, 2, \dots, n)$, 则 n 个测点的同步支持度为:

$$\text{Sup}(\delta_1, \delta_2, \dots, \delta_n) = \frac{n_{\text{count}}(x_1 = x_2 = \dots = x_n)}{T} \quad (3)$$

通过同步支持度可以定量评定不同测点之间时空数据的相似性,在关联挖掘时需要设定最小同步支持度和最小置信度作为强关联测点的筛选条件。

(4) 挖掘强关联测点。最后当测点间满足设定的同步支持度和置信度阈值时,输出强关联测点;否则认为测点关联性较弱,予以删除。

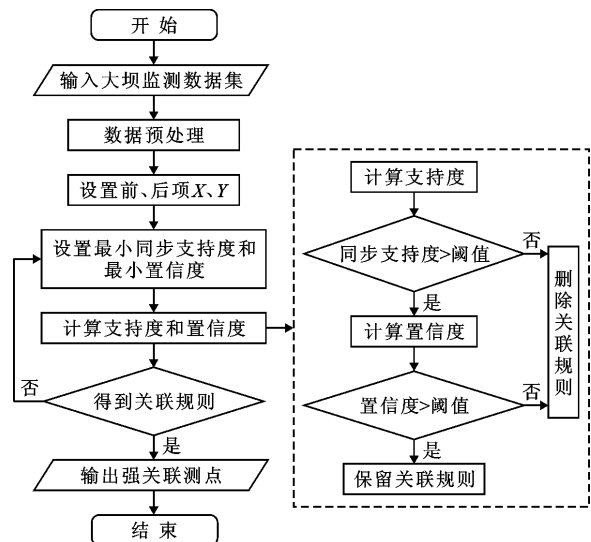


图2 大坝监测信息空间关联挖掘流程图

3 基于 GP - XGBoost 的监测信息空缺填补

3.1 基于关联测点的测值残缺填补

如图 3 所示,假设某混凝土拱坝上存在测点①、②、③,测点②存在残缺信息,经关联分析后建立了测点①→②,测点③→②的强关联规则,即考虑采用测点①、③对测点②的空缺数据进行填补,将测点②的监测序列表示为测点①和测点③的函数如下:

$$\delta_2 = g(\delta_1) + g(\delta_3) + \varepsilon_0 \quad (4)$$

式中: g 为测点①和测点③关于测点②的效应量函数; ε_0 为用测点①和测点③拟合测点②的残差。

传统的基于空间近邻的数据缺失填补方法仅考虑了测点的空间位置,未从测值本身特征去分析各测点间的相关性,且从空间位置上来看,缺失测值的目标测点附近布置了多个测点,很难评定选用哪些测点填补数据最为合适^[18]。为解决这个问题,本文首先对目标测点的邻近测点监测效应量进行关联分析,然后按照与目标测点的同步支持度进行排序,最后优先选取关联度高的测点对目标测点进行填补。为了进一步提高填补精度,采用贝叶斯优化的 XGBoost 模型对缺失值进行填补。

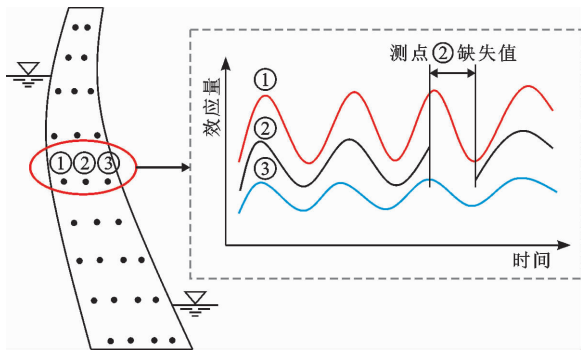


图 3 基于关联测点的测值残缺填补方法示意图

3.2 GP - XGBoost 算法原理

XGBoost 算法是由传统梯度提升树算法 (gradient boosting decision tree, GBDT) 发展而来,相比于 GBDT 算法,XGBoost 算法对损失函数加以改进,支持并行计算,大幅提高了算法效率^[19-20]。

本文主要利用 XGBoost 算法高效、精准的回归预测能力结合强关联测点的已知信息对目标测点的未知信息进行填补。XGBoost 算法的核心思想是多个基础模型的线性组合^[21],对于一棵含 t 个基础模型的集成树来说,可表示为:

$$\hat{y}_t^{(t)} = \sum_{k=1}^t f_k(x_i) = \hat{y}_t^{(t-1)} + f_t(x_i) \quad (5)$$

式中: $\hat{y}_t^{(t)}$ 为第 t 轮迭代后的模型预测值; $\hat{y}_t^{(t-1)}$ 为已知第 $t-1$ 个基础模型的预测值; $f_t(x_i)$ 为第 t 个基础模型,表示由多测点已知信息构造的决策树函数。

构建目标函数如下:

$$\begin{aligned} Obj^{(t)} &= \sum_{i=1}^n L(y_i, \hat{y}_t^{(t)}) + \sum_{j=1}^t \Omega(f_j) \\ &= \sum_{i=1}^n L(y_i, \hat{y}_t^{(t-1)} + f_t(x_i)) + \sum_{j=1}^t \Omega(f_j) \end{aligned} \quad (6)$$

其中, $\Omega(f_j)$ 为正则项,可表示为:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2 \quad (7)$$

式中: γ 和 λ 为正则化系数; T 为叶子节点的个数; ω 为叶子节点的权重。

对函数式(6)中损失函数 L 作二阶泰勒展开:

$$f(x + \Delta x) \approx f(x) + f(x)' \Delta x + \frac{1}{2} f(x)'' \Delta x^2 \quad (8)$$

令损失函数中 $\hat{y}_t^{(t-1)}$ 项为泰勒公式中的 x ,则目标函数可近似表示为:

$$\begin{aligned} Obj^{(t)} &\approx \sum_{i=1}^n [L(y_i, \hat{y}_t^{(t-1)}) + g_i f_t(x_i) + \\ &\quad \frac{1}{2} h_i f_t(x_i)^2] + \Omega(f_t) \end{aligned} \quad (9)$$

式中: g_i 和 h_i 分别为损失函数 $L[y_i, \hat{y}_t^{(t-1)}]$ 关于 $\hat{y}_t^{(t-1)}$ 的一阶导数和二阶导数。

g_i 和 h_i 可以表示为:

$$\begin{cases} g_i = \frac{\partial L(y_i, \hat{y}_t^{(t-1)})}{\partial \hat{y}_t^{(t-1)}} \\ h_i = \frac{\partial^2 L(y_i, \hat{y}_t^{(t-1)})}{\partial \hat{y}_t^{(t-1)^2}} \end{cases} \quad (10)$$

通过迭代找到使目标函数最大化降低的 f_t ,即完成模型的训练过程。

为了进一步提高 XGBoost 模型的精度,本文结合基于高斯过程 (Gaussian Process, GP) 的贝叶斯优化算法优化 XGBoost 模型的重要参数,以确定最佳的回归模型。贝叶斯优化框架是一个迭代的过程,主要通过采集函数选择下一个最佳评估点 x_t 并计算目标函数值 y_t ,更新历史观测集和概率代理模型,以此不断循环得到最优解^[22-23]。贝叶斯优化框架最核心的两部分是概率代理模型和采集函数,概率代理模型用于代理目标函数,其中迭代速度快、扩展性高的高斯过程应用较为广泛,考虑到大坝监测数据集庞大,以高斯过程作为优化内核可以明显提高效率^[24];采集函数由后验概率分布决定,作用是选择下一个潜力最高的评估数据。以下详细介绍概率代理模型 (高斯过程) 和采集函数具体内容:

(1) 高斯过程。在以高斯过程作为内核进行参数优化时,认为目标函数 $f(x)$ 满足 GP 过程,构成如下:

$$f(x) \sim \text{GP}(E(f(x)), k(x, x')) \quad (11)$$

式中: $E(x)$ 为期望,通常取值为 0; $k(x, x')$ 为 x 的协方差。

假设一个满足均值为 0 的先验分布 $p(f|X, \theta)$:

$$p(f|X, \theta) = N(0, \sum) \quad (12)$$

式中: X 为训练集 $\{x_1, x_2, \dots, x_t\}$; f 为目标函数值集合 $\{f(x_1), f(x_2), \dots, f(x_t)\}$; $\sum$ 为 $k(x, x')$ 构成的协方差矩阵 ($\sum_{i,j} = k(x_i, x_j)$); θ 为超参数。

考虑存在满足相同高斯分布的观测噪声 ε , 得到似然分布:

$$p(\varepsilon) = N(0, \sigma^2) \quad (13)$$

$$P(y|f) = N(f, \sigma^2 I) \quad (14)$$

式中: y 为观测值集合 $\{y_1, y_2, \dots, y_t\}$ 。

进一步地,结合公式(12)和(14),可以得到边际似然分布:

$$\begin{aligned} p(y|X, \theta) &= \int P(y|f)p(f|X, \theta)df \\ &= N(0, \sum + \sigma^2 I) \end{aligned} \quad (15)$$

根据高斯过程,存在如下联合分布:

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sum + \sigma^2 I & K_* \\ K_*^T & K_{**} \end{bmatrix}\right) \quad (16)$$

$$K_*^T = \{k(x_1, X_*), k(x_2, X_*), \dots, k(x_t, X_*)\} \quad (17)$$

$$K_{**} = k(X_*, X_*) \quad (18)$$

式中: f_* 为函数预测值; X_* 为输入值。

根据公式(16),得到预测分布如下:

$$p(f_*|X, y, X_*) = N(\langle f_* \rangle, \text{Cov}(f_*)) \quad (19)$$

$$\langle f_* \rangle = K_*^T [\sum + \sigma^2 I]^{-1} y \quad (20)$$

$$\text{Cov}(f_*) = K_{**} - K_*^T [\sum + \sigma^2 I]^{-1} K_* \quad (21)$$

式中: $\langle f_* \rangle$ 为预测均值; $\text{Cov}(f_*)$ 为预测协方差。

(2) 采集函数。采集函数作为评估潜力参数的重要依据,通常优化目标便是在全集参数 A 中寻找 $f(x)$ 达到极值时的 X 集合,其表达式如下:

$$x^* = \arg \max_{x \in A} f(x) \quad (22)$$

为了对贝叶斯参数优化过程中的模型精度进行定量评价,本文以 K -折交叉验证产生的均方误差 MSE 作为训练过程中模型精度的评价指标。

假设通过关联分析得到目标测点 δ_{n+1} 的 n 个强关联测点 $\delta_1, \delta_2, \dots, \delta_n$, 将 n 个强关联测点的测值作

为输入,目标测点 δ_{n+1} 的测值作为输出,则建立的 GP-XGBoost 算法的训练优化过程如图 4 所示。

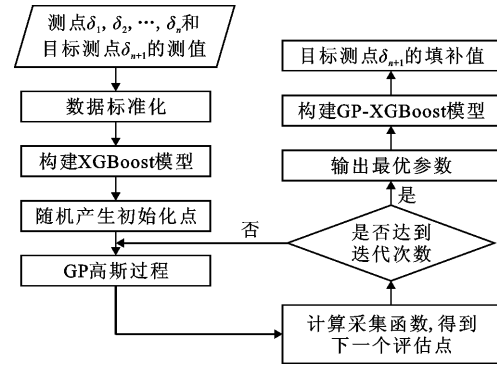


图4 GP-XGBoost 算法的训练优化流程图

4 工程实例

锦屏一级水电站特高拱坝地处深山峡谷河段,坝基高程为 1 580.00 m,最大坝高为 305.00 m,正常蓄水位为 1 880.00 m,属于大(1)型工程。该工程目前处于稳定运行阶段,为了监测坝体变形,分别在 5[#]、9[#]、11[#]、13[#]、16[#]和 19[#]坝段不同高程布置垂线测点,如图 5 所示,收集了该特高拱坝近 5 年的水平径向位移监测数据进行研究。

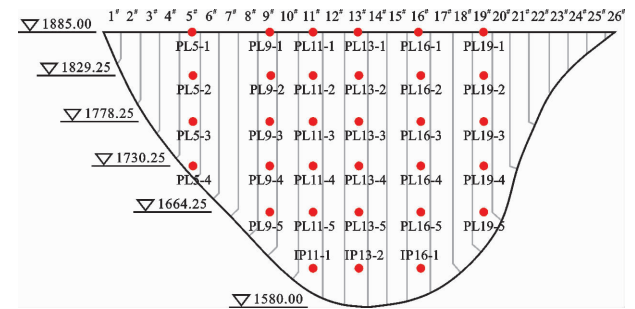


图5 锦屏一级水电站特高拱坝正垂线测点布置(高程单位:m)

4.1 测点变形序列关联分析

基于 Apriori 算法进行测点序列的关联分析时,首先从空间位置上选取与目标测点位置相邻的多个测点,即以目标测点为中心,关联分析窗口取目标测点相邻的 8 个测点(图 6),分别计算与中心目标测点的关联程度。首先将变形数据从定量测值转换为定性属性,考虑用变形测值大小和增量衡量测点变形序列的关联程度,按照测值大小划分为 A、B、C 3 个属性区间,分别表示变形较大、变形一般和变形较小;又按照增量情况划分为 a、b、c 3 个属性区间,分别表示变形相对上一时刻增大、不变和减小,在此基础上展开关联挖掘工作。

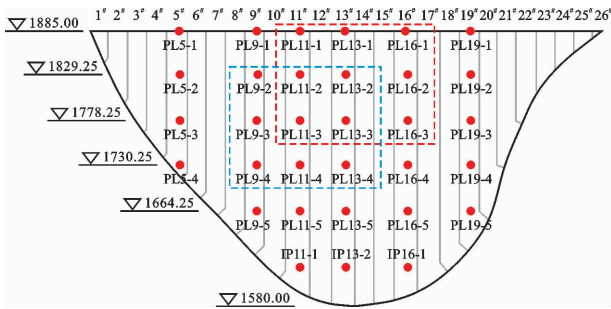


图6 关联分析测点选取窗口示意图(高程单位:m)

以 PL11-3 测点作为目标测点,选取 2016 年 1 月至 2021 年 3 月共 1 791 组水平径向位移监测数据,分别与其空间相邻测点 PL9-2、PL9-3、PL9-4、PL11-2、PL11-4、PL13-2、PL13-3、PL13-4 进行关联分析,计算结果如表 2、图 7 所示。结果表明,当设置最小同步支持度为 85% 时,与 PL11-3 测点强关联测点为 PL11-2、PL9-3、PL13-3 和 PL11-4;当设置最小同步支持度为 90% 时,强关联测点为 PL11-4。

为了进一步验证强关联测点的可信度,计算最小同步支持度为 85% 时的置信度,设置最小置信度为 80%,表 3 为 PL11-3 测点与强关联测点置信度计算表。由表 3 可见,PL11-4 和 PL13-3 测点满足置信度要求,PL11-2 和 PL9-3 测点在属性集 Ba 区间与目标测点 PL11-3 的关联规则略小于最小置信度,但

考虑到这两个测点出现在 Ba 属性集所占的比例较低,因此保留这两个强关联测点,用于填补目标测点 PL11-3 的残缺数据。

表2 各测点关联同步支持度计算表 %

目标测点	相邻测点	同步支持度
PL11-3	PL11-4	92.58
PL11-3	PL11-2	89.18
PL11-3	PL9-3	85.26
PL11-3	PL13-3	85.15
PL11-3	PL13-2	84.12
PL11-3	PL13-4	82.49
PL11-3	PL9-2	80.62
PL11-3	PL9-4	79.87

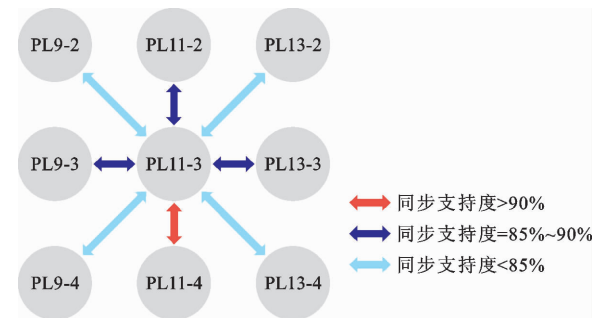


图7 基于 Apriori 算法的关联挖掘结果示意图

表3 PL11-3 测点与强关联测点置信度计算表 %

X/Y 划分 属性	PL11-4		PL11-2		PL9-3		PL13-3	
	$X \Rightarrow Y$	$Y \Rightarrow X$	$X \Rightarrow Y$	$Y \Rightarrow X$	$X \Rightarrow Y$	$Y \Rightarrow X$	$X \Rightarrow Y$	$Y \Rightarrow X$
Aa	88.89	97.44	87.34	88.46	87.34	88.46	82.80	91.55
Ab	/	/	/	/	/	/	/	/
Ac	98.17	95.54	90.12	97.77	90.12	97.77	87.95	95.17
Ba	95.51	85.86	75.00	78.79	75.00	78.79	87.76	80.37
Bb	/	/	/	/	/	/	/	/
Bc	97.49	94.17	81.71	97.57	81.71	97.57	86.41	89.00
Ca	96.71	91.98	93.52	90.41	93.52	90.41	87.52	85.33
Cb	/	/	/	/	/	/	/	/
Cc	92.49	94.30	94.11	87.82	94.11	87.82	86.63	83.62

注:“/”表示关联规则中的非频繁项集或未出现的属性,不参与置信度的计算。

4.2 GP-XGBoost 模型对缺失值的填补结果

目标测点 PL11-3 存在部分连续的残缺信息,为了验证本文提出数据填补方法的有效性,首先假定该测点在 2020 年 12 月 20 日测点数据丢失,测值为 42.540,其他关联测点数据集较完整。基于上述关

联分析的结果,按照与目标测点 PL11-3 关联度从高到低的顺序,依次选取不同数量的测点变形序列作为 XGBoost 模型的输入样本,输出为目标测点 PL11-3 除缺失数据以外的完整数据,建立 XGBoost 模型进行训练,再采用训练好的模型对 PL11-3 测点缺失

数据进行填补。

XGBoost 模型基于 python 平台进行构建,主要参数 max_depth (基模型包含最大深度) 和 $n_estimators$ (基模型数量) 对 XGBoost 模型的回归精度起着关键作用,故采用贝叶斯优化,迭代 100 次,图 8(a)、8(b) 分别为贝叶斯优化和随机搜索过程中均方误差 MSE 的变化曲线。由图 8 可见,贝叶斯优化每次迭代过程的 MSE 均较小,具有主动寻优的能力,而随机搜索由于参数的随机组合会出现局

部 MSE 较大的情况。贝叶斯优化在迭代 23 次后 MSE 达到最小,为 $7.64 \times 10^{-6} \text{ mm}^2$,而随机搜索在迭代 68 次后 MSE 才达到最小,为 $2.60 \times 10^{-5} \text{ mm}^2$ 。因此,贝叶斯优化采用贝叶斯优化参数能够高效寻优,具有良好的表现。

基于 4.1 节关联分析的结果,按照与目标测点 PL11-3 关联程度由高到低的顺序选取不同的学习样本对其缺失值进行填补,XGBoost 参数由贝叶斯优化,结果如表 4 所示。

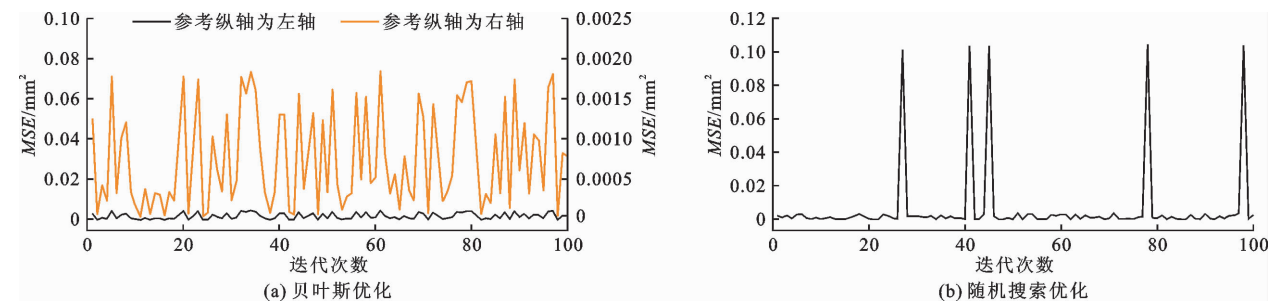


图 8 贝叶斯优化和随机搜索过程中 MSE 的变化曲线

表 4 不同学习样本对目标测点 PL11-3 缺失数据的填补结果

序号	学习样本	模型参数	XGBoost 估计值/mm
		$\text{max_depth}/n_estimators$	
1	{ PL11-2、PL11-4 }	23/56	42.535
2	{ PL11-2、PL11-4、PL9-3 }	31/62	42.518
3	{ PL11-2、PL11-4、PL9-3、PL13-3 }	49/68	42.493
4	{ PL11-2、PL11-4、PL9-3、PL13-3、PL13-2 }	13/44	42.494
5	{ PL11-2、PL11-4、PL9-3、PL13-3、PL13-2、PL13-4 }	17/100	42.478
6	{ PL11-2、PL11-4、PL9-3、PL13-3、PL13-2、PL13-4、PL9-2 }	15/38	42.479
7	{ PL11-2、PL11-4、PL9-3、PL13-3、PL13-2、PL13-4、PL9-2、PL9-4 }	13/66	42.462
8	{ PL9-2、PL9-4 }	15/50	42.820

由表 4 可见,当采用强关联测点 PL11-2、PL11-4 作为学习样本时,填补值与缺失值(42.540)最为接近,即精度最高。除此以外,采用其他的强关联测点作为填补测点时,也能够取得较好的填补结果,如 { PL11-2、PL11-4、PL9-3 }、{ PL11-2、PL11-4、PL9-3、PL13-3 } 等。值得注意的是,当学习样本中出现邻近的非关联测点时,会导致模型精度产生不同程度的降低,影响填补值的精度,如当学习样本为与目标测点关联程度最低的 PL9-2 和 PL9-4 测点时,模型的估计值为 42.820,是估计偏差最大的情况。因此在采用空间邻近关系填补缺失测值之前进行关联挖掘是很有必要的。

表 5 给出了不同方法对测点 PL11-3 的数据缺

失填补结果比较,可见本文采用的贝叶斯优化 XGBoost 算法的填补精度要优于传统的插值方法。

表 5 采用不同方法对目标测点 PL11-3 数据缺失的填补结果比较

缺失真实值	mm		
	线性插值	样条插值	XGBoost 插值
42.540	42.443	42.480	42.535

大坝监测数据中的粗差被剔除后,会出现多个单点空缺,另外,由于监测仪器损坏或环境等因素影响,监测数据中也会存在连续的残缺信息,因此采用本文提出的方法,利用强关联测点 PL11-2、PL11-4 对目标测点 PL11-3 存在的多段连续空缺测值进行填补,结果见表 6。

表 6 基于 GP-XGBoost 的测点 PL11-3 多测值填补结果

日期	原始测值/mm	填补测值/mm
2016-03-24	25.73	25.731
2016-03-25	25.53	25.532
2016-03-26	25.24	25.246
2016-08-21	32.32	32.325
2016-08-22	31.83	31.831
2016-08-23	31.52	31.524
2016-08-24	31.30	31.306
2016-08-25	31.07	31.073
2016-08-26	30.84	30.845
2017-06-21	17.78	17.781
2017-06-22	18.54	18.544
2017-06-23	19.27	19.278
2017-06-24	20.08	20.080
2019-12-21	43.07	43.072
2019-12-22	43.17	43.176
2019-12-23	43.28	43.284
2019-12-24	43.35	43.354
2019-12-25	43.38	43.381

5 结 论

针对传统缺失测值填补效率与精度的不足,借助关联挖掘规则,构建了目标测点空间分布的强关联信息,以此为基础利用优化的回归模型进行数据拟合,实现了从时空角度结合关联规则挖掘、机器学习等技术手段对特高拱坝长期变形缺失数据的高效精准填补,为大坝安全监控信息的挖掘、建模及评价工作提供了可靠基础,主要成果如下:

(1)考虑各测点间的空间邻近关系,引进空间关联规则,基于 Apriori 算法挖掘了目标测点的邻近测点监测效应量在空间维度上的关联性,按照与目标测点的同步支持度排序结果,优选出与目标测点关联度高的测点,获取了测值填补的强关联信息。

(2)结合基于高斯过程的贝叶斯优化算法优化 XGBoost 模型的重要参数,确定了最佳的回归模型。根据强关联测点的已知信息,利用 XGBoost 算法高效、精准的回归预测能力对目标测点中的残缺测值进行填补,高精度还原了变形监测数据中的缺失信息,保障了监测信息的完整性与准确性,为后续的分析工作提供了高质量的数据基础。

(3)以锦屏一级水电站特高拱坝 PL11-3 测点为

例,可得采用强关联测点作为学习样本时,填补值与缺失值最为接近,当学习样本中出现邻近的非关联测点时,会导致模型精度产生不同程度的降低,进而验证了采用空间关联挖掘填补缺失测值的必要性。

参考文献:

- [1] 曾 晖. 大数据挖掘在工程项目管理中的应用[J]. 科技进步与对策, 2014, 31(11): 46-48.
- [2] 渠孟飞. 基于数据挖掘的三峡库区滑坡整治工程设计参数估计研究[D]. 成都: 西南交通大学, 2017.
- [3] SONG Wei, GAO Chao, ZHAO Yue, et al. A time series data filling method based on LSTM—Taking the stem moisture as an example[J]. Sensors, 2020, 20(18): 5045.
- [4] 李正欣, 张凤鸣, 王 瑛, 等. 多元时间序列缺失数据填补方法[J]. 系统工程与电子技术, 2018, 40(1): 225-230.
- [5] 王一蓉, 王瑞杰, 陈文刚, 等. 基于遗传优化的调控系统缺失数据填补算法[J]. 电力系统保护与控制, 2016, 44(21): 182-186.
- [6] 邱德俊, 周 洋, 仲静文, 等. 基于空间邻近点与极限学习机的大坝位移缺失数据补齐[J]. 水力发电, 2021, 47(12): 98-101.
- [7] XU Huan, SUN Guangpei, JIANG Peng, et al. Water quality monitoring missing data filling method based on improved OCS-FCM[C]// 2019 Chinese Automation Congress (CAC). New York, USA: IEEE, 2019.
- [8] 郑霞忠, 陈国梁, 邹 韬. 考虑时间序列关联的大坝监测异常数据清洗[J]. 水力发电, 2020, 46(4): 111-114+125.
- [9] 毛宁宁, 苏怀智, 高建新. 基于 FP-growth 的大坝安全监测数据挖掘方法[J]. 水利水电科技进展, 2019, 39(5): 78-82.
- [10] 苏振华, 周宜红, 赵春菊, 等. 基于数据挖掘技术的溪洛渡大坝施工期温度监测数据分析[J]. 水电能源科学, 2016, 34(3): 70-73.
- [11] 商 桑, 赵春菊, 周宜红, 等. 溪洛渡拱坝施工期混凝土中期降温速率与通水冷却参数关系的数据挖掘模型[J]. 长江科学院院报, 2019, 36(6): 116-120.
- [12] CHEN Bo, HU Tianyi, HUANG Zishen, et al. A spatio-temporal clustering and diagnosis method for concrete arch dams using deformation monitoring data[J]. Structural Health Monitoring, 2018, 18(5-6): 1355-1371.
- [13] 陈 悦, 汪 程, 尹文中. 基于 Kohonen 聚类的特高拱坝变形分区[J]. 三峡大学学报(自然科学版), 2019, 41(1): 1-4.
- [14] 胡添翼. 基于面板数据分析方法的混凝土拱坝变形数据时空聚类模型[J]. 长江科学院院报, 2021, 38(2): 39-45.
- [15] 李 俊, 王 康, 张 群. 基于关联规则的大坝变形性态预判模型[J]. 西北水电, 2022(1): 73-77.

(下转第 166 页)

- factor of multiple crack-hole problem[J]. Theoretical and Applied Fracture Mechanics, 2020, 107: 102535.
- [11] 王强胜,李孝滔,咎晓东,等. 分布位错法研究移动赫兹压力作用下次表面裂纹的力学行为[J]. 表面技术, 2019, 48(6): 252–260.
- [12] 郑安兴,罗先启,钱镜林. 基于互补理论和扩展有限元法的摩擦接触裂纹应力强度因子计算[J]. 固体力学学报, 2020, 41(1): 50–58.
- [13] 杨恩光,杨立云,胡恒宁,等. 单轴压缩荷载下闭合裂纹扩展的试验和数值研究[J]. 岩土力学, 2022, 43(S1): 613–622.
- [14] 李立,焦敬品,吴斌,等. 结构中微裂纹与超声波的混频非线性作用数值仿真研究[J]. 计算力学学报, 2018, 35(6): 725–731.
- [15] 张玉良,吴必胜,赵高峰. 基于声发射监测的岩石热损伤实时演化研究[J]. 中南大学学报(自然科学版), 2021, 52(8): 2945–2958.
- [16] LIU Xiaohui, ZHAO Xiaoping, ZHANG Shishu, et al. Research on the failure evolution process of rock mass base on the acoustic emission parameters[J]. Frontiers in Physics, 2021, 9(8): 1–15.
- [17] REN Zhangyu, XIE Huimin, JU Yang. Determination of the stress and strain fields in porous structures by photoelasticity and digital image correlation techniques[J]. Polymer Testing, 2021, 102: 107315.
- [18] ZHANG Peng, DU Chengbin, BIRK C, et al. A scaled boundary finite element method for modelling wing crack propagation problems[J]. Engineering Fracture Mechanics, 2019, 216: 106466.
- [19] YI Cai, WANG Dong, ZHOU Lu, et al. A simulation investigation on the influence of pantograph crack defect on graphite contact strip wear[J]. Engineering Failure Analysis, 2022, 131: 105889.
- [20] WANG Lanwen, SHENG Xuanyu, LUO Jianbin. A peridynamic frictional contact model for contact fatigue crack initiation and propagation[J]. Engineering Fracture Mechanics, 2022, 264: 108338.
- [21] VALVO P S. A physically consistent virtual crack closure technique accounting for contact and interpenetration[J]. Procedia Structural Integrity, 2020, 28: 2350–2369.
- [22] ANKIT A. A sideface traction approach for cohesive and frictional crack growth problems using SBFEM[J]. Computer Methods in Applied Mechanics and Engineering, 2021, 386: 114076.
- [23] 申伟,郑芳彤,李冬生. 钢桥面板疲劳裂纹损伤非线性 Lamb 波检测的数值模拟与试验研究[J]. 土木与环境工程学报(中英文), 2021, 43(3): 135–141.
- [24] 倪敏,苟小平,王启智. 压缩双裂纹和单裂纹圆孔板应力强度因子公式[J]. 力学学报, 2013, 45(1): 94–102.
- [25] 祝青钰,周锐. 椭圆孔边任意长度双裂纹复合型应力强度因子复变函数解[J]. 机械强度, 2019, 41(4): 983–987.
- [26] 于耀庭,何芝仙,陈曦. 基于权函数法的齿根裂纹应力强度因子求解[J]. 机械设计, 2020, 37(9): 71–77.

(上接第 158 页)

- [16] 王露. 大坝施工期温度监测数据关联规则挖掘与预测[D]. 北京:清华大学, 2017.
- [17] 王丹丹. 基于 Apriori 的关联规则挖掘算法的研究[D]. 南京:河海大学, 2018.
- [18] 黄潇霏. 混凝土坝健康状态的实测资料时空诊断方法[D]. 南京:河海大学, 2018.
- [19] 闵恺艺. 基于多测点模型的面板堆石坝沉降变形预测研究[D]. 西安:西安理工大学, 2021.
- [20] 杨晨蕾,包腾飞. 基于 FCM-XGBoost 的大坝变形预测模型[J]. 长江科学院院报, 2021, 38(8): 66–71.
- [21] 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.
- [22] 崔佳旭,杨博. 贝叶斯优化方法和应用综述[J]. 软件学报, 2018, 29(10): 3068–3090.
- [23] 杨洋. 贝叶斯优化和关联规则挖掘的若干问题研究[D]. 北京:清华大学, 2020.
- [24] 李亚茹. 面向超参数估计的贝叶斯优化算法研究[D]. 杭州:浙江科技学院, 2021.