

基于孤立森林算法的压力钢管实时 在线人工智能分析

尚鑫¹, 方超群², 胡木生³

(1. 国网新源控股有限公司北京十三陵蓄能电厂, 北京 102200;

2. 水利部水工金属结构质量检验检测中心, 河南 郑州 450044;

3. 中国南水北调集团有限公司, 北京 100097)

摘要:为实现水利水电工程压力钢管实时在线的人工智能状态分析,在对压力钢管实时在线监测数据的结构特点进行分析的基础上,探索在健康状态监测时适用的人工智能算法,并在对无监督机器学习算法的讨论中,确定采用孤立森林算法进行实时在线人工智能分析。通过对北京十三陵抽水蓄能电站1号引水系统压力钢管1[#]支管和2[#]支管监测数据计算后表明,孤立森林算法在不需要对传感器实时数据进行预处理情况下,具备良好的异常发现能力,且运行速度较快。工程实例表明,孤立森林算法为智能在线分析提供了条件,且特别适用于实时在线监测分析和健康状态安全预测预警。

关键词: 压力钢管;孤立森林算法;数据预处理;人工智能

中图分类号: TV732⁺.41

文献标识码: A

文章编号: 1672—1144(2022)04—0058—06

Real-time Online Artificial Intelligence Analysis of Penstock Based on Isolated Forest Algorithm

SHANG Xin¹, FANG Chaoqun², HU Musheng³

(1. Beijing Shisanling Power Station, State Grid Xinyuan Company, Beijing 102200, China;

2. National Center of Quality Inspection & Testing for Hydraulic Metal Structure Ministry of Water Resources, Zhengzhou, Henan 450044, China;

3. China South-to-North Water Diversion Corporation Limited, Beijing 100097, China)

Abstract: In order to realize the real-time online artificial intelligence status analysis of penstocks in water conservancy and hydropower projects, based on analyzing the structural characteristics of the real-time online monitoring data of penstocks, the artificial intelligence algorithms applicable to the health status monitoring are explored, and the artificial intelligence algorithms applicable to unsupervised machines are explored. In the discussion of learning algorithms, it is determined to use the isolated forest algorithm for real-time online artificial intelligence analysis. The calculation of the monitoring data of the penstock 1[#] branch and 2[#] branch pipe of the No. 1 diversion system of Beijing Shisanling Pumped Storage Power Station shows that the isolated forest algorithm has good abnormality detection without the need to preprocess the real-time sensor data. Engineering examples show that the isolated forest algorithm provides conditions for intelligent online analysis, and is especially suitable for real-time online monitoring and analysis and health state safety prediction and early warning.

Keywords: pressure steel pipes; isolation forest algorithm; data reprocessing; artificial intelligence

收稿日期:2022-04-02

修稿日期:2022-05-06

基金项目:十三陵电厂引水系统压力钢管明管段在线监测系统研究(525711190002)

作者简介:尚鑫(1989—),女,高级工程师,主要从事水利水电工程新技术应用等方面工作。E-mail:41380433@qq.com

通讯作者:方超群(1983—),男,硕士,高级工程师,主要从事水工金属结构安全评估及人工智能应用等方面工作。

E-mail:fangchaoqun@163.com

压力钢管是水利水电工程用来输送有压水流的钢制管路,在保障水电站正常安全生产运行或引水管路安全使用方面具有重要的作用。压力钢管在服役过程中,随着金属材料疲劳、锈蚀等情况的发生,性能会不断的劣化。如何以实时在线的方式评估压力钢管在常态运行中的稳定性和可靠性,及时发现和监测运行过程中的动态风险,降低微小故障随运行工况和水情条件变化不断汇聚成重大安全生产隐患的风险,是当前水利水电工程安全运行亟需解决的问题。

目前,针对压力钢管安全运行的关键性因素研究,国内外常规方法仍集中在传统的人工检查和定期安全检测的方式进行安全性评估和维护。其中,国内在压力钢管安全评估方面集中在理论计算和短暂无损试验后的研究分析,如杨绿峰等^[1]通过弹性模量缩减法从理论角度探讨压力钢管的安全评估方法;黄波等^[2]从观测数据提出D-S理论进行安全评估。在人工智能分析基础技术和应用方面,国内外的研究成果较多,研究重点在算法模型、数据结构以及应用环境验证等方面,但在基于实时数据驱动并采用人工智能分析的水利水电工程压力钢管安全评估或预警方面,国内还没有相关研究成果公布。

孤立森林算法在不平衡样本中的应用具有优势,在船舶、坝体、水质、风电、高铁等类似设备上进行了较多探索,如操江能等^[3]在船舶柴油机故障监测方面对孤立森林算法进行了试验验证;张海龙等^[4]通过对大坝监测数据的分解与重构,将孤立森林算法引入坝体安全监测的应用;谢炎昆^[5]通过孤立森林算法实现了用水的异常监测;汤婷等^[6]通过风力发电产生的自然异常点验证了孤立森林算法的有效性;王冲等^[7]将孤立森林算法应用至高铁制动系统的实时异常监测等。

本文以北京十三陵抽水蓄能电厂1号引水系统压力管道明管段为实例,通过压力管道明管段实时在线监测系统所采集的实时应力、振动数据为数据源,采用孤立森林的算法实现对压力钢管的人工智能分析,实现对压力钢管安全运行进行安全评估和预测预警的目的。

1 数据集的获取

北京十三陵抽水蓄能电站位于北京市昌平区蟒山,是我国北方地区建成的第一座大型抽水蓄能电站。水电站利用已经建成的十三陵水库为下库,左岸蟒山的上寺沟为上库,由引水系统和尾水系统连

接上库和下库。发电厂房位于蟒山岩体内,电站装机容量为 4×200 MW,采用“一管两机”布置方式,由两个相互独立的水道系统组成,包括上库进/出水口、引水事故闸门、引水隧洞、引水调压井、压力管道、尾水支管、尾水事故闸间室、尾水调压井、尾水隧洞、下库进/出水口等建筑物^[8]。

水电站引水压力钢管在运行过程中,设备在线监测数据对设备的异常检测和安全评估具有现实意义。通过对压力钢管进行实时在线监测,可获得压力钢管在运行过程中的振动、应力、温度、位移、地震等参数信息,这些信息经过数据保存,可形成完备的压力钢管安全运行历史数据。

对于一种算法的验证,往往基于一定的数据集,而数据集的制作则需要进行数据的二次采样、数据剔除和数据填补,并且需要对数据进行归一化处理,这些工作一般是由人工完成,但对于实时在线运行的设备来讲,用历史数据制作的数据集往往只能进行事后的分析或基于历史数据的预测,无法实现实时的预测预警。本文通过具体的工程实例,讨论以实时在线的方式进行机器学习应用的可行性。

1.1 对数据库的读取

采用北京十三陵抽水蓄能电站1号引水系统压力钢管的1[#]支管和2[#]支管在运行时所产生的数据,进行智能分析。在工程试验中,由于压力钢管健康状况实时在线监测的传感器较多,且每个传感器监测的数据量较大,因此,本文仅提取对压力钢管安全运行安全相关性较大且随动性较好的振动、应力两个参数进行数据读取。

实时在线监测系统为线上运行,产生的为连续数据且在不断的增加,所以在进行数据的运算分析时应进行数据提取操作。为保证特征数据不被漏掉,也最大可能的包含压力钢管在机组发电、抽水以及停机状态下的参数信息,根据压力钢管的运行特性和电站开、关机的频度信息,本次将以5 min作为采样周期对数据库信息进行提取和运算。

1.2 数据预处理

机器学习算法优势得益于数据集的标准化。而在线监测系统实时感知的数据往往存在着一些缺失值,如传感器故障、通讯延迟等现象,这将导致数据库数据的数据缺失,存在缺失数据的数据集往往与机器学习不兼容,甚至获得与期望值相反的结论,这就需要在机器学习算法的导入之前加入相关的前置运算,以实现数据集的标准化。

传感器感知的数据以数字的形式存在于数据库

中,代表着压力钢管此刻运行的物理特征。数据处理前,将不完整的数据整行丢弃,可能会丢失有价值的时间切片。更好的策略则是估算缺失值,也就是从数据的已知部分推断未知部分。估算缺失值的方法有很多,有单变量特征插补、多变量特征插补和最近邻插补等等,每个估算方法均有特定的意义,在选择估算缺失值的方法时,应根据具体的应用条件进行选择。

对实时在线监测系统来说,采用任何参与运算的估算均会影响到运行参数的时效性,所以在缺失值的插补时,可采用常数插补,即将缺失值用0来进行插补。

2 基于无监督学习的异常值检测

2.1 异常值检测

异常检测(anomaly detection)是对不匹配预期模式或数据集中其他项目的项目、事件或观测值的识别,异常也被称为离群值、噪声、偏差等。异常检测是人工智能算法中机器学习算法的一种,它通过大量的测试值以检测是否有异常值的存在,可应用于各个行业,如设备故障的检测以及大型设备设施的实时在线监测和发现。

异常检测的算法是从数据集中学习数据特征,并为其设定“参考基准值”,当新进入的数据超过预定义为正常范围的阈值时,则将新数据与“参考基准值”偏离的程度作为输出。监督学习、半监督学习、无监督学习的方法均可用来进行异常监测,但对于传感器获得的实时在线监测数据,无法采用监督学习或半监督学习,应采用无监督学习的方法来进行智能分析。

2.2 无监督学习的异常检测算法应用场景

无监督学习异常检测算法适用于异常数据过滤、未标记数据筛选以及对严重不平衡数据的计算等场景。在这些场景中,都有一个共同的特征,即异常的数据量都是很少的部分,诸如:支持向量机、逻辑回归等基于监督或半监督学习的分类算法都不适用。无监督学习的异常检测算法适用于有大量的正向样本和大量负向样本的数据集,有足够的样本让算法去学习其特征,且未来新出现的样本与训练样本分布一致。

2.3 异常检测算法分类

异常检测的目的是找到数据集中与大多数数据不同的数据,常用的算法可分为三类,第一类是基于统计学的方法来处理异常数据,这种方法一般会构

建一个概率分布模型,并计算对象符合该模型的概率,把具有低概率的对象视为异常点;第二类是基于聚类的方法来做异常点检测;第三类则是基于专门的异常检测算法来运算,其目的极具针对性,这类算法的代表是单类支持向量机(One Class SVM)和孤立森林(Isolation Forest)^[9]。

2.4 样本特征与算法选择

工程应用中,很多样本无法找到负样本,或即使有负样本,该样本个数也非常的少,会造成训练集的不平衡。如果人为制造一些负样本或通过数据筛查整理出一些负样本,也会因为负样本的污染而无法达到预期的目标。所以在只有正样本而没有负样本的场景中,使用单分类算法则会更加合适。本项目采用的压力钢管健康状态实时监测数据均为正样本,即该数据也为单样本数据集。

人工智能的大部分算法对数据集的要求相对较高,数据集的质量也是影响其算法准确率的保障,所以一般用于训练的数据集均需要有标签。但压力钢管健康状态实时监测数据集没有标签,也不可能对实时在线的数据进行一一标注。对于没有标签且为单样本的数据集,只能采用单样本分类算法。单样本分类算法只关注与样本的相似或者匹配程度,对偏离的未知数据不作评估。

3 孤立森林算法

孤立森林是一种非常高效的异常检测算法,与随机森林算法相似,但孤立森林每次选择划分属性或划分点(值)时都是随机的,而不是根据信息增益或基尼指数来选择^[10]。

孤立森林算法通过随机选择一个特征来“隔离”观察结果,然后在最大值和最小值之间随机选择一个分割值,所选特征的值,由于递归分区可以用树结构表示,所以分离样本所需的分裂次数等于路径从根节点到终止节点的长度,这个路径长度在这些随机树的森林上平均,用于度量常态和实现我们的决策功能。当随机树组成的森林共同产生更短的路径长度时,它将很可能是异常。

孤立森林算法适用于连续数据的异常检测,将异常定义为“容易被孤立的离群点”。可理解为分布稀疏且离密度高的群体较远的点,在数据空间里面,分布稀疏的区域表示数据发生在此区域的概率很低,因此可认为落在这些区域里的数据是异常的^[11]。这就实现了实时数据的异常筛查。

3.1 孤立森林算法原理

孤立森林不定义数学模型也不对数据进行标记,而是去查找哪些点容易被“孤立”,它使用了一套非常高效的策略,例如我们用一个随机超平面来切割数据空间,切一次可以生成两个子空间,之后我们再继续用一个随机超平面来切割每个子空间,循环下去,直到每个子空间里面只有一个数据点为止。直观上来讲,我们可以发现那些密度很高的簇是被切分很多次才会停止切割,但是那些密度很低的点很容易很早就停到一个子空间里了。

孤立森林算法得益于随机森林的思想,与随机森林由大量决策树组成一样,孤立森林也由大量的“树”组成,但和决策树不太一样,其构建过程也比决策树简单,是一个完全随机的过程^[12]。

3.2 孤立森林算法模型

如何分割数据空间是孤立森林算法的核心思想。由于切割是随机的,所以需要特定的方法(如蒙特卡洛方法)来得到一个收敛值,即反复从头开始切,然后平均每次切的结果。

孤立森林由 n 个孤立树组成,每棵孤立树是一个二叉树结构。该算法分两个阶段,第一个阶段需要训练出 n 棵孤立树,组成孤立森林,然后将每个样本点带入森林中的每棵孤立树,计算平均高度,之后再计算每个样本点的异常值分数。

孤立森林算法模型的第一阶段算法步骤如下:

(1) 从训练数据中随机选择 Ψ 个样本点作为样本子集,放入树的根节点。

(2) 随机指定一个维度,在当前节点数据中随机产生一个切割点 p 。

(3) 以切割点生成一个超平面,然后将当前节点数据空间划分为2个子空间:把指定维度里小于 p 的数据放在当前节点的左子节点,把大于等于 p 的数据放在当前节点的右子节点。

(4) 在子节点中递归步骤(2)和步骤(3),不断构造新的子节点,直到子节点中只有一个数据或子节点已到达限定高度。

(5) 循环步骤(1)至步骤(4),直至生成 n 个孤立树。

孤立森林算法模型的第二阶段算法步骤如下:

获得 n 个孤立树之后,训练结束,然后用生成的数据来评估。用每一个数据点 (x_i) 遍历每一棵孤立树,计算数据点在森林中的平均高度值 $h(x_i)$,对所有点的平均高度做归一化处理。异常值分数的计算如式(1)所示:

$$S(x, \Psi) = 2^{\frac{E(h(x))}{C(\Psi)}} \quad (1)$$

其中:

$$C = \begin{cases} 2H(\Psi - 1) - \frac{2(\Psi - 1)}{\Psi} & \Psi > 2 \\ 1 & \Psi = 2 \\ 0 & \text{otherwise} \end{cases} = \begin{cases} 2[\ln(\Psi - 1) + 0.5772156649] - \frac{2(\Psi - 1)}{\Psi} & \Psi > 2 \\ 1 & \Psi = 2 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

式中: x 为输入的数据; Ψ 为子采样大小; S 为异常值的分数; E 为样本从树的根节点到叶节点的过程中经历的边的个数; C 为修正值,表示样本构建一个二叉树的平均路径长度; 0.5772156649 为欧拉常数。

通过计算,异常值分数越小表示数据越为异常。

3.3 数据分析环境

采用 Python + Numpy + Sklearn + Matplotlib 的开源组合框架进行数据处理并绘图,首先通过 Python 连接数据库,并对数据库相对应的数据通道进行二次数据采样,并将采样点以多维矩阵的形式保存在 Numpy 数组中。当遇到数据缺失时,直接标记为 0。完成数据二次采样后,即形成用于计算的数据集,不再进行数据的预处理。

3.4 数据集运算

将北京十三陵抽水蓄能电站 1 号引水系统压力钢管的 1[#] 和 2[#] 支管的“振动-应力”两个不同量纲作为数据的两个维度进行运算,获得 1[#]、2[#] 支管在 X 轴和 Y 轴方向上的四组计算数值,并在二维平面内予以展示。

由于孤立森林算法每次选择划分属性和划分点时都是随机的,为了数据能够复现,指定随机数种子来生成伪随机数。从相同的随机数种子出发,可以得到相同的随机数序列^[13]。

计算采用 IsolationForest() 函数,该函数仅需要设置异常值比例和伪随机数,其它均采用默认参数,即:

contamination = 0.00005;

random = np.random.RandomState(42)。

运算后结果如图 1—图 4 所示。

3.5 异常值分析

本次运算采用配置为 Core i5-9400F 处理器的普通计算机,每个数据集均包含了 13 000 行数据,运算时间分别为 8.22 s、8.54 s、8.09 s 和 8.45 s,可见在运算速度方面,孤立森林有明显的速度优势。

从 3.4 小节中可视化结果看,黑色的点为正常点,红色的点为计算出的异常点。通过查询函数将异常点进行定位,可获得四组异常点的时间坐标,四个数据集异常值的时间坐标如表 1 所示。

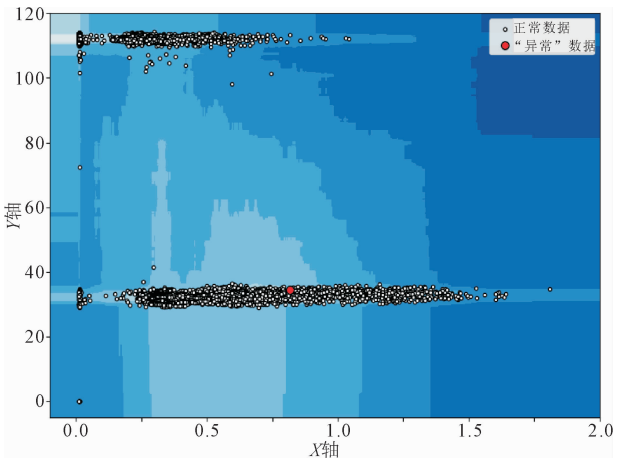


图 1 1#支管 X 轴方向“振动-应力”数据集异常检测

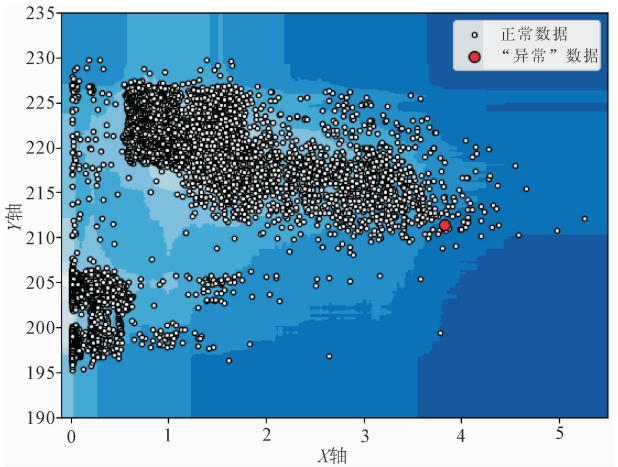


图 2 1#支管 Y 轴方向“振动-应力”数据集异常检测

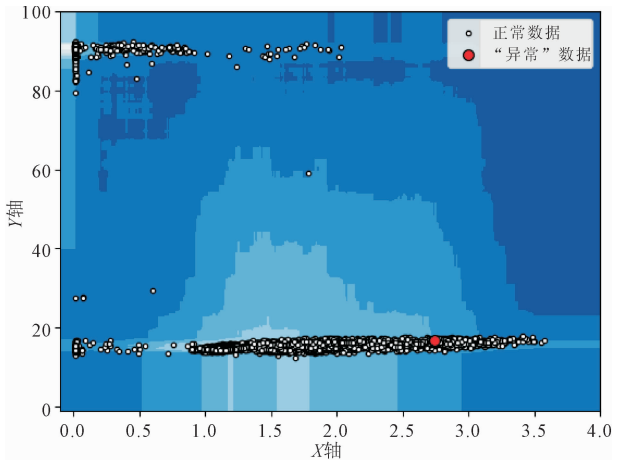


图 3 2#支管 X 轴方向“振动-应力”数据集异常检测

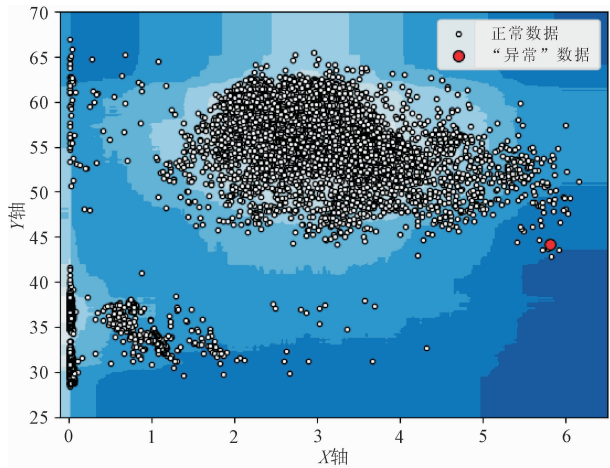


图 4 2#支管 Y 轴方向“振动-应力”数据集异常检测

表 1 四组数据集的异常值时间坐标					
位置	数据 ID	数值	时间	状态	特征
1 号 X 轴	216961	0.8163, 34.53	9 月 5 日 20:26:54	发电	停机前 26 min
1 号 Y 轴	329401	3.8236, 211.34	9 月 18 日 20:44:10	发电	停机前 34 min
2 号 X 轴	223771	2.7460, 16.62	9 月 6 日 20:30:36	发电	停机前 11 min
2 号 Y 轴	48091	5.8090, 44.12	8 月 16 日 21:15:59	发电	停机前 26 min

从表 1 可以看出,孤立森林算法计算后发现的异常点具有非常明显的特征,四组异常值所处的时间范围集中在晚上 20:30 时左右,且压力钢管均处于停机前 30 min 左右的发电状态。作为抽水蓄能电站,停机前的工况是机组运行中对压力钢管最不利的工况,存在着流态的急剧变化和水锤影响^[14-16],由此可以判断,该实时在线监测系统具有发现压力钢管异常状态的能力。同时需要说明的是,本次发现的“异常”点只是通过调整 contamination 参数的值从压力钢管正常运行数据中抽取的,该“异常”点仅说明了压力钢管此时的工况发生了微小变化,并不说明压力钢管出现了故障。

4 结 语

孤立森林算法作为人工智能技术中的数据分析方法,它的特点是不需要对数据集进行标注,能够以在线的方式实现对压力钢管监测数据的异常分析,这就为发现压力钢管运行时的异常表征和及时发现事故前的征兆信息提供了计算基础,从而对实现压力钢管的安全评估和预测预警提供技术手段。

根据水利水电工程压力钢管运行数据的结构特征及数据特点,选择孤立森林算法进行实时在线的分析具备工程实施的可行性。通过与采用 OCSVM 计算^[17]的结果相比较,采用孤立森林计算获取的异常值与采用 OCSVM 计算获取的异常值高度的一致,说明孤立森林算法对运行中的设备异常具有较为可靠的发现能力。

人工智能技术中的每一类算法均有各自的特征和应用范围,孤立森林算法具有异常发现能力,并且对于数据量较大的数据集具备快速计算能力,但孤立森林由于不定义数学模型,所以对发现的“异常”点也不会进行评价和判断,所以在使用该算法时,可作为实时数据的前置分析手段,对于如何将孤立森林发现的异常点进行评价,则需要用其它手段来共同完成。

参考文献:

- [1] 杨绿峰,张伟,韩晓凤.水电站压力钢管整体安全评估方法研究[J].水力发电学报,2011,30(5):149-156.
- [2] 黄波,阳萍华,肖剑,等.水电厂压力钢管运行安全监测及预警关键技术研究[J].湖南电力,2016,36(3):15-18.
- [3] 操江能,尚前明,杨安声,等.基于优化孤立森林的船舶柴油机故障监测[J].船舶工程,2021,43(11):125-132.
- [4] 张海龙,范振东,陈敏.孤立森林算法在大坝监测数据异常识别中的应用[J].人民黄河,2020,42(8):154-157.
- [5] 谢炎昆.基于孤立森林算法的用水异常监测研究[D].北京:北京邮电大学,2022.
- [6] 汤婷,张岩,安宗文.基于孤立森林算法的风电齿轮箱轴承故障检测方法[J].轴承,2022(2):68-74.
- [7] 王冲,刘杰.基于宽度学习的高铁制动系统实时异常检测[J].北京信息科技大学学报(自然科学版),2022,37(2):1-7.
- [8] 翟洁,张毅,李萌,等.十三陵抽水蓄能电站上水库面板堆石坝抗震复核分析[J].大坝与安全,2018(3):7-12.
- [9] 蔡兴旭,刘以安,肖颖.基于改进 LSTM 的桥梁传感器异常数据的检测方法[J].计算技术与自动化,2021(2):8-11.
- [10] 胡仁健.面向桥结构健康管理的大数据分析方法及实现[D].南京:东南大学,2020.
- [11] 朱佳俊.机器学习在智能入侵检测系统中的应用[D].上海:上海交通大学,2020.
- [12] 李新鹏,高欣,阎博,等.基于孤立森林算法的电力调度流数据异常检测方法[J].电网技术,2019(4):1447-1456.
- [13] 章立亮.IFS 分形吸引子的多尺度随机控制[J].纺织高校基础科学学报,2017(1):92-95.
- [14] 付亮,鲍海艳,田海平,等.基于实测甩负荷的水轮机力矩特性曲线拟合[J].农业工程学报,2018,34(19):66-73.
- [15] 张飞,朱晨,徐静,等.抽水蓄能电站压力钢管焊缝应力测试与评价分析[J].人民长江,2017,48(17):91-95.
- [16] 赵慧荣.上标水电站压力明钢管的安全运行与监控[J].大坝与安全,2009(1):39-44.
- [17] 尚鑫,方超群,武东辉.基于 OCSVM 的压力钢管健康监测数据的异常检测[J].水电与抽水蓄能,2022(1):106-112.