基于自编码器和隔离森林的水处理系统递进式异常检测方法

胡向东[1,2],刘 浪[2]
(1. 重庆邮电大学现代邮政学院,重庆 400065;2. 重庆邮电大学自动化学院/工业互联网学院,重庆 400065)

摘 要: 集成了工业互联网技术的水处理系统随着信息化程度的加深而面临着愈加严峻的异常行为入侵挑战. 针对传统异常检测方法常用单一阈值检测、检测准确率低、误报率高等问题,提出一种融合自编码器和隔离森林的水处理系统递进式异常检测方法. 首先,通过降采样过滤重复数据,加快递进式异常检测模型的训练和测试效率;其次,构建自编码器隐含层神经元捕捉数据关键特征,优化自编码器的权重和偏置,设定重构误差阈值作为输入与重构之间的差异度量进行基础性检测;最后,构建以平均路径长度为异常度量阈值的隔离树并生成隔离森林,针对基础性检测发现的异常数据进一步遍历隔离树完成高级检测;基于两阶段递进式异常检测提升检测效果. 实验结果表明,本文方法在安全水处理系统数据集下的异常检测准确率、F_1值均超过95%,准确率相比于传统方法平均提升31.86个百分点,特别是异常检测误报率被较大幅度降至0.30%. 对配水系统数据集进行泛化性分析取得的精确率、召回率等指标均超过94%. 模型的训练和测试时间相较于对比方法具有综合性能上的突出优势.

关键词: 水处理系统;异常检测;自编码器;隔离森林;递进式
基金项目: 重庆市高校创新研究群体(No.CXQT20016)
中图分类号: TN918.91;TP183 **文献标识码:** A **文章编号:** 0372-2112(2024)11-3823-12
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230704

A Progressive Abnormal Detection Method for Water Treatment System Based on Autoencoder and Isolation Forest

HU Xiang-dong[1,2], LIU Lang[2]
(1. *School of Modern Posts*, *Chongqing University of Posts and Telecommunications*, *Chongqing* 400065, *China*;
2. *School of Automation/School of Industrial Internet*, *Chongqing University of Posts and Telecommunications*, *Chongqing* 400065, *China*)

Abstract: With the deepening of informatization of water treatment systems integrated industrial internet technology are facing increasingly severe challenges of abnormal behavior intrusion. Aiming at such problems as single threshold detection, low detection accuracy, high false alarm rate and so on in traditional anomaly detection methods, a progressive anomaly detection method for water treatment systems that integrates autoencoders and isolation forests is proposed. Firstly, by downsampling to filter duplicate data, the training and testing efficiency of the progressive anomaly detection model is accelerated; Secondly, the hidden layer neurons of the autoencoder are constructed to capture the key features of the data, optimize the weight and bias of the autoencoder, and set the reconstruction error threshold as the difference measurement between input and reconstruction for basic detection; Finally, construct an isolation tree with the average path length as the anomaly measurement threshold to form an isolation forest, and further traverse the isolation tree to complete advanced detection based on the anomaly data discovered by basic detection; Improving detection performance based on two-stage progressive anomaly detection. The experimental results show that the accuracy and F_1 score of the proposed method in the secure water treatment dataset exceeds 95%, compared with the traditional method, the accuracy is improved by 31.86 percentage points on average, especially, the false positive rate of anomaly detection is significantly reduced to 0.30%. The precision rate, recall rate and other indicators obtained by the generalization analysis of the water distribution dataset are all over 94%. The training and testing time of the model has outstanding advantages in terms of comprehensive performance compared to comparative methods.

收稿日期:2023-07-26;修回日期:2024-02-02;责任编辑:梅志强

Key words: water treatment system; abnormal detection; autoencoder; isolation forest; progressive
Foundation Item(s): Chongqing University Innovation Research Group (No.CXQT20016)

1 引言

水处理系统作为关系国计民生的关键基础设施之一,其安全可靠运行具有重要的保障性意义. 随着工业化与信息化的深度融合及工业互联网等新一代信息技术的不断迭代及其在水处理系统中的应用渗透,水处理系统逐渐呈现出规模大型化、功能智能化和控制复杂化的特点[1]. 但工业控制系统中先进信息和通信工具的日益集成大大增加了水处理系统被入侵的潜在风险[2].

水处理系统由多个可编程逻辑控制器(Programmable Logic Controller,PLC)、监视和数据采集系统(Supervisory Control And Data Acquisition,SCADA)以及人机接口(Human Machine Interface,HMI)组成. PLC通过通信网络与其他PLC、SCADA和HMI进行通信. 水处理系统与其他信息系统不同的是,攻击者通常不会直接攻击PLC、SCADA,因为他们的目的不是使PLC或者SCADA瘫痪,而是通过欺骗PLC、SCADA去执行恶意程序代码或指令进行敲诈勒索等. 攻击者通过篡改、操纵PLC、SCADA以及HMI的控制器和传感器读数以及代码对水处理系统发起网络攻击,进一步形成难以检测的流量异常、压力异常等现象[3]. 系统中的某些攻击手段不会立即生效,但对传感器、执行器的攻击会在一定时间延迟后影响整个系统的性能. 由于水处理系统的不同功能节点资源差异、指令运行实时性要求高等特殊性,传统异常检测方法很难直接运用于水处理系统,多技术方法融合是适应该类异常检测未来发展的新趋势之一.

自编码器因其可学习数据中的有用特征而受到广泛关注. 文献[4]采用堆叠自编码神经网络对数据进行降维,并根据数据间的相互关联性设计长短期记忆神经网络异常检测模型. 该模型提高了异常检测准确率并与对比模型相比降低了误报率. 但其模型结构相对复杂,所占资源较多. 文献[5]利用无向图结构提取数据间的关联特征,由双路自编码器对原始特征和关联特征进行融合,然后基于高斯混合模型估计数据的概率密度,最后设定阈值进行异常检测. 该模型的异常检测各项指标均提升2%左右,但模型步骤过于复杂,很难迁移到工业互联网应用场景. 文献[6]针对攻击检测效果不佳问题,首先利用注意力机制采集特征信息,再通过变分自编码器设定阈值判断异常情况. 该方法提升了检测精度,提高了低频次攻击的检测率. 文献[4~6]均使用单一阈值检测,其准确率还有待提高.

由于隔离森林适用于高维数据的异常检测,其成为异常检测的研究热点之一. 文献[7]由一维卷积神经网络实现的特征提取模型和用于检测的隔离森林模型组成. 该框架能够检测更广泛的攻击场景,但单一的隔离森林阈值检测导致误报率较高,准确率较低. 文献[8]基于双隔离森林模型进行异常检测,该模型分别将原始数据进行规范化和主成分分析,再将处理好的两种数据进行独立训练,其在攻击检测能力、计算要求等方面有所提高,但召回率与F_1值表现不佳. 文献[9]针对冶金能源数据污染问题,利用隔离森林选取正常数据输入自编码器进行训练,重构误差大于阈值的数据被判定为异常. 该方法所涉及的隔离森林仅用于选取正常数据,并不参与后续数据的训练和检测,虽提高了F_1值,但单一自编码器提高了攻击检测对于阈值的敏感性,导致其他指标表现不佳.

另外,基于深度学习的异常检测方法也在不断发展. 文献[10]针对海量不平衡数据问题,使用残差块构建特征模块以此得到高质量的数据特征,并通过动态路由算法对数据特征进行聚类,该方法有效提高了检测准确率. 文献[11]利用改进的灰狼优化算法去优化支持向量回归模型的参数,并通过最优参数构建支持向量回归预测模型,该模型有更高的预测准确率和预测精度.

由以上分析可知,针对水处理系统的攻击往往呈现的是过程异常,即潜伏在系统中随着系统继续运行,在一定时间延迟后对系统产生破坏,被发现时往往已经造成了损害. 异常检测技术会对水处理系统中各个传感器、执行器的读数进行实时检测,不符合系统预期的行为就会被检测出来. 但目前的异常检测方法常用单一阈值进行检测,导致攻击检测对阈值的敏感性过高,准确率和误报率难以满足需求.

综上,本文设计出一种融合自编码器和隔离森林(Auto Encoder-Isolation Forest,AE-IF)的水处理系统递进式异常检测方法. 结合自编码器的重构误差和隔离森林的异常度量进行综合评估,解决准确率低及误报率高的问题;利用自编码器学习数据的内在表示来捕捉潜在的异常模式,计算重构误差进行基础性检测;对于基础性检测中被归类为异常的数据利用隔离森林进行更精细的异常度量实现递进式高级检测,突破传统方法单一阈值检测不够精准的局限性. 综合使用两阶段递进式检测提高整体的异常检测性能,该方法相较于对比方法有着更高的准确率和较低的误报率.

2 水处理系统递进式异常检测方法

2.1 水处理系统递进式异常检测模型

基于工业互联网的水处理系统因大量传感器节点

的广泛分布和信息化程度的加深,系统数据通常具有量大、种类多、维数高、重复率高等特点,尤其是正常数据量远比异常数据量多,二者极不平衡,且水处理系统的检测和控制还表现出实时性要求较高等需求. 因此,水处理系统的异常检测面临着数据关键特征捕捉和异常度量困难等问题.

本文针对基于工业互联网的水处理系统的数据特点和高实时性需求,提出融合自编码器和隔离森林算法的递进式异常检测方法,充分利用两种算法各自的优势和适应场景. 自编码器作为无监督学习的神经网络,其最大的优势是结构简单、训练快,能够从海量的数据中自主学习,及时获取数据的有效特征;隔离森林的突出优势[12]在于避免了计算距离、密度,其时间开销不随数据维度的增加而增加,可以有效解决高维数据的不利影响,且隔离森林算法属于集成学习方法,对大型数据集的适应性较好,特别是对于不平衡数据集,其有助于减少对异常数据的掩盖.

基于自编码器和隔离森林的水处理系统异常检测框架如图1所示,主要包含三个模块:数据预处理、模型训练和异常检测.

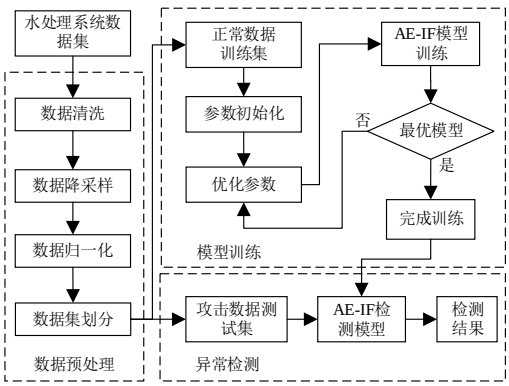


图1　水处理系统异常检测框架

在图1中,异常检测阶段将攻击数据测试集输入经训练优化得出的AE-IF检测模型. 首先通过检测模型的自编码器进行基础性检测,对于该阶段输出的被归类为异常的数据,再调用检测模型的隔离森林进行高级检测;基于两阶段递进式检测来改善检测准确率、降低误报率.

2.2　数据预处理

来自真实应用场景的水处理系统的原始网络流量数据、传感器读数和执行器数据等往往存在无效数据、缺失值,样本数据量过大、数据重复率过高、不同类别数据差异大等特点. 为了提高本文构建的水处理系统递进式异常检测模型的训练效率、检测准确率和及时性等,有必要对其进行预处理,预处理环节包括数据清洗、数据降采样、数据归一化和数据集划分.

数据清洗主要是对水处理系统数据集存在的缺失值和无效数据进行删除操作. 数据缺失值是指在数据集中某些观测变量的数值没有记录. 分析数据时发现数据集中缺失值的数量相对较少,对后续实验分析影响较小,本文选择删除包含缺失值的观测变量记录,以确保数据的完整性和一致性,改善原始数据的质量.

数据降采样主要针对水处理系统中可能存在传感器和执行器的读数较长时间没有变化,从而导致数据集样本量过大、数据重复率过高,影响训练效率的问题,对数据集数据引入降采样方法进行原始数据的遴选,即在间隔1 s的原始采样数据中选取每间隔5 s的数据备用,以此减少数据处理量,降低因采样数据高度重复带来处理效率等方面的不利影响.

数据归一化主要针对水处理系统中不同类别数据的取值范围差别大、度量单位不统一问题,采用Min-Max标准化方法对数据进行归一化,确保不同类别和量级的数据可以进行比较和分析. Min-Max归一化方法如式(1)所示:

$$x^* = \frac{x - \min}{\max - \min} \tag{1}$$

其中,x^*是归一化后的数据;x是原始数据;min代表原始数据的最小值;max代表原始数据的最大值.

数据集划分的目的在于将经过数据清洗、数据降采样和数据归一化后形成的更高质量水处理系统数据集根据模型训练和测试需要分解为正常数据训练集和攻击数据测试集两部分,分别用于模型训练和异常检测.

2.3　自编码器检测方法

2.3.1　自编码器算法原理

自编码器包括编码器和解码器[13],其中输入层和隐藏层构成编码器,隐藏层与输出层构成解码器. 自编码器的网络结构如图2所示.

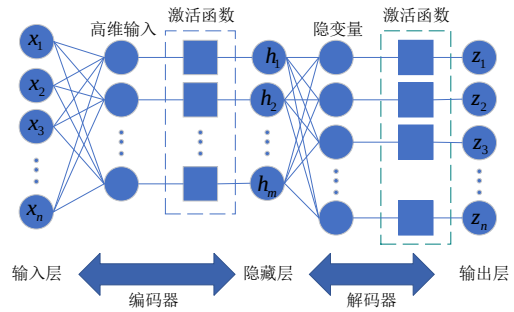


图2　自编码器网络结构

自编码器的处理过程为:编码器通过激活函数把高维输入数据编码成低维的隐变量;解码器把隐变量还原为初始状态. 自编码器最好的状态就是解码器的输出能够最大限度地还原出原始输入.

2.3.2 自编码器算法实现

自编码器算法实现主要包括2个过程：编码与解码、实例检测.

编码与解码 本文中的自编码器使用Relu作为激活函数. 编码器如式(2)所示：

$$\boldsymbol{h}=\sigma(W_1\boldsymbol{x}+b_1) \tag{2}$$

其中，W_1和b_1是编码器的权重和偏置；$\sigma(\cdot)$是非线性变换函数；$\boldsymbol{h}$是隐藏变量；$\boldsymbol{x}$是输入矢量. 式(2)表示编码器通过非线性变换将输入矢量$\boldsymbol{x}$映射到隐藏变量$\boldsymbol{h}$中.

解码器如式(3)所示：

$$\boldsymbol{z}=\sigma(W_2\boldsymbol{h}+b_2) \tag{3}$$

其中，W_2和b_2是解码器的权重和偏置；$\boldsymbol{z}$是重构矢量. 式(3)表示解码器通过与编码器相同的非线性变换将隐藏变量$\boldsymbol{h}$映射回原始输入空间进行重构[14]，以便计算重构数据与原始输入数据的误差.

重构误差$\boldsymbol{e}$的计算如式(4)所示：

$$\boldsymbol{e}=\boldsymbol{x}-\boldsymbol{z} \tag{4}$$

即重构误差等于原始输入矢量$\boldsymbol{x}$与重构矢量$\boldsymbol{z}$之差.

实例检测 自编码器算法根据训练数据计算重构误差，并以此作为基础性异常检测的阈值. 在编码与解码过程结束后，计算实例的重构误差并判断是否大于阈值，若重构误差大于阈值则为异常，否则为正常. 自编码器检测流程如图3所示.

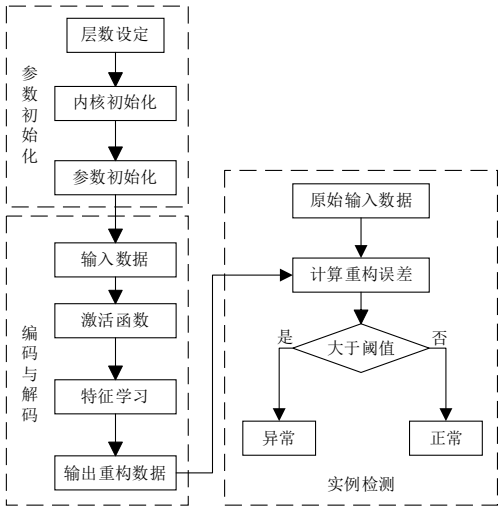


图3 自编码器检测流程

2.4 隔离森林检测方法

2.4.1 隔离森林算法原理

隔离森林算法首先使用随机超平面对数据空间进行切割，每次切割都会生成两个子空间. 使用随机超平面重复切割子空间，直到每个子空间中只剩下一个数据点. 由于高密度数据集需要多次切割才能停止，所以在切割后低密度点很早就会被分割出来，从而停留在子空间中. 隔离森林算法通过创建隔离树(isolation Tree，iTree)来隔离样本，每棵隔离树对随机数据子集执行递归二进制分裂，直到所有样本都被隔离. 其分裂隔离过程如图4所示.

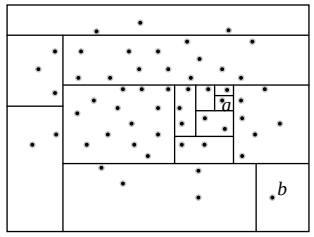


图4 隔离森林算法的分裂隔离过程

图4是二维数据空间中一棵隔离树的生成过程. 在该空间中不断选取随机超平面来分割空间并隔离数据，高密度点需要多次隔离才会被划分出去，导致路径较长，如点a. 低密度点只需较少隔离次数便会被划分出去，所以路径较短，如点b.

2.4.2 隔离森林算法实现

隔离森林算法通过隔离树的子采样进行建立，隔离森林算法流程如图5所示.

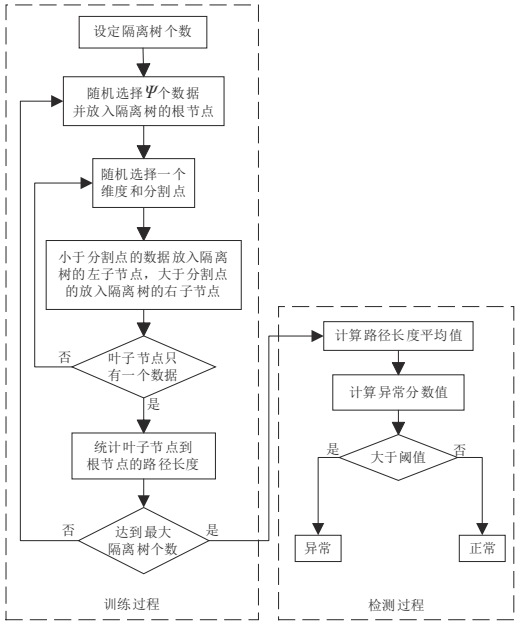


图5 隔离森林算法流程

由图5可知，隔离森林算法实现包括两个过程：训练过程和检测过程. 训练过程使用从训练数据集中随机选取的实例来构建隔离树并生成隔离森林. 检测过程则将自编码器基础性检测方法判断为异常的数据通过隔离树的传递进行高级检测.

训练过程 在训练过程中，隔离树通过递归划分给定的数据集，直到每个实例都被隔离或达到隔离树

高度限制. 训练过程分为4个步骤:

步骤1　从训练数据中随机选取ψ个实例作为子样本并放入隔离树的根节点.

步骤2　随机选择一个维度,在当前数据范围内随机生成一个分割点p,该分割点产生于当前数据中的最大值和最小值之间.

步骤3　在隔离树中把小于p的点放在当前分支的左边,大于p的点放在右边.

步骤4　当前节点的左右分支重复步骤2和3,直到叶子节点只有一个数据或隔离树达到高度限制为止. 隔离树训练时的迭代过程如算法1所示.

算法1　隔离树训练算法

输入: 输入数据X,当前隔离树高度m,限制高度l

输出: 隔离树

1: if $m \geqslant l$ or $|X| \leqslant 1$ then
2: return 外节点{Size ← $|X|$}
3: else
4:　设Q是X中的一个属性列表
5:　随机选择一个属性$q \in Q$
6:　从X中属性q的最大值和最小值中随机选择一个分割点p
7:　X_l ← 分离 $(X, q < p)$
8:　X_r ← 分离 $(X, q \geqslant p)$
9: return 内节点{左节点 ← 隔离树(X_l, m+1, l),
10:　右节点 ← 隔离树(X_r, m+1, l),
11:　分割属性←q,
12:　分割值←p}
13: end for

检测过程　在检测过程需要通过检测点的路径长度推导出异常分数. 数据的平均路径长度是一个重要指标,其计算式为

$$c(n)=2H(n-1)-\frac{2(n-1)}{n} \tag{5}$$

其中,n为给定的数据集实例个数;$c(n)$为n的平均路径长度;$H(\cdot)$为谐波数,可由欧拉常数0.577 215 664 9+ln$(\cdot)$估计.

隔离森林算法根据得到的平均路径长度计算异常分数,实例x的异常分数计算式为

$$s(x,n)=2^{-\frac{E(h(x))}{c(n)}} \tag{6}$$

其中,$h(x)$为实例x的路径长度;$E(h(x))$是实例x在一组隔离树中路径长度的平均值.

$E(h(x))$是通过隔离森林中每棵隔离树递归传递实例得到. 通过遍历隔离树计算实例x从根节点到终止叶子节点的边数s,以此推导出实例x在单棵隔离树上的路径长度$h(x)$. 当对集合中的每棵隔离树计算得到$h(x)$时,就可通过式(6)计算x的异常分数. 路径长度计算过程如算法2所示.

算法2　路径长度算法

输入: 实例x,隔离树T,当前路径长度s

输出: x的路径长度

1: if T是一个外节点 then
2: return s+c(T.size)
3: end if
4: a←T.分割属性
5: if x_a < T.分割值 then
6: return 路径长度(x, T.left, s+1)
7: else {$x \geqslant T$.分割值}
8: return 路径长度(x, T.right, s+1)
9: end for

2.5　两阶段递进式异常检测

由2.1节的水处理系统递进式异常检测模型可知,自编码器和隔离森林算法基于各自的优势和适用场景,能够较好地满足水处理系统异常检测的需要. AE-IF两阶段异常检测流程如图6所示.

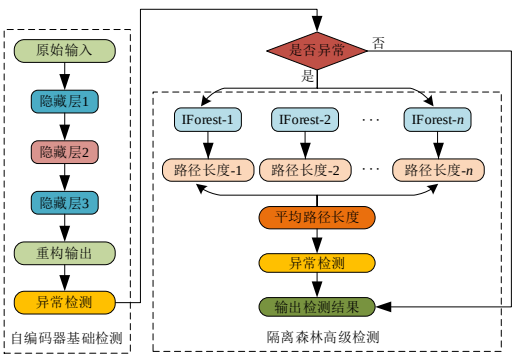


图6　AE-IF两阶段异常检测流程

图6中输入层、隐藏层以及输出层均为全连接层,分别为1、3、1层. 基于自编码器的基础检测通过隐藏层1、隐藏层2以及隐藏层3充分学习数据的特征表示,将测试数据经过编码器和解码器转换后,最大限度地重构原始输入数据,完成异常数据的检测识别.

基于隔离森林算法的高级检测将自编码器识别为异常的数据传递到预先设置好的n棵隔离树中,通过每棵隔离树的分割选择后输出每棵隔离树的路径长度,对得到的n个路径长度取平均值,得出它们的平均路径长度,基于该平均路径长度对数据进行异常判定,输出检测结果.

3　实验分析

3.1　水处理试验台及其分析

为了综合可靠地验证本文方法的检测效果,采用

两个完全不同的水处理试验台产生的基准数据集进行独立实验验证，分别是由新加坡科技与设计大学iTrust网络安全研究中心开发的安全水处理（Secure Water Treatment，SWaT）试验台[15]和Ahmed、Palleti、Mathur构建的水分配（Water Distribution，WADI）试验台[16].

3.1.1 SWaT水处理试验台

SWaT水处理试验台隶属于一个全面运行的小型水处理厂，该试验台由阶段1~6共6个工艺过程组成，6个阶段共同处理水资源（如图7所示），以此模拟城市大型水厂处理过程.

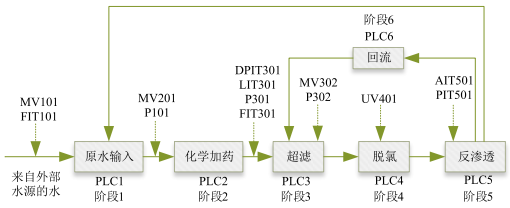


图7 SWaT试验台水处理过程

图7中MV101、P101、FIT101等为涉及的传感器或执行器型号，其中Pxxx代表泵，AITxxx代表化学传感器，DPITxxx代表压差指示器，MVxxx代表电动阀，FITxxx代表流量指示器，LITxxx代表液位传感器，UVxxx代表紫外线灯，PITxxx代表PH值传感器.

阶段1中，来自外部的水源进入原水箱，以足够的水资源用于后续处理过程；阶段2加入化学物质进行预处理以保持水质安全；阶段3使用超滤装置过滤掉水中残留的物质；阶段4使用紫外线灯进行脱氯，清除残留的氯气；阶段5通过反渗透去除水中存在的无机杂质；阶段6将水重新回流到阶段3进行循环过滤和脱氯. 完成处理后的水被回收返回到第1阶段.

SWaT水处理试验台的每一个阶段由独立的可编程逻辑控制器（PLC）控制. 每个PLC分别连接到各自阶段的一组传感器或执行器，传感器负责采集平台内部组件的状态变化信息，PLC基于传感器获取的信息向执行器发送控制命令. SWaT水处理试验台传感器与执行器型号如表1所示.

表1 SWaT水处理试验台各个阶段传感器与执行器型号

阶段	传感器型号	执行器型号
阶段1	FIT101 LIT101	MV101 P101 P102
阶段2	AIT201 AIT202 AIT203 FIT201	MV201 P201 P202 P203 P204 P205 P206
阶段3	DPIT301 FIT301 LIT301	MV301 MV302 MV303 MV304 P301 P302
阶段4	AIT401 AIT402 FIT401 LIT401	P401 P402 P403 P404 UV401
阶段5	AIT501 AIT502 AIT503 AIT504 FIT501 FIT502 FIT503 FIT504 PIT501 PIT502 PIT503	P501 P502
阶段6	FIT601	P601 P602 P603

3.1.2 WADI配水试验台

WADI配水试验台代表城市中大型配水系统的缩小版本. 整体配水过程如图8所示，该配水试验台包含三个不同的控制进程，即图中阶段1到阶段3，每个进程由各自独立的一组PLC控制. 阶段1的主网格包含两个储水罐，分别为T-001和T-002，其水源来自外部水和阶段3的循环水. 阶段2时主网格中的水依序向高架水箱和消费水箱供水，这是根据需求向用户水箱分配水的过程. 阶段3中，一旦用户水箱满足用户需求，循环水就会被传输回主网格.

图8中，S和A分别代表传感器和执行器的集合；1-LT-001代表阶段1的储水罐1中的液位传感器；1-FS-001代表阶段1中的流量计1；1-T-001代表阶段1中的1号储水罐；2-MV-001代表阶段1中的电动阀2；2-MCV-101代表阶段1中的电动消费阀2；3-P-004代表阶段3中的水泵4.

3.1.3 水处理试验台攻击数据分析

由SWaT水处理试验台与WADI配水试验台的构成可知，来自外部的拒绝服务、身份欺骗或者数据伪造、篡改等攻击导致系统异常的可能入口在于PLC、传感器或执行器，被攻击的传感器发送错误的数据给PLC，或者直接攻击PLC，导致PLC依据错误的数据做出错误的决策，或者攻击执行器，向其发送错误的指令导致非预期的水处理操作或供水中断等破坏性效果.

图9是SWaT水处理试验台中执行器MV101受到攻击时传感器LIT101测量值的变化情况，以此为例进行说明. 图中红色线段标记为攻击. 受到攻击前，执行器MV101处于关闭状态，攻击者通过欺骗PLC打开执行器MV101，使得本身处于高水位的储水罐仍然进水，导致水溢出储水罐. 传感器LIT101的对应测量值表明储水罐的液位在不断升高，揭示出异常. 图9所示攻击是通过欺骗PLC实现的，很难从执行器MV101入手检测发现该攻击，但通过检测传感器LIT101测量值的变化有助于发现. 要对SWaT水处理试验台与WADI配水试验台进行有效的攻击检测，就需要充分利用试验台中各个传感器和执行器信息. 此外，不同攻击的持续时间不同，故水处理试验台的时间序列特征不可忽视.

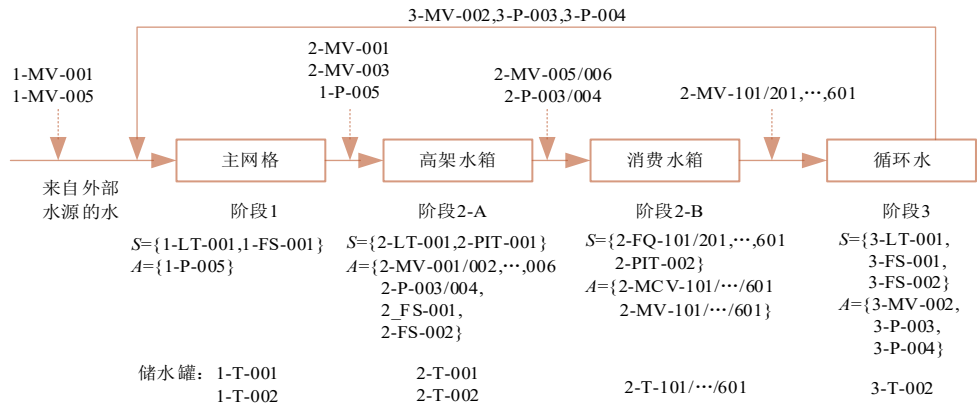
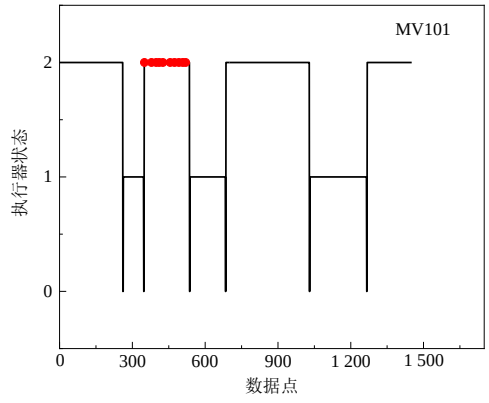
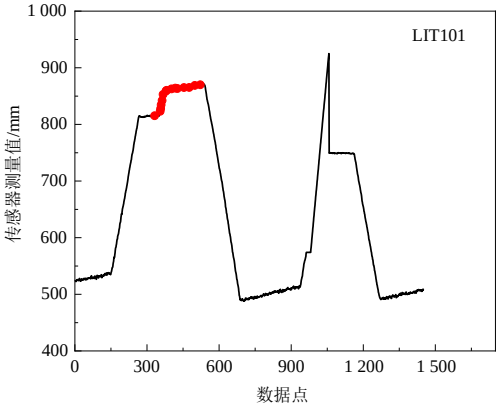


图 8　WADI 试验台配水过程

基于上述分析,本文利用自编码器包含的编码与解码过程学习数据的时间序列特征,且隔离森林算法可用于包含连续变量(传感器测量值)和离散变量(执行器信号)的混合数据集,利用隔离森林算法的隔离性质,可用于分析不同传感器测量值和执行器信号之间的关联关系,实现对系统异常的判定.



(*a*) MV101 受到攻击时的变化情况



(*b*) LIT101 传感器测量值变化情况

图 9　MV101 受到攻击时 LIT101 的变化情况

3.2　实验环境与评估指标

3.2.1　实验环境

本文的算法测试实验均在操作系统 Windows10、处理器 Intel(R) Core(TM) i7-10700 CPU @ 2.90 GHz、RAM 为 16.0 GB 的环境下进行,模型的算法实现调用了 Python 中 Scikit-learn、Keras 库等方法.

本文实验所用的 SWaT 和 WADI 数据集都是通过搭建真实的水处理系统采集而得,能够体现实际水处理系统的数据特征及其关联关系和复杂性,包含的攻击数据多为过程异常. SWaT 数据集源于 SWaT 试验台连续运行 11 天,其中 7 天在正常环境下运行,4 天在攻击场景下运行;在收集数据过程中,从 25 个传感器和 26 个执行器中收集网络流量、传感器读数和执行器数据. WADI 数据集源于 WADI 试验台连续运行 16 天,其中 14 天在正常环境下运行,2 天在攻击场景下运行;数据来自于 123 个传感器或执行器.

为了确保模型训练与测试效果,本文数据集划分将 SWaT 数据集中在 7 天正常环境下运行所采集的数据用作正常数据训练集,在 4 天攻击场景下运行所采集的数据用作攻击数据测试集. 将 WADI 数据集在 14 天正常环境下运行所采集的数据用作正常数据训练集,在 2 天攻击场景下运行所采集的数据用作攻击数据测试集.

SWaT 数据集包含的正常数据集样本和攻击数据集样本数量分别为 496 800 条和 449 919 条,基于 2.2 节所述降采样方法对数据进行降采样,降采样后的 SWaT 数据集中正常数据集样本和攻击数据集样本数量分别为 99 360 条和 89 984 条;WADI 数据集包含的正常数据集样本和攻击数据集样本数量分别为 784 571 条和 172 801 条,由于二者规模悬殊,攻击数据集样本数量相对较少,故对 WADI 数据集中的攻击数据集样本不做降采样处理,保留其原始的数据量,只对 WADI 数据集中的正常数据集样

本进行降采样,处理后其样本数量降为241 921条.故本文使用的样本类别分布如表2所示,其中,WADI数据集用于泛化性测试.

表2 SWaT和WADI数据集样本类别分布

数据集	训练集样本数	测试集样本数	维度
SWaT	99 360	89 984	51
WADI	241 921	172 801	123

3.2.2 评估指标

为了全面地评估模型的性能,选取准确率(Accuracy)、精确率(Precision)、召回率(Recall)、F_1值、误报率(FPR)以及漏报率(FNR)作为评价指标,它们的定义如下:

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{7}$$

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{8}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{9}$$

$$F_1=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{10}$$

$$\text{FPR}=\frac{\text{FP}}{\text{FP}+\text{TN}} \tag{11}$$

$$\text{FNR}=\frac{\text{FN}}{\text{FN}+\text{TP}} \tag{12}$$

其中,TP表示正常样本被正确分类为正常样本的数量;FP表示正常样本被错误分类为异常样本的数量;TN表示异常样本被正确分类为异常样本的数量;FN表示异常样本被错误分类为正常样本的数量.

3.3 模型训练

3.3.1 自编码器训练

本文的自编码器训练使用Keras库进行,Keras库是一个用Python编写的开源神经网络库.具体的实验参数如表3所示.

表3 自编码器实验参数

实验参数名称	解释说明	参数设置
activation	激活函数	Relu
batch_size	每个训练批次中的样本数量	64
epochs	迭代次数	50
learning_rate	学习率	0.001
loss	损失函数	MSE
optimizer	优化器	Adam

隐藏层是自编码器的核心部分,负责将输入数据转换为紧凑的表示形式,并在解码器中将其还原回原始数据.本文自编码器的隐藏层共有3层,隐藏层中的神经元通过训练过程学习如何提取和表示输入数据的有用信息,因此,每层隐藏层中神经元个数尤为重要.基于表3的参数设置进行隐藏层神经元个数寻优实验.

预实验中发现神经元个数在8~12之间时可保持较好收敛状态,故神经元个数寻优试验以此为基础,隐藏层各层神经元个数的训练损失值随迭代次数变化曲线如图10所示.

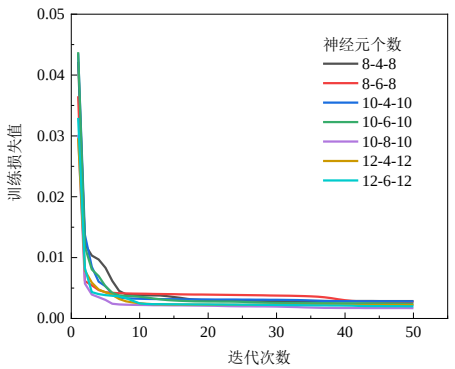


图10 隐藏层神经元个数训练损失值变化曲线

由图10可知,不同神经元个数组合实验中,3个隐藏层的神经元个数分别在10、8、10时,其训练损失曲线收敛最快且损失值最小,且在后续迭代中保持平稳状态.因此本文自编码器隐藏层的神经元个数分别设置为10、8、10.

由2.3节可知,自编码器通过计算实例的重构误差来判断该实例是否为异常,若重构误差大于阈值则为异常,否则为正常,所以需要利用训练数据寻找自编码器阈值.正常数据训练集的自编码器训练误差损失值分布如图11所示.横坐标代表自编码器误差损失值分布区间,纵坐标代表落在不同区间内的误差损失值的实例个数.

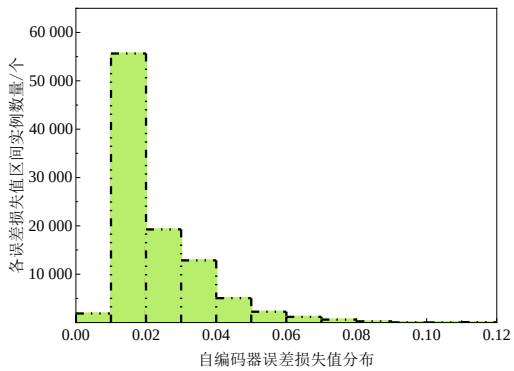


图11 自编码器训练数据误差损失值分布

由图11可知,自编码器训练数据误差损失值分布于0~0.12范围内.通过选取不同误差损失值进行实验发现阈值为0.08时效果最佳,故本文以此作为自编码器阈值用于基础性检测.训练好的自编码器对异常样本可有效识别,但存在将部分正常样本错误判断为异常样本的情形,为消除误判,本文针对基础性检测判断为异常的样本采用隔离森林检测方法做递进式再次鉴别.

3.3.2 隔离森林训练

隔离森林训练使用Scikit-learn库进行,Scikit- learn库是Python的开放软件机器学习库. 具体的实验参数如表4所示.

表4 隔离森林实验参数

实验参数名称	解释说明	参数设置
n_estimators	隔离森林中隔离树的数量	100
max_samples	每棵隔离树使用的样本数	auto
contamination	预期的异常样本比例	0.1
max_features	每棵隔离树使用的特征数	1.0
bootstrap	是否在构建隔离树时使用自助采样	False
random_state	控制随机数生成	42

由2.4节可知,隔离森林需要通过检测点的路径长度推导出异常分数,以此作为阈值用于异常判断. 正常数据训练集通过隔离森林训练后,其异常分数值分布如图12所示. 横坐标代表隔离森林异常分数值分布区间,纵坐标代表落在不同区间异常分数值的实例个数.

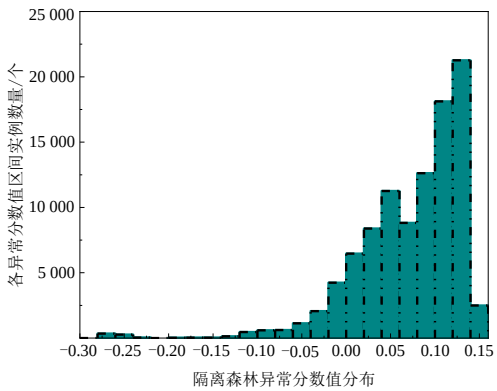


图12 隔离森林训练数据异常分数值分布

由图12可知,训练数据异常分数值绝大部分在大于−0.15的范围内,极少数小于−0.20. 通过选取不同异常分数值进行实验发现阈值为−0.15时可取得较理想的综合效果,所以本文选取−0.15作为隔离森林的阈值用于异常检测.

3.4 实验结果与分析

3.4.1 单一网络检测模型对比

为了验证基于两阶段递进式异常检测方法的优越性,首先将本文方法与单一网络检测模型进行实验对比. 用于对比的模型包括:不考虑自编码器可学习特征信息的基础性检测,仅使用单一隔离森林(Single Isolation Forest,Single-IF);不考虑隔离森林可学习过程信息的递进式检测,仅使用单一自编码器(Single Auto Encoder,Single-AE);单一阈值检测方法单类支持向量机(One Class Support Vector Machines,OCSVM)和局部异常因子(Local Outlier Factor,LOF)基线模型. LOF和OCSVM基线模型的实验参数如表5所示,Single-IF和Single-AE的参数与AE-IF的参数设置相同. 单一网络检测模型对比结果如表6所示.

表5 基线模型实验参数

模型	实验参数名称	解释说明	参数设置
LOF	n_neighbors	最近邻样本数量	3
	p	距离度量的幂参数	2
	algorithm	计算最近邻的算法	auto
	leaf_size	叶子节点的大小	30
	contamination	预期的异常点比例	auto
OCSVM	kernel	核函数类型	rbf
	degree	多项式核函数度数	3
	nu	异常点的比例	0.05
	cache_size	缓存的数据大小	200
	tol	控制算法的收敛性	0.001

由表6可知,与其他四个单一网络检测模型对比,本文提出的AE-IF除漏报率高于Single-AE外,其余指标均表现更佳,准确率、误报率和F_1值相较平均值分别改善了31.86、27.15和27.54个百分点. Single-AE除漏报率优于本文方法外,其余3个指标均有较大幅度落后,特别是Single-AE的误报率在所有对比模型中最高,达到81.74%,难以满足实际应用需求;LOF和OCSVM原理简单且易于实现,但误报率和漏报率相对较高,说明在训练过程中针对数据样本的特征信息丢失较多;同样采用了隔离森林算法的Single-IF在准确率、漏报率和F_1值3项指标上与本文提出的AE-IF接近,但仍然分别落后0.11、0.01和0.07个百分点,特别是其误报率是本文方法的6.1倍. 这说明本文的AE-IF使用自编码器进行基础性检测后再用隔离森林进行递进式高级检测可有效提升检测效果,在利用自编码器充分学习数据特征信息的同时,利用隔离森林优化异常分数值选择进一步强化精确检测,提高了样本的分类质量,误报率明显降低,这是本文的核心贡献.

表6 单一网络检测模型对比　　单位:%

模型	准确率	漏报率	误报率	F_1
LOF	33.11	75.60	2.96	39.11
OCSVM	78.72	21.01	23.27	86.73
Single-IF	95.02	5.23	1.83	97.24
Single-AE	46.25	0	81.74	56.00
AE-IF	95.13	5.22	0.30	97.31

3.4.2 融合网络检测模型对比

为了验证本文方法融合的有效性,选取相关融合网络检测模型进行精确率、召回率及F_1值的异常检测结果对比. 用于对比的融合模型包括:以隔离森林为基础的

双隔离森林检测模型DIF[8]，基于自编码器并引入自注意力机制的STAE-AD[17]模型，EPCA-HG-CNN[18]模型以及SDA-1D_CNN-GRU[19]模型. 对比结果如图13所示.

由图13可知，与融合网络检测模型相比，AE-IF的精确率最高达到99.98%，相较于DIF和STAE-AD分别提高6.48个百分点和3.98个百分点，这说明融合自编码器和隔离森林的递进式异常检测方法显著提高了异常检测性能，AE-IF在充分学习正常数据特征信息的同时可有效识别异常数据. SDA-1D_CNN-GRU使用SDA进行数据降维和特征提取，再使用1D_CNN-GRU进行异常检测，但SDA-1D_CNN-GRU的3项指标均低于AE-IF，主要原因在于特征提取过程中舍弃的特征可能包含重要信息，导致在训练过程中无法充分学习各个数据特征的信息，而AE-IF使用全部特征进行训练学习，保留了丰富的数据信息. AE-IF的召回率和F_1值要低于EPCA-HG-CNN，原因是超图技术可识别数据的信息性和非信息性特征，相较于AE-IF对异常样本有更好的分类效果，但AE-IF的精确率相较于EPCA-HG-CNN提升2.27个百分点.

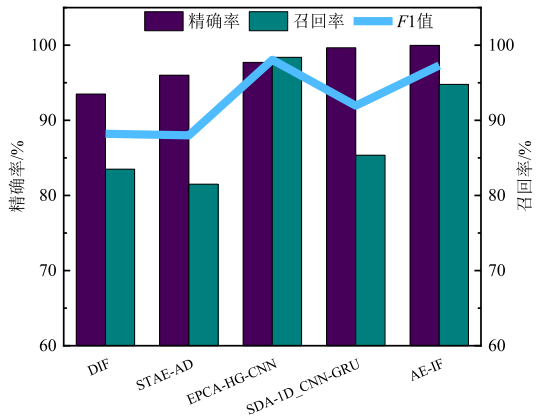


图13 检测结果对比

3.4.3 模型运行时间对比

模型运行时间对比实验均在操作系统为Windows10，处理器为Intel(R) Core(TM) i7-10700 CPU@2.90 GHz，RAM为16.0 GB的环境中实现. 通过实验记录了LOF、OCSVM、Singer-IF、Singer-AE以及AE-IF的训练时间及测试时间，结果如图14所示. 值得指出的是，这里的时间数据均是针对数据集规模的训练与测试时间统计，并非单一数据样本的实验结果，但可以根据实验用数据集样本数估算单一数据样本的平均训练或检测时间.

由图14可知，AE-IF相较于OCSVM，训练时间和检测时间均有大幅下降，同时检测时间与LOF相比也有所减少，原因是OCSVM和LOF需分别计算距离和密度，所以计算复杂度较高，导致检测时间增加. 相反，AE-IF中引入的IF可避免计算距离或密度，所以其时间开销不随数据维度的增加而增加. AE-IF相比于单一检测模型的复杂度有所上升，所以其训练时间要长于Singer-IF和Singer-AE. 但检测时间相较于Singer-IF有所下降，因为通过自编码器的基础性检测后仅需将异常数据输入到隔离森林中进行检测，数据量的减少使检测时间下降. 结合表6、图13可知，本文提出的AE-IF检测效果更具优势.

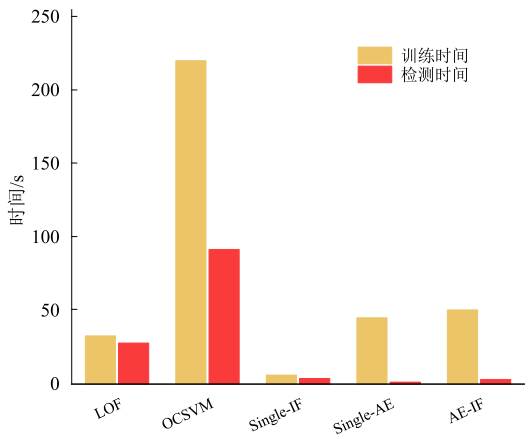


图14 模型训练与检测时间对比

3.4.4 泛化性分析

为了检验AE-IF的泛化性能，采用WADI数据集进行AE-IF泛化能力检验. 选取DIF、UAE[20]、MAD-GAN[21]、Single-IF、Single-AE进行对比. 其中，AE-IF、Single-IF、Single-AE重新对WADI数据集进行训练和检测. 对比结果如图15所示.

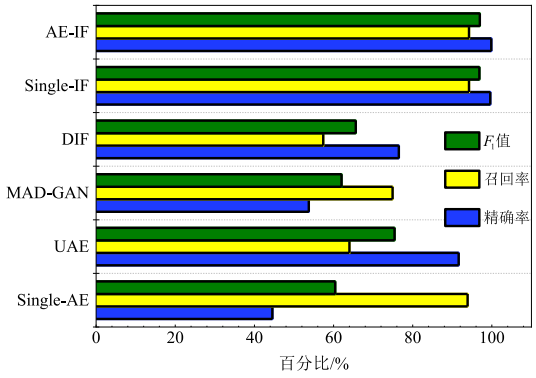


图15 模型泛化性指标对比

由图15可知，在WADI数据集下，AE-IF的精确率、召回率及F_1值均为最优值，分别为99.86%、94.23%、96.96%. 相比于Single-IF在精确率上提升0.26个百分点，相较于Single-AE在F_1值上提升36.55个百分点，说明基于两阶段递进式检测的AE-IF同样可有效学习WADI数据特征信息. DIF和USE的3项指标均比AE-IF差，主要原因是DIF没有考虑到数据间特征信息的重要性，而USE忽略了高维数据的特性. MAD-GAN使用

GAN训练的生成器和鉴别器进行异常检测并同时考虑整个变量集来捕获变量之间的潜在交互,AE-IF与其相比召回率和精确率分别提升19.31个百分点和46.11个百分点. 以上结果证明两阶段递进式异常检测方法AE-IF在WADI数据集下依然有效,即AE-IF泛化性较好.

4 结论

为了应对正在兴起的以集成工业互联网等新一代信息技术实现转型升级的水处理系统所面临日益严峻的入侵威胁,并进行有效的异常检测,提出了融合自编码器和隔离森林的水处理系统递进式异常检测方法. 该方法利用自编码器的重构误差和隔离森林的异常度量进行递进式异常检测;自编码器对于明显异常的样本有较高的敏感性,利用其捕获数据的内在特征表示进行基础性检测;对于被归类为异常的数据利用隔离森林算法可处理更复杂异常情况的能力进行高级检测,综合使用两阶段递进式异常检测实现了整体性能提升. 该方法在SWaT数据集下的异常检测准确率为95.13%,准确率相较于传统方法平均提升31.86个百分点,同时异常检测误报率取得了较大幅度的下降,仅为0.30%,相较于其他方法表现出明显优势. 且实验测试结果表明:本文方法在WADI数据集下有较优的泛化性,能更好地适应数字化程度越来越高的基于工业互联网的水处理系统应用场景.

参考文献

[1] ALIMI O A, OUAHADA K, ABU-MAHFOUZ A M, et al. A review of research works on supervised learning algorithms for SCADA intrusion detection and classification[J]. Sustainability, 2021, 13(17): 9597.

[2] 孙海丽, 龙翔, 韩兰胜, 等. 工业物联网异常检测技术综述[J]. 通信学报, 2022, 43(3): 196-210.
SUN H L, LONG X, HAN L S, et al. Overview of anomaly detection techniques for industrial Internet of Things[J]. Journal on Communications, 2022, 43(3): 196-210. (in Chinese)

[3] BHAMARE D, ZOLANVARI M, ERBAD A, et al. Cybersecurity for industrial control systems: A survey[J]. Computers & Security, 2020, 89: 101677.

[4] 尚文利, 石贺, 赵剑明, 等. 基于SAE-LSTM的工艺数据异常检测方法[J]. 电子学报, 2021, 49(8): 1561-1568.
SHANG W L, SHI H, ZHAO J M, et al. An anomaly detection method of process data based on SAE-LSTM[J]. Acta Electronica Sinica, 2021, 49(8): 1561-1568. (in Chinese)

[5] 席亮, 王瑞东, 樊好义, 等. 基于样本关联感知的无监督深度异常检测模型[J]. 计算机学报, 2021, 44(11): 2317-2331.
XI L, WANG R D, FAN H Y, et al. Sample-correlation-aware unsupervised deep anomaly detection model[J]. Chinese Journal of Computers, 2021, 44(11): 2317-2331. (in Chinese)

[6] 施媛波. 变分自编码器和注意力机制的异常入侵检测方法[J]. 重庆邮电大学学报(自然科学版), 2022,34(6):1071-1078.
SHI Y B. Anomaly intrusion detection method based on variational autoencoder and attention mechanism[J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2022, 34(6): 1071-1078. (in Chinese)

[7] ELNOUR M, MESKIN N, KHAN K M. Hybrid attack detection framework for industrial control systems using 1D-convolutional neural network and isolation forest[C]//2020 IEEE Conference on Control Technology and Applications (CCTA). Piscataway: IEEE, 2020: 877-884.

[8] ELNOUR M, MESKIN N, KHAN K, et al. A dual-isolation-forests-based attack detection framework for industrial control systems[J]. IEEE Access, 2020, 8: 36639-36651.

[9] XIONG Z M, ZHU D F, LIU D F, et al. Anomaly detection of metallurgical energy data based on iForest-AE[J]. Applied Sciences, 2022, 12(19): 9977.

[10] 胡向东, 李之涵. 基于胶囊网络的工业互联网入侵检测方法[J]. 电子学报, 2022, 50(6): 1457-1465.
HU X D, LI Z H. Intrusion detection method based on capsule network for industrial Internet[J]. Acta Electronica Sinica, 2022, 50(6): 1457-1465. (in Chinese)

[11] 胡向东, 吕高飞, 白银. 基于优化支持向量回归的工业互联网安全态势预测方法[J]. 电子学报, 2023, 51(2): 446-454.
HU X D, LÜ G F, BAI Y. A method of security situation prediction for industrial Internet based on optimized support vector regression[J]. Acta Electronica Sinica, 2023, 51(2): 446-454. (in Chinese)

[12] 杨晓晖, 张圣昌. 基于多粒度级联孤立森林算法的异常检测模型[J]. 通信学报, 2019, 40(8): 133-142.
YANG X H, ZHANG S C. Anomaly detection model based on multi-grained cascade isolation forest algorithm[J]. Journal on Communications, 2019, 40(8): 133-142. (in Chinese)

[13] TSAI D M, JEN P H. Autoencoder-based anomaly detection for surface defect inspection[J]. Advanced Engineering Informatics, 2021, 48: 101272.

[14] 袁非牛, 章琳, 史劲亭, 等. 自编码神经网络理论及应用综述[J]. 计算机学报, 2019, 42(1): 203-230.
YUAN F N, ZHANG L, SHI J T, et al. Theories and applications of auto-encoder neural networks: A literature survey[J]. Chinese Journal of Computers, 2019, 42(1):

203-230. (in Chinese)
[15] GOH J, ADEPU S, JUNEJO K N, et al. A dataset to support research in the design of secure water treatment Systems[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017: 88-99.
[16] AHMED C M, PALLETI V R, MATHUR A P. WADI: A water distribution testbed for research in the design of secure cyber physical systems[C]//Proceedings of the 3rd International Workshop on Cyber-Physical Systems for Smart Water Networks. New York: ACM, 2017: 25-28.
[17] MACAS M, WU C M. An unsupervised framework for anomaly detection in a water treatment system[C]//2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA). Piscataway: IEEE, 2019: 1298-1305.
[18] PRIYANGA S, KRITHIVASAN K, PRAVINRAJ S, et al. Detection of cyberattacks in industrial control systems using enhanced principal component analysis and hypergraph- based convolution neural network (EPCA-HG-CNN) [J]. IEEE Transactions on Industry Applications, 2020, 56(4): 4394-4404.
[19] XIE X, WANG B, WAN T C, et al. Multivariate abnormal detection for industrial control systems using 1D CNN and GRU[J]. IEEE Access, 2020, 8: 88348-88359.
[20] PANG G S, SHEN C H, CAO L B, et al. Deep learning for anomaly detection: A review[J]. ACM Computing Surveys, 2022, 54(2): 1-38.
[21] LI D, CHEN D C, JIN B H, et al. MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019: 703-716.

作者简介

胡向东 男，1971 年生，四川广安人．博士，重庆邮电大学教授，博士生导师．主要研究方向为智能感知、网络化测量与工业互联网安全等．
E-mail: huxd@cqupt.edu.cn

刘 浪 男，1999 年生，重庆江津人．重庆邮电大学硕士研究生．主要研究方向为工业互联网安全．
E-mail: liul1652725@163.com