DOI:10.11918/202412064

考虑代价敏感性的增量式入侵检测机制

金志刚,陈旭阳,刘泽培,武晓栋,刘 凯,丁 禹
(天津大学 电气自动化与信息工程学院,天津 300072)

摘 要:增量式入侵检测系统训练过程中存在数据不平衡和遗忘不平衡问题,易导致高错分代价样本漏检,为此,提出一种考虑代价敏感性的增量式入侵检测机制(incremental intrusion detection mechanism considering cost-sensitive,IIDM-CCS)。首先,提出一个改进的生成对抗网络,引入检测模型和目标增强信息熵损失函数,抑制非特定类生成,增加生成样本中特定类的比例;其次,提出一种多教师知识蒸馏方法,通过基础检测模型和旧检测模型为增量训练过程施加约束,指导新检测模型学习如何检测新特定类,并缓解对旧特定类的灾难性遗忘;最后,在知识蒸馏过程中设计代价敏感梯度加权损失函数引导模型关注特定类,利用软标签增量蒸馏损失函数提高模型对旧特定类的泛化能力。在 CSE-CIC-IDS2018 和 EDGE-IIoT 数据集上与其他增量式入侵检测方法进行对比,实验结果表明:增量训练后,所提机制检测特定类时的平均召回率分别比对比方法高 51.45% 和 60.32%;同时,检测旧特定类时的召回率较常规增量式入侵检测方法有明显提升。所提机制可以适配于现有的增量式入侵检测系统,使其具备代价敏感性,进一步提升了入侵检测系统对特定类别的识别精度并缓解了其在增量过程中的灾难性遗忘。

关键词:入侵检测系统;增量学习;生成对抗网络;数据增强;代价敏感
中图分类号:TP393.08 **文献标志码**:A **文章编号**:0367-6234(2026)09-0017-10

Incremental intrusion detection mechanism considering cost-sensitive

Jin Zhigang, Chen Xuyang, Liu Zepei, Wu Xiaodong, Liu Kai, Ding Yu
(School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China)

Abstract: Aiming at the issues of data imbalance and forgetting imbalance during training, which makes incremental intrusion detection system prone to missed detection of the specific attack classes with high-cost misclassification, an incremental intrusion detection mechanism considering cost-sensitive (IIDM-CCS) is proposed. First, an improved generative adversarial network that incorporates a detection model and an information entropy loss function for target augmentation is proposed to suppress the generation of non-specific classes and increase the proportion of specific classes in the generated samples. Second, a multi-teacher distillation method is proposed, which employs the base and old detection models to constrain the incremental training process, thereby guiding the new detection model to detect the new specific classes and alleviate the catastrophic forgetting of old specific classes. Finally, a cost-sensitive-based gradient weighted loss is designed to guide the model focus on specific classes, and a soft-label incremental distillation loss is leveraged to enhance the generalization capability on old specific classes. The comparative experimental results with other incremental intrusion detection methods show that the average recall rate of the proposed mechanism in detecting specific classes is 51.45% and 60.32% higher than those of the compared methods after incremental training. The proposed mechanism can be adapted to existing incremental intrusion detection systems to introduce cost-sensitive, thereby improving the detection accuracy for specific attack classes and mitigating the catastrophic forgetting during the incremental training.

Keywords: intrusion detection system; incremental learning; generative adversarial network; data augmentation; cost sensitive

信息技术的发展使得越来越多的设备和应用活跃于网络中,发生攻击行为的频率也随之上升[1]。入侵检测系统(intrusion detection system,IDS)可以及时发现攻击行为,基于深度学习的 IDS 因自动化程度和准确性高等优势而成为了研究的热点。传统 IDS 多采用离线学习,无法处理开放网络环境中不

收稿日期:2024-12-23;**录用日期**:2025-03-04;**网络首发日期**:2025-03-25
网络首发地址:https://link.cnki.net/urlid/23.1235.t.20250324.1711.006
基金项目:国家自然科学基金(52171337)
作者简介:金志刚(1972—),男,教授,博士生导师
通信作者:金志刚,zgjin@tju.edu.cn

断涌现的新攻击[2],需要不断地在全部数据上重新训练模型。但在有隐私保护和存储限制的情境中,当出现新攻击时,获取全部旧数据重新训练是不可行的。仅用新数据微调模型参数会导致灾难性遗忘[3],即 IDS 会逐渐遗失检测旧类的能力。在此情况下,基于增量学习的 IDS 受到了关注。

增量学习中,模型接收一系列仅含有新类样本的任务以学习新知识[4],并通过一些算法维持其在旧类上的表现。基于数据回放的增量式 IDS 使用内存中少量旧类示例约束模型更新,这些示例来自于历史任务的采样或是生成样本[5-7]。基于知识蒸馏(knowledge distillation,KD)的增量式 IDS 将旧模型作为教师进行知识蒸馏,将检测旧类的知识传递至作为学生的新模型[8-9]。Nugraha 等[10]提出一种使用滑动窗口存储示例的增量式 IDS;Jin 等[11]将示例用于 KD 中,实现了增量式联邦 IDS。然而增量式 IDS 的流量监测具有数据不平衡和遗忘不平衡等特点,导致了对高错分代价攻击的漏检问题。

一方面,以上特点中的数据不平衡[12]指网络中正常流量的数量远高于攻击流量,以及各攻击类别间流量的数量失衡,导致 IDS 对攻击流量的检测效果下降。研究者大多利用过采样或数据增强方法解决数据不平衡问题。其中,数据增强可以利用生成对抗网络(generative adversarial network,GAN)补充稀缺的攻击数据,更适合 IDS 使用。Park 等[13]利用边界均衡生成对抗网络生成样本为 IDS 进行数据增强;Ding 等[14]在 IDS 中集成多个 GAN 以提高生成样本质量。以上研究平等地对待所有少数类,但现实中的 IDS 有代价敏感的特殊问题[15],即出于设备或服务的原因,在某些特定少数类上的错分代价会更高[16]。例如,物联网中 heartbleed 攻击会窃取敏感信息,使智能家居、工业控制等有隐私保护要求的业务受到致命打击[17];车联网中 sybil 攻击可引发错误决策,造成严重事故[18]。一旦 IDS 漏检这样的特定攻击,将会对网络造成比其他攻击更大的危害,因此,在应对数据不平衡时,IDS 需根据网络特性重点提高检测高错分代价的特定类的准确率。

另一方面,增量学习不能完全消除灾难性遗忘。更进一步地,Xu 等[19]首次揭示类间存在遗忘不平衡现象,即各个旧类被遗忘的程度不同。现有增量式 IDS 研究尚未针对遗忘不平衡问题提出应对措施,相关研究中缓解遗忘的算法依然平等地对待所有类。经对相关研究[10-11,20-23]实验环节的调研发现,一个类别被遗忘的程度与该类样本在训练集的数量比例呈一定相关性,旧类样本较少时,灾难性遗忘的影响更为严重。遗忘不平衡与数据不平衡的共同作用使增量式 IDS 检测某些高错分代价的特定类时难以令人满意。因此,增量式 IDS 会因遗忘不平衡进一步扩大特定类的漏检风险。

综上,为提高现有增量式 IDS 对少数高错分代价特定类的检测效果,本文提出一种考虑代价敏感性的增量式入侵检测机制(incremental intrusion detection mechanism considering cost-sensitive,IIDM-CCS)。该机制通过 GAN 模块增强特定类样本,通过增量 KD 模块提取用于检测和记忆特定类的关键信息,得到一个代价敏感的增量式 IDS 模型。设计实验探究各模块及损失函数的效用,并与其他增量式 IDS 方法进行对比,验证所提机制的适用性和有效性。

1 整体设计

IIDM-CCS 结构如图 1 所示,由 GAN 模块和增量 KD 模块两部分构成。IIDM-CCS 训练时,首先在 GAN 模块中使用随机信号和存储器中的真实样本示例训练一个 GAN,并通过生成器得到增强特定类的生成样本;之后在增量 KD 模块中采取多教师蒸馏方法,利用基础检测模型和旧学生模型作为教师指导当前任务的学生模型进行知识蒸馏,使用生成

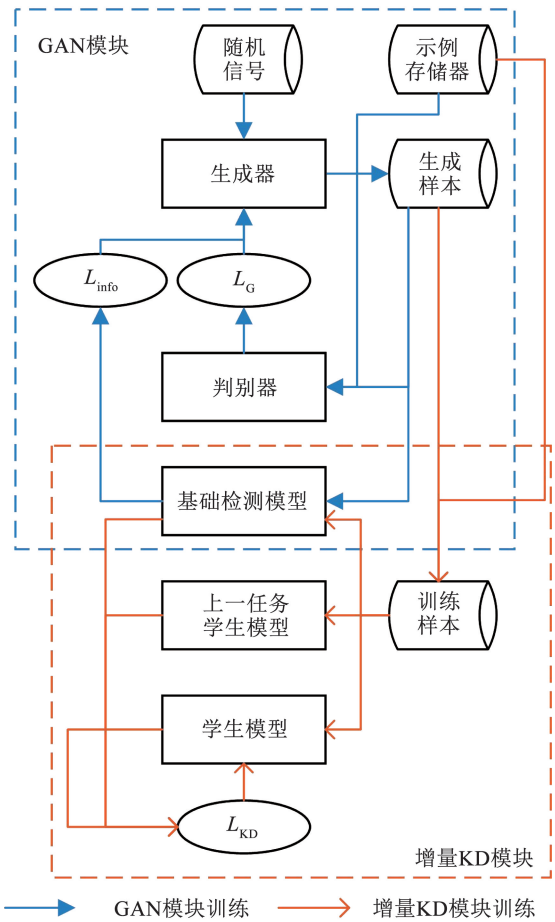


图 1 IIDM-CCS 结构

Fig. 1 Structure of IIDM-CCS

样本和示例训练得到一个对特定类更敏感的检测模型。其中，学生模型的输出维度与分类任务类别数量相同，并随着增量训练的过程不断增加。

IIDM-CCS 具有通用性，可以应用于任何增量式 IDS 中。在原增量式 IDS 检测模型的基础上，IIDM-CCS 无需再额外获取全部样本即可训练一个代价敏感的检测模型，增强 IDS 对高错分代价特定类的检测能力。训练开始前，获取欲改善的增量式 IDS 中的检测模型作为基础检测模型。示例存储器在基础检测模型的训练阶段存储了少量真实样本，用于辅助 GAN 学习真实样本的特征。

2　各模块设计

2.1　GAN 模块

一般地，给定真实样本 x 及其标签 y，向生成器 G 输入服从分布 $p(z)$ 的随机噪声 z，输出生成样本 $G(z)$，G 的损失函数为

$$L_{\mathrm{G}} = -E_{z\sim p(z)}[\log(D(G(z)))] \tag{1}$$

式中：$E_{z\sim p(z)}$ 为当 z 服从分布 $p(z)$ 时对输入变量的期望值；$D(G(z))$ 为判别器 D 区分 $G(z)$ 是否为真样本的概率，以此尽可能地使生成样本近似于真实样本。

原始的 GAN 学习并模仿的对象是全体类别，但这一方面不满足本文增强特定类的需要，另一方面非特定类的知识会为特定类的生成引入干扰。相关研究[24-25]表明，判别器 D 可以有效提取样本的特征，因此可以从区分任务泛化到分类任务中，区分任务和分类任务可以相互促进。相应地，分类模型也具备在区分任务中提取特征的能力[26]，可以与判别器 D 共同引导生成器 G 生成质量更高的样本。因此本文将基础检测模型 M_{base} 加入 GAN 指导生成器 G 学习特定类的知识。具体而言，在判别器 D 判断生成样本是否接近真实样本的同时，M_{base} 输出概率预测向量 $\boldsymbol{P}(G(\boldsymbol{z}_i), M_{\mathrm{base}}) = [p_1, p_2, \cdots, p_i, \cdots, p_n]$，其中 p_i 为样本属于类别 Y_i 的概率。本文使用信息熵引导生成器 G 增加特定类的比例。在信息论中，信息熵可以量化事件的不确定性，当各类出现的概率相等时取得最大值。令特定类的标签集合为 Y_{spec}，其他非特定类作为背景类，其标签集合为 Y_{bg}。目标增强信息熵损失函数为

$$L_{\mathrm{info}} = -\left[-\left(\sum_{Y_i\in Y_{\mathrm{bg}}} p_i\right)\log\left(\sum_{Y_i\in Y_{\mathrm{bg}}} p_i\right) - \sum_{Y_i\in Y_{\mathrm{spec}}}(p_i\log p_i)\right] \tag{2}$$

L_{info} 将各非特定类看作一个背景类，从而在一定程度上对其进行了抑制。当生成器 G 可以均匀地生成背景类与各个特定类时，L_{info} 得到最小值，使生成样本中特定类的比例得到增加，同时保证类别之间的平衡，避免 GAN 模式崩溃。

最终，训练生成器 G 的损失函数为

$$L_{\mathrm{total}} = L_{\mathrm{G}} + \alpha L_{\mathrm{info}} \tag{3}$$

式中 α 为权重系数。

2.2　增量 KD 模块

KD 是一种常用的模型压缩方法，学生模型通过模仿教师模型输出的分布来传递教师模型提供的解释、比较等特权信息[27]，从而使轻量级模型逼近复杂模型的性能。本节着眼于提升 IIDM-CCS 的分类性能和抗遗忘能力，提出一种多教师 KD 方法，并根据不同教师模型的引导目标设计损失函数。

2.2.1　增量 KD 过程

在任务 T^i 中，增量 KD 模块使用一个轻量化模型作为学生模型 M_{s}^i，并选取 M_{base}^i 和上一任务的学生模型 M_{s}^{i-1} 共同作为教师模型进行知识蒸馏。用 $G(z)$ 与示例构建训练数据集，获得 3 个模型的输出。M_{s}^i 模仿 M_{base}^i 的输出以学习分类所需丰富的知识，采用代价敏感梯度加权损失函数 L_{we} 对不同类的训练速度进行差异化；M_{s}^i 模仿 M_{s}^{i-1} 的输出以记忆旧类，采用软标签增量蒸馏损失函数 L_{sl} 提高对旧类的泛化能力。需注意的是，M_{s}^i 仅需辨别流量是否属于特定类，而 M_{base}^i 则针对全体类别完成分类，二者输出维度不同。M_{base}^i 包含丰富非特定类的知识，可以赋予学生模型较好的泛化能力，但是对非特定类的关注又会为学生模型引入过多干扰。因此，将全部非特定类作为一个背景类 Y_{bg}，其标签重新标注为 y_{bg}。上述过程如图 2 所示。

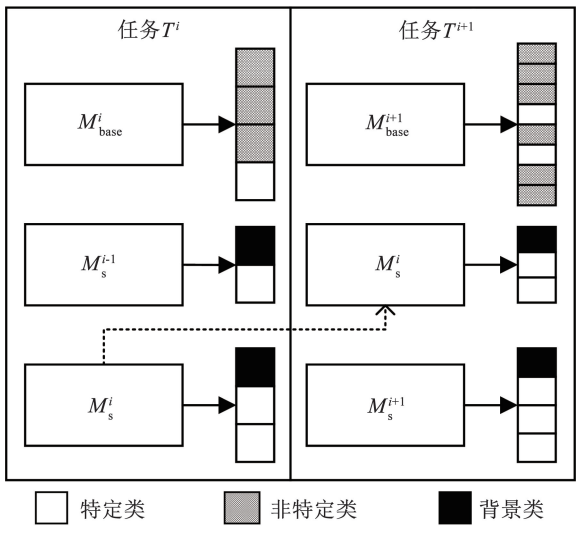


图 2　增量 KD 过程

Fig. 2　Process of incremental KD

2.2.2　代价敏感梯度加权损失函数

KD 常用 Kullback-Leibler 散度或交叉熵损失函

数比较学生与教师输出之间的距离。本文设计了代价敏感梯度加权损失函数 L_{we},以测量 M_s^i 与 M_{base}^i 输出间的差距。在任务 T^i 中,由 b 个样本 x_j^i 及其标签 y_j^i 构成的 mini batch 在训练中的分类损失函数为

$$L_{ce} = \frac{1}{b}\sum_{j=1}^{b} H_{ce}(\boldsymbol{P}(x_j^i, M_s^i), \boldsymbol{O}(y_j^i)) \tag{4}$$

式中:H_{ce} 为交叉熵损失,$\boldsymbol{P}(x_j^i, M_s^i)$ 为 M_s^i 对 x_j^i 输出的概率分布,$\boldsymbol{O}(y_j^i)$ 为 y_j^i 的独热向量。在一个 mini batch 中,传统的交叉熵损失函数在背景类及特定类上形式相同。但在 IDS 中,背景类样本容量远多于特定类,简单地使用交叉熵损失函数会导致检测模型偏向于学习背景类知识。在增量的过程中,背景类和特定类的样本容量有着不同的增长速度,加剧了数量不平衡。

因此,本文采用代价敏感梯度加权损失函数 L_{we} 来调整分类网络对背景类和特定类的学习速度。对于属于 y_j^i 的样本 x_j^i,M_s^i 的隐层输出向量 $\boldsymbol{h}$ 中第 k 个输出值 h_k 的梯度信息为:

$$g_j^i(k) = \frac{\partial H_{ce}(\boldsymbol{P}(x_j^i, M_s^i), \boldsymbol{O}(y_j^i))}{\partial h_k} \tag{5}$$

$$\frac{\partial H_{ce}(\boldsymbol{P}(x_j^i, M_s^i), \boldsymbol{O}(y_j^i))}{\partial h_k} = \begin{cases} P(x_j^i, M_s^i)_k - 1, & k = y_j^i \\ P(x_j^i, M_s^i)_k, & k \neq y_j^i \end{cases} \tag{6}$$

式中 $P(x_j^i, M_s^i)_k$ 为 M_s^i 对第 k 类预测的 Softmax 概率。损失函数中特定类和背景类的权重分别为:

$$W_{spec} = \frac{\sum_{j=1}^{b} | g_j^i(y_j^i) | \cdot I(y_j^i \neq y_{bg})}{\sum_{j=1}^{b} I(y_j^i \neq y_{bg})} \tag{7}$$

$$W_{bg} = \frac{\sum_{j=1}^{b} | g_j^i(y_j^i) | \cdot I(y_j^i = y_{bg})}{\sum_{j=1}^{b} I(y_j^i = y_{bg})} \tag{8}$$

式中 $I(\cdot)$ 为指示函数,括号内条件为真时取 1,否则取 0。

用教师模型的输出 $\boldsymbol{P}(x_j^i, M_{base}^i)$ 作为学生模型模仿的目标以进行 KD。最终得到代价敏感梯度加权损失函数,公式为

$$L_{we} = \frac{1}{b}\sum_{j=1}^{b} \frac{| g_j^i(y_j^i) | \cdot H_{ce}(\boldsymbol{P}(x_j^i, M_s^i), \boldsymbol{P}(x_j^i, M_{base}^i))}{W_{spec} \cdot I(y_j^i \neq y_{bg}) + W_{bg} \cdot I(y_j^i = y_{bg})} \tag{9}$$

L_{we} 通过增强隐藏层对应神经元输出的梯度信息分别调整特定类和背景类的学习速度,避免检测模型过于偏向学习背景类,增加对高错分代价特定类的关注。

2.2.3 软标签增量蒸馏损失函数

考虑到维持旧类的检测性能,借助旧模型 M_s^{i-1} 为新任务的模型 M_s^i 建立监督关系,抵抗灾难性遗忘现象。训练时,将训练样本 x_j^i 给予 M_s^{i-1},得到其输出的概率预测 $\boldsymbol{P}(x_j^i, M_s^{i-1})$。然后利用 $\boldsymbol{P}(x_j^i, M_s^{i-1})$ 替换 $\boldsymbol{O}(y_j^i)$ 对应的部分作为软标签,记为 $\boldsymbol{S}(\boldsymbol{P}(x_j^i, M_s^{i-1}), \boldsymbol{O}(y_j^i))$,如图 3 所示。最后,通过 Kullback-Leibler 散度 D_{KL} 比较 M_s^i 的输出与软标签间的距离以完成 KD。软标签增量蒸馏损失函数 L_{sl} 补充了 KD 阶段其他损失函数没有考虑增量变化的不足,其计算式为

$$L_{sl} = D_{KL}(\boldsymbol{P}(x_j^i, M_s^i) \parallel \boldsymbol{S}(\boldsymbol{P}(x_j^i, M_s^{i-1}))) \tag{10}$$

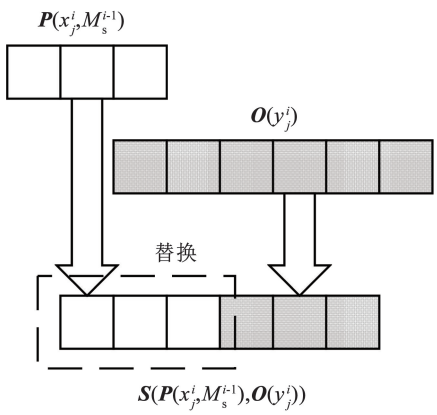


图 3 软标签构造过程

Fig. 3 Process of build soft labels

综上,增量 KD 模块的总损失函数为

$$L_{KD} = L_{ce} + \beta L_{we} + (1-\beta) L_{sl} \tag{11}$$

式中 β 为权重系数。

增量 KD 模块最终得到的学生模型 M_s^i 即为 IIDM-CCS 中用于检测流量中特定类的模型。

3 实 验

3.1 数据集与模型结构

本文在 CSE-CIC-IDS2018[28] 和 EDGE-IIoT[29] 数据集上进行实验,两者的样本分布见表 1。为检验所提机制在数据不平衡时的效果,选取样本量最少的几种攻击作为特定类,并添加至各个任务中进行学习。CSE-CIC-IDS2018 是在 IDS 研究中被广泛使用的数据集,涵盖了多种攻击类型,能模拟现实中复杂的网络情况。因数据量庞大,随机抽取 1% 的样本进行实验并划分为两个任务[23]。其中,正常流量占比高达 73.79%,而被选为特定类的 Infiltration 类和 Bot 类占比极低,分别为 1.95% 和 3.46%。物联网中大量的隐私数据通常会使 IDS 面临代价敏感问题[17],因此,选取物联网安全数据集 EDGE-IIoT 对所提机制在物联网场景中的表现进行实验。将 EDGE-IIoT 划分为 3 个任务,抽取 20% 的样本作为

测试集。其中，正常流量占比达到 42.94%，被选为特定类的 Password 类、Fingerprinting（FP）类和 MITM 类占比分别为 17.29%、1.73% 和 2.10%。基础检测模型采用全部训练样本，示例存储器依据训练集数据分布的比例抽取 1 000 条示例用于 IIDM-CCS 训练。

表 1　数据集样本分布

Tab. 1　Sample distribution of dataset

数据集	任务	类别	是否为特定类	训练集样本量	测试集样本量	样本量占比/%
CSE-CIC-IDS2018	1	Normal	否	48 898	12 224	73.79
		DDoS	否	5 502	1 375	8.30
		Infiltration	是	1 295	324	1.95
	2	Bot	是	2 290	572	3.46
		Brute Force	否	3 047	762	4.60
		DoS	否	5 234	1 309	7.90
EDGE-IIOT	1	Normal	否	19 949	4 852	42.94
		Password	是	7 984	2 005	17.29
	2	DDoS	否	8 401	2 160	18.28
		FP	是	814	187	1.73
	3	Backdoor	否	8 172	2 023	17.65
		MITM	是	988	226	2.10

为避免 IIDM-CCS 应用于其他增量式 IDS 时引入过多额外开销，本文为 GAN 模块和增量 KD 模块涉及的模型设计了轻量化的结构，在特征提取过程中仅使用了全连接（fully connected，FC）层、批量归一化（batch normalization，BN）层和修正线性单元（rectified linear unit，ReLU）激活函数，如图 4 所示。图 4 中各模块结构仅作参考，在实际应用中可根据具体情况和需要进行调整。

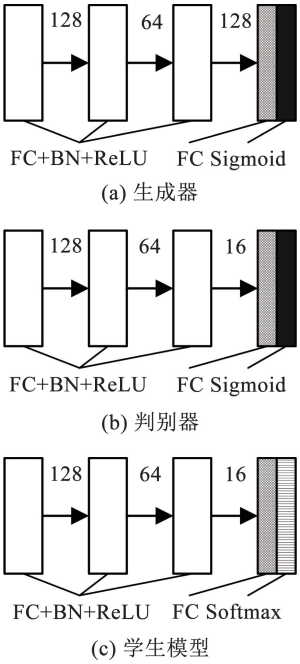


图 4　实验所用模型结构

Fig. 4　Structure of models for experiments

3.2　实验环境与评价指标

实验在安装了 NVIDIA RTX2060 的主机上运行，相关程序利用 Python 3.8 编写。标记正确分类的异常样本为 T_P，正确分类的正常样本为 T_N，错误分类的正常样本为 F_P，错误分类的异常样本为 F_N。采用准确率 A、召回率 R、精确率 P 和 F_1 分数作为评价指标，计算式分别为：

$$A=\frac{T_P+T_N}{T_P+F_N+T_N+F_P} \tag{12}$$

$$R=\frac{T_P}{T_P+F_N} \tag{13}$$

$$P=\frac{T_P}{T_P+F_P} \tag{14}$$

$$F_1=\frac{2PR}{P+R} \tag{15}$$

本文的研究目标是提升 IDS 对高错分代价特定类的防护能力，需要模型能够降低对特定类的漏检风险，因此将着重评估模型在召回率上的表现。

3.3　消融实验

为探究 GAN 模块和增量 KD 模块对模型整体表现的影响，本文利用 CSE-CIC-IDS2018 进行消融实验。消融实验涉及 3 个可变条件：是否加入 GAN 模块，是否有基础检测模型参与 KD 和是否有旧学生模型参与 KD。实验所涉及模型及差别见表 2，结果见表 3，基础检测模型由其他增量式 IDS 方法[23]训练获得。

表 2 消融实验模型

Tab. 2 Models of ablation experiment

模型	GAN 模块	增量 KD 模块	
		基础检测模型	旧学生模型
Baseline			
A1		●	●
A2	●		
A3	●	●	
A4	●	●	●

注:●表示模型包含此部分。

实验中模型 Baseline 出现了过拟合,几乎将所有 Infiltration 样本分为了背景类,导致 F_1 和 P 无法获取。说明直接使用原始数据训练会使检测模型被数据不平衡影响,无法学习少数类的特征,导致 IDS 失去检测高错分代价特定类的能力。另外,模型 Baseline 准确率依然超过模型 A1,说明数据不平衡的情况下,IDS 模型只需重视数据量较大的类别即可获得较高的准确率,如果仅依靠整体表现评价 IDS 是有缺陷的。单独加入 GAN 模块或增量 KD 模块均会导致模型在特定类上的 R 升高,同时加入时可达到最佳效果。

表 3 消融实验结果

Tab. 3 Results of ablation experiment %

任务	模型	A	R		P		F_1	
			Infiltration	Bot	Infiltration	Bot	Infiltration	Bot
1	Baseline	97.67	0		—		—	
	A1	97.38	18.52		37.27		24.74	
	A2	98.41	89.51		60.80		72.42	
	A3	99.50	96.27		84.32		89.91	
	A4	99.46	98.15		82.17		89.45	
2	Baseline	97.79	0	99.83	—	63.80	—	77.85
	A1	96.00	1.97	50.70	8.82	96.99	3.23	66.59
	A2	98.16	44.75	99.83	64.44	52.82	2.52	76.44
	A3	97.97	61.11	100.00	48.41	83.87	54.02	91.23
	A4	98.87	64.81	99.83	78.07	88.25	70.83	93.68

对比模型 Baseline 与 A2、模型 A1 与 A4 的实验结果可知:在其他条件相同时,加入 GAN 模块可以使特定类的 R 升高;模型 A4 较模型 A1 在任务 1 中 Infiltration 类的 R 高 79.63%,在任务 2 中 Infiltration 类和 Bot 类的 R 分别高 62.84% 和 49.13%,表明 GAN 模块能有效地对特定类进行数据增强,提升模型检测特定类的能力;模型 A4 除 Bot 类外的 P 均高于模型 A1,所有特定类 F_1 亦高于模型 A1,说明 GAN 模块生成样本时可以权衡特定类与其他类之间的平衡,不会使检测模型对其他类的样本过度地错检。以上结果证明加入 GAN 模块可在一定程度上抵抗数据不平衡问题的负面影响。

为确定增量 KD 模块的效用,对模型 A2、A3 和 A4 的实验结果进行分析。由于开始训练时尚不存在旧学生模型,模型 A3 和 A4 的训练条件是相同的,在任务 1 中的表现只因随机初始化而有轻微差别。模型 A3 较模型 A2 在任务 1 中 Infiltration 类的 R 高 6.76%,在任务 2 中 Infiltration 类和 Bot 类的 R 分别高 16.36% 和 0.17%,表明增量 KD 模块可以将基础检测模型中的特权信息传递给学生模型,提升其对特定类的检测能力。因此,即使没有在全部样本上进行训练,模型仍可取得较好的检测性能。而 Bot 类上的提升效果没有 Infiltration 类显著,可能是由于对旧类的遗忘减少了 Bot 类受到的干扰,使得所有模型在 Bot 类上都能有较好的表现。两个任务间,模型 A3 上 Infiltration 类的 R 比模型 A4 多下降了 1.82%,A 多下降了 0.94%,表明对旧模型的 KD 和 L_{sl} 可以为新任务提供在旧类上的约束,从而对抗灾难性遗忘。以上结果说明增量 KD 模块可以指导检测模型提升对特定类的检测能力,并在一定程度上缓解增量过程中的灾难性遗忘问题。

3.4 损失函数效果实验

为确定权重系数 β 对 L_{KD} 效果的影响,采用 CSE-CIC-IDS2018 训练集进行调参实验,结果见表 4。由表 4 可知:L_{we} 可以帮助模型识别当前任务中的特定类,增大 β 可使当前任务中特定类的 R 升高。L_{sl} 可帮助 IDS 模型保持对旧类的识别性能,减小 β 可使旧类的 R 升高。简单地增大两个损失函数的权重会导致 IIDM-CCS 面临增量学习中的稳定性-可塑性困境,即记忆旧知识和学习新知识的能

力难以兼得。因此，考虑到学习过程中对二者的恰当平衡，根据表 4 确定 $\beta=0.3$ 进行后续实验。

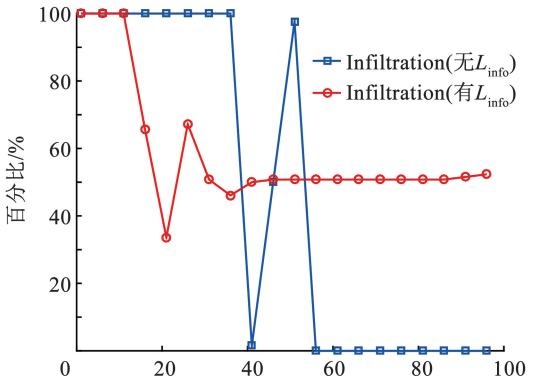
表 4　调参实验结果

Tab. 4　Results of parameter adjust experiment

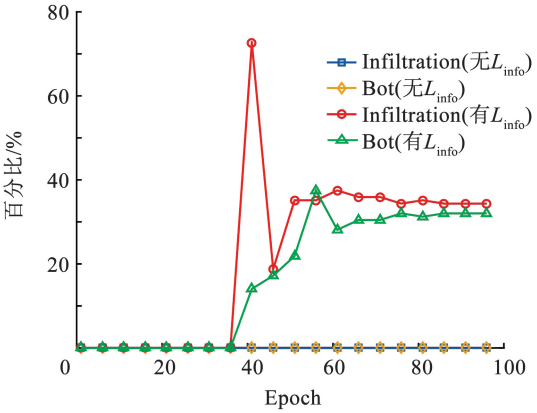
任务	β	A/%	R/%	
			Infiltration	Bot
1	0.1	99.29	92.12	
	0.3	99.29	92.28	
	0.5	99.29	92.59	
	0.7	99.28	92.66	
	0.9	99.28	98.82	
2	0.1	95.87	80.62	48.86
	0.3	97.47	70.66	99.78
	0.5	96.98	42.39	99.78
	0.7	96.78	38.22	99.87
	0.9	97.02	37.22	99.83

为验证损失函数是否达到预期功能，在 CSE-CIC-IDS2018 上对 L_{info} 和 L_{we} 进行实验。由于消融实验中模型 A3 和 A4 的对比已经体现 L_{sl} 的作用，不再重复实验。

比较有 L_{info} 和无 L_{info} 两种情况下特定类样本数量在生成样本中的占比，结果如图 5 所示。



(a) 任务1



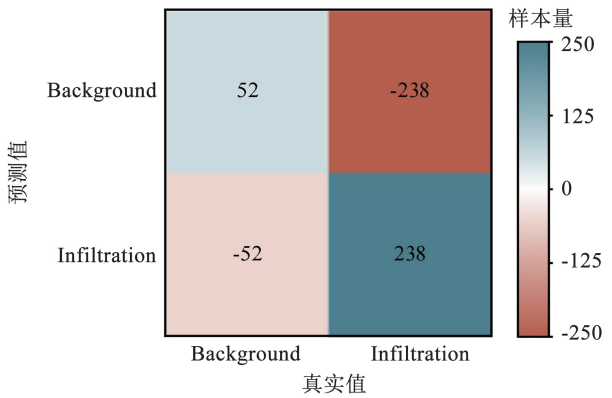
(b) 任务2

图 5　特定类样本的数量占比

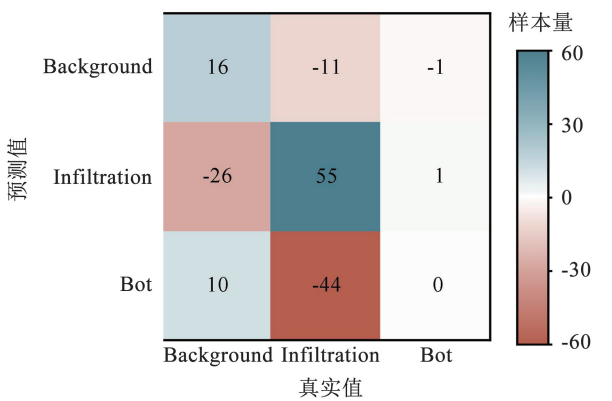
Fig. 5　The proportion of specific class samples

由图 5 可知：任务 1 中，虽然训练开始时无 L_{info} 的 GAN 的生成样本全部为 Infiltration 类，但此时生成器和判别器的损失值较高，生成样本质量偏低；伴随着 GAN 的训练，在有 L_{info} 的情况下，Infiltration 类样本在生成样本中的占比会稳定在 50% 左右，而在无 L_{info} 的情况下，生成器不再生成 Infiltration 类样本。任务 2 中，在无 L_{info} 的情况下，特定类的占比为 0，在有 L_{info} 的情况下，特定类的占比会稳定在 33% 左右，表明 L_{info} 可以引导 GAN 模块学习真实数据中数量较少的类。以上结果符合在两个任务中分别按二分之一和三分之一生成特定类样本的预期，表明 L_{info} 可使 GAN 均匀生成各类样本，实现在数据不平衡问题下对特定类的数据增强，避免模式崩溃而只生成数量占比大的类别。

通过实验比较 L_{we} 和 L_{ce} 对基础检测模型进行 KD 的效果，以热力图的形式展示了是否加入梯度加权的混淆矩阵间的成对差异，如图 6 所示。



(a) 任务1



(b) 任务2

图 6　不同损失函数下成对差异热力图

Fig. 6　Heatmaps based on pairwise differences in different loss

由图 6可以看出：采用 L_{we} 的模型在任务 1 和任务 2 中均比采用 L_{ce} 的模型检出更多 Infiltration 类样本；采用 L_{we} 的模型使得被漏检的 Bot 类样本从倾向于被分为背景类改变为倾向于被分为特定类。可

见,在训练中对背景类和特定类分开进行梯度加权,将数量少的特定类赋予更大的权重可以大幅提高 IDS 检出高错分代价特定类的能力,并缓解增量过程中的遗忘问题。

3.5 模型性能对比实验

为探究 IIDM-CCS 的有效性与普适性,与不同增量式 IDS 方法 FL-IIDS[22] 和 AAE-IIDS[23] 进行对比,实验结果见表 5、表 6 和图 7,其中 IIDM-CCS 的上标为选用的基础检测模型。在新任务中,对比方法对部分旧类产生严重遗忘,导致 F_1 分数无法获取。

表 5 CSE-CIC-IDS2018 数据集上模型对比结果

Tab. 5 Comparison results of the models on CES-CIC-IDS2018 dataset %

任务	模型	R		F_1	
		Infiltration	Bot	Infiltration	Bot
1	FL-IIDS	0		—	
	AAE-IIDS	28.70		44.60	
	IIDM-CCS[22]	96.27		83.31	
	IIDM-CCS[23]	98.15		89.45	
2	FL-IIDS	0.62	0	1.19	—
	AAE-IIDS	0	100.00	—	66.71
	IIDM-CCS[22]	86.72	52.45	58.42	67.11
	IIDM-CCS[23]	64.81	99.82	70.83	93.68

表 6 EDGE-IIoT 数据集上模型对比实验

Tab. 6 Comparison results of the models on EDGE-IIoT dataset %

任务	模型	R			F_1		
		Password	FP	MITM	Password	FP	MITM
1	FL-IIDS	31.42			43.66		
	AAE-IIDS	14.61			25.50		
	IIDM-CCS[22]	43.49			50.51		
	IIDM-CCS[23]	63.34			49.83		
2	FL-IIDS	0	87.17		—	62.10	
	AAE-IIDS	0	0		—	—	
	IIDM-CCS[22]	35.46	100.00		44.29	26.39	
	IIDM-CCS[23]	16.06	95.19		26.32	26.32	
3	FL-IIDS	0.05	0	100.00	0.10	—	19.28
	AAE-IIDS	6.13	0	0	11.56	—	—
	IIDM-CCS[22]	30.47	56.68	100.00	39.65	69.28	80.28
	IIDM-CCS[23]	26.13	59.36	95.58	32.85	32.85	27.00

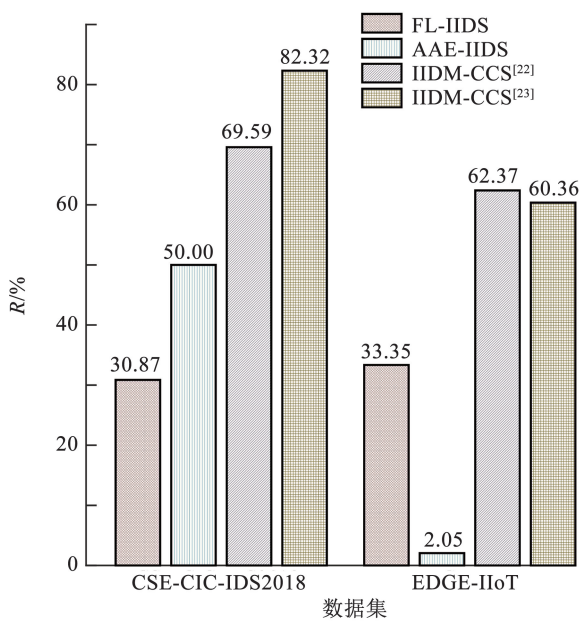


图 7 增量训练结束后对特定类的平均召回率

Fig. 7 Average recall rate of specific classes after incremental training

由表 5、表 6 和图 7 可以看出:在同一任务中,IIDM-CCS 对各特定类的 R 均超过或接近对比方法;增量训练结束后,在 CSE-CIC-IDS2018 和 EDGE-IIoT 数据集上,特定类的平均 R 分别比对比方法高 51.45% 和 60.32%,说明所提机制在一定程度上缓解了数据不平衡和遗忘不平衡的影响,有效提高了增量式 IDS 对高错分代价特定类的识别能力。其中,表 5 任务 2 中 IIDM-CCS 较 AAE-IIDS 对 Bot 类的 R 低 0.18%,原因是 IIDM-CCS 为维持对旧类的记忆会牺牲在新类上较少的可塑性。但此时 IIDM-CCS 较 AAE-IIDS 对 Infiltration 类的 R 高 64.81%,说明在对新特定类的检测性能与对比方法接近的情况下,IIDM-CCS 能显著缓解对旧特定类的遗忘。虽然对比方法有对抗灾难性遗忘的机制,但在增量训练结束后,已无法识别旧任务中的特定类,说明遗忘不平衡依然会导致对比方法对样本数量较少的部分特定类产生严重的遗忘,造成漏检,而 IIDM-CCS 在

旧特定类上的性能受基础检测模型的影响而衰退，但不会完全丧失，进而缓解原 IDS 因灾难性遗忘失去对高错分代价特定类识别能力的问题。同时，IIDM-CCS 对各个特定类的 F_1 分数均高于对比方法，说明 IIDM-CCS 也能适当兼顾对其他类别的准确分类，避免产生过分的错检。

综上，所提机制可以应用于现有增量式 IDS 中，在动态开放与数据不平衡的情景下，可有效提高 IDS 识别高错分代价特定类的能力。

4　结　论

1）本文提出了一种考虑代价敏感性的增量式入侵检测机制 IIDM-CCS，由 GAN 模块和增量 KD 模块两部分组成。

2）GAN 模块可以进行数据增强补充样本，使 IDS 模型能有效应对数据不平衡问题。引入基础检测模型和目标增强信息熵损失函数改进的 GAN 可以均匀提升生成样本中特定类的比例，避免模式崩溃。

3）基础检测模型和旧检测模型作为教师模型进行多教师知识蒸馏，可以为新任务中的检测模型提供特权信息和在旧类上的约束，仅在示例和生成样本上训练即可提升对特定类的检测能力，并降低灾难性遗忘造成的负面影响。

4）代价敏感梯度加权损失函数可以为特定类赋予更大的权重，提高 IDS 检出高错分代价特定类的能力，并帮助缓解增量过程中的遗忘问题。

5）在原增量式 IDS 检测模型的基础上，IIDM-CCS 无需再额外获取全部样本即可训练代价敏感的检测模型，增强了 IDS 对高错分代价特定类的检测能力。

参考文献

[1] Butun I, Osterberg P, Song Houbing. Security of the internet of things: vulnerabilities, attacks, and countermeasures[J]. IEEE Communications Surveys and Tutorials, 2020, 22(1): 616. DOI: 10.1109/COMST.2019.2953364

[2] Farrukh Y A, Wali S, Khan I, et al. AIS-NIDS: an intelligent and self-sustaining network intrusion detection system[J]. Computers and Security, 2024, 144: 103982. DOI: 10.1016/j.cose.2024.103982

[3] Li Ziwei, Xu Jiayi, Chao Weilun, et al. Visual analytics on network forgetting for task-incremental learning[J]. Computer Graphics Forum, 2023, 42(3): 437. DOI: 10.1111/cgf.14842

[4] Masana M, Liu Xialei, Twardowski B, et al. Class-incremental learning: survey and performance evaluation on image classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(5): 5513. DOI: 10.1109/TPAMI.2022.3213473

[5] Weiflog M, Protzel P, Neubert P. FETCH: a memory-efficient replay approach for continual learning in image classification[C]// 24th International Conference on Intelligent Data Engineering and Automated Learning. Évora: Springer, 2023: 418. DOI: 10.1007/978-3-031-48232-8_38

[6] Tran M T, Le T, Le X M, et al. Text-enhanced data-free approach for federated class-incremental learning[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR2024. Seattle: IEEE, 2024: 23870. DOI: 10.1109/CVPR52733.2024.02253

[7] Chen Chen, Li Ruizhe, Hu Yuchen, et al. Overcoming catastrophic forgetting by exemplar selection in task-oriented dialogue system[C]//62nd Annual Meeting of the Association for Computational Linguistics. Bangkok: ACL, 2024: 48. DOI: 10.18653/v1/2024.findings-acl.5

[8] Yu Longhui, Weng Zhenyu, Wang Yuqing, et al. Multi-teacher knowledge distillation for incremental implicitly-refined classification[C]//2022 IEEE International Conference on Multimedia and Expo, Taipei: IEEE, 2022. DOI: 10.1109/ICME52920.2022.9859936

[9] Kang Mengxue, Zhang Jinpeng, Zhang Jinming, et al. Alleviating catastrophic forgetting of incremental object detection via within-class and between-class knowledge distillation[C]//Proceedings of the IEEE International Conference on Computer Vision. Paris: IEEE, 2023: 18848. DOI: 10.1109/ICCV51070.2023.01732

[10] Nugraha B, Yadav K, Patil P, et al. Improving the detection of unknown DDoS attacks through continual learning[C]//3rd IEEE International Conference on Cyber Security and Resilience, CSR 2023. Hybrid: IEEE, 2023: 477. DOI: 10.1109/CSR57506.2023.10224989

[11] Jin Dong, Chen Shuangwu, He Huasen, et al. Federated incremental learning based evolvable intrusion detection system for zero-day attacks[J]. IEEE Network. 2023, 37(1): 125. DOI: 10.1109/MNET.018.2200349

[12] Zhao Qian, Yang Yang, Zhao Longjun, et al. Unbalanced data oversampling method for traffic multi-classification in convergence network[C]//11th International Conference on Computer Engineering and Networks, CENet2021. Hechi: Springer, 2022: 1549. DOI: 10.1007/978-981-16-6554-7_171

[13] Park C, Lee J, Kim Y, et al. An enhanced AI-based network intrusion detection system using generative adversarial networks[J]. IEEE Internet of Things Journal, 2023, 10(3): 2330. DOI: 10.1109/JIOT.2022.3211346

[14] Ding Hongwei, Sun Yu, Huang Nana, et al. TMG-GAN: generative adversarial networks-based imbalanced learning for network intrusion detection[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 1156. DOI: 10.1109/TIFS.2023.3331240

[15] 万建武，杨明. 代价敏感学习方法综述[J]. 软件学报，2020，31(1)：113
Wan Jianwu, Yang Ming. Survey on cost-sensitive learning method[J]. Journal of Software, 2020, 31(1): 113. DOI: 10.13328/j.cnki.jos.005871

[16] Al-Garadi M A, Mohamed A, Al-Ali A K, et al. A survey of machine and deep learning methods for internet of things (IoT) security[J]. IEEE Communications Surveys and Tutorials, 2020,

22(3): 1646. DOI:10.1109/COMST.2020.2988293

[17]廖丽云, 张伯雷, 吴礼发. 基于代价敏感学习的物联网异常检测模型[J]. 信息网络安全, 2023, 23(11): 94
Liao Liyun, Zhang Bolei, Wu Lifa. IoT anomaly detection model based on cost-sensitive learning[J]. Netinfo Security, 2023, 23(11): 94. DOI:10.3969/j.issn.1671-1122.2023.11.010

[18]Zhang Zhaoyi, Lai Yingxu, Chen Ye, et al. Fast tracing method for Sybil attack in VANETs[C]//98th IEEE Vehicular Technology Conference, VTC2023-Fall. Hong Kong: IEEE, 2023. DOI:10.1109/VTC2023-Fall60731.2023.10333449

[19]Xu Shixiong, Meng Gaofeng, Nie Xing, et al. Defying imbalanced forgetting in class incremental learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver: AAAI, 2024: 16211. DOI:10.1609/aaai.v38i14.29555

[20]Du Lei, Gu Zhaoquan, Wang Ye, et al. A few-shot class-incremental learning method for network intrusion detection[J]. IEEE Transactions on Network and Service Management, 2024, 21(2): 2389. DOI:10.1109/TNSM.2023.3332284

[21]Wang Tingting, Lv Qiujian, Hu Bo, et al. A few-shot class-incremental learning approach for intrusion detection[C]//30th International Conference on Computer Communications and Networks (ICCCN). Athens: IEEE, 2021. DOI: 10.1109/ICCCN52240.2021.9522260

[22]Jin Zhigang, Zhou Junyi, Li Bing, et al. FL-IIDS: a novel federated learning-based incremental intrusion detection system[J]. Future Generation Computer Systems, 2024, 151: 57. DOI:10.1016/j.future.2023.09.019

[23]武晓栋, 金志刚, 陈旭阳, 等. 对抗学习辅助增强的增量式入侵检测系统[J]. 哈尔滨工业大学学报, 2024, 56(9): 31
Wu Xiaodong, Jin Zhigang, Chen Xuyang, et al. Adversarial learning-augmented incremental intrusion detection system[J]. Journal of Harbin Institute of Technology, 2024, 56(9): 31. DOI:10.11918/202403066

[24]Odena A. Semi-supervised learning with generative adversarial networks[PP/OL]. V1. arXiv (2016-06-05)[2024-11-12]. https://doi.org/10.48550/arXiv.1606.01583

[25]Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[PP/OL]. V1. arXiv (2015-11-19)[2024-11-12]. https://doi.org/10.48550/arXiv.1511.06434

[26]Chen Hanting, Wang Yunhe, Xu Chang, et al. Data-free learning of student networks[C]//Proceedings of the IEEE International Conference on Computer Vision. Seoul: IEEE, 2019: 3513. DOI:10.1109/ICCV.2019.00361

[27]Vladimir V, Akashay V. A new learning paradigm: learning using privileged information[J]. Neural Networks, 2009, 22(5/6): 544. DOI:10.1016/j.neunet.2009.06.042

[28]Sharafaldin I, Lashkari A H, Ghorbani A A. Toward generating a new intrusion detection dataset and intrusion traffic characterization[C]//Proceedings of the 4th International Conference on Information Systems Security and Privacy. Funchal: SciTePress, 2018: 108. DOI:10.5220/0006639801080116

[29]Ferrag M A, Friha O, Hamouda D, et al. Edge-IIoTset: a new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning[J]. IEEE Access, 2022, 10: 40281. DOI: 10.1109/ACCESS.2022.3165809