DOI:10.11918/202403066

对抗学习辅助增强的增量式入侵检测系统

武晓栋,金志刚,陈旭阳,刘 凯

(天津大学 电气自动化与信息工程学院,天津 300072)

摘 要:为解决增量式入侵检测系统(intrusion detection system,IDS)在检测新类攻击过程中存在的新类过拟合、旧类泛化能力弱、灾难性遗忘问题,提出了对抗辅助增强的增量式IDS。在增量训练过程中,利用对抗样本的正则化能力约束检测模型在新类攻击上的过拟合,设计同时存储旧类攻击的原样本及对抗样本的双分布模拟缓存器以增强检测模型对旧类的泛化能力,引入加权交叉熵损失缓解灾难性遗忘问题。在CSE-CIC-IDS2018数据集和UNSW-NB15数据集上的实验结果表明:对抗样本直接参与训练会导致模型的识别性能恶化,而对抗样本以数据分布分离的形式参与训练则增强了模型的识别性能;对抗样本在缓存器中的存储有效抑制了模型对旧类泛化能力的丢失;加权交叉熵损失对学习权重的调整缓解了新类及缓存器内数据间的不平衡所导致的灾难性遗忘。所提方法为识别动态复杂网络环境中的真实攻击提供了可行方案,具有潜在的应用价值。

关键词:入侵检测;深度学习;增量学习;对抗学习;灾难性遗忘

中图分类号:TP393.08 **文献标志码**:A **文章编号**:0367-6234(2024)09-0031-07

Adversarial learning-augmented incremental intrusion detection system

WU Xiaodong, JIN Zhigang, CHEN Xuyang, LIU Kai

(School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China)

Abstract: To address the issues of overfitting on new classes, limited generalization ability on old classes, and catastrophic forgetting of incremental learning-based incremental intrusion detection system (IDS) when dealing with attacks of new classes, an adversarial assistance-augmented incremental IDS is proposed. In the incremental training, the regularization of adversarial samples is leveraged to mitigate the overfitting on new classes. A dual distribution simulation buffer that stores both clean and adversarial samples of old classes is proposed to enhance the generalization ability to old classes. In addition, weighted cross-entropy loss is introduced into the training process to alleviate the catastrophic forgetting. Experimental results on the CSE-CIC-IDS2018 dataset and the UNSW-NB15 dataset show that direct participation of adversarial samples in training leads to deterioration of the recognition performance, while participation in the form of detached data distribution enhances the recognition performance of the model. The storage of adversarial samples in the buffer effectively suppresses the loss of the model's generalization ability for old classes, and the adjustment of learning weights by weighted cross-entropy loss alleviates the catastrophic forgetting caused by the imbalance between the new classes and the data in the buffer. The proposed method offers a viable strategy for detecting real attacks within dynamic and complex networks, presenting substantial practical applicability.

Keywords: intrusion detection; deep learning; incremental learning; adversarial learning; catastrophic forgetting

网络中频繁的数据交互为恶意网络攻击提供了更多机会,攻击者利用各种手段[1]入侵网络以窃取数据甚至实现其破坏性目的。为应对这一问题,入侵检测系统(intrusion detection system,IDS)被广泛应用到网络攻击的检测中。其中,数据驱动式入侵检测方法充分挖掘网络流量中的特征信息,利用机器学习方法学习相关特征以构建检测模型,是目前研究的一个热点。主流的IDS基本采用离线训练模式[2-3],即在训练前获取所有训练样本,模型在训练结束后无法更新。但开放动态的现实网络环境中,数据流具备时变特性,新类别攻击层出不穷[4],离线训练的IDS难以有效应对。重新训练IDS是一种有效解决措施,但可能存在两方面阻碍:一是历史数据量过大导致训练成本激增;二是历史数据因隐私问题而不再可用。此外,利用新类别样本微调IDS模型也可检测新类别攻击。但样本量增大会引发模

收稿日期:2024-03-28;**录用日期**:2024-04-19;**网络首发日期**:2024-06-17
网络首发地址:https://link.cnki.net/urlid/23.1235.t.20240615.2327.002
基金项目:国家自然科学基金(52171337)
作者简介:武晓栋(1996—),男,博士研究生;金志刚(1972—),男,教授,博士生导师
通信作者:金志刚,zgjin@tju.edu.cn

型对旧类知识的遗忘,即灾难性遗忘会导致 IDS 丧失对旧类的识别能力。

增量学习(incremental learning,IL)中,模型在学习新类的同时能够保留旧类知识[5]。基于增量学习的 IDS 能够有效应对上述问题。付子燨等[6]提出了支持向量机与 K 最近邻(k-nearest neighbor,KNN)相结合的增量式入侵检测方法;刘强等[7]提出基于自组织增量神经网络的物联网增量式入侵检测系统。考虑到入侵检测中攻击的不确定性,本文研究更为困难的类增量学习[8]场景。类增量学习相关方法主要分为两类:一是存储式方法,即将旧类样本存储至缓存器;二是生成式方法[9],即在训练中生成旧类样本。入侵检测任务中,生成式方法可能丢失数据特征间的隐含约束,更适合作为一种存储式方法的辅助手段。存储式方法需要随新类数据的到来不断增大存储空间,这将使增量学习研究降级为模型重新训练。因此,缓存器的存储量应固定,随新类到来而变化的应是缓存器中各类的占比。但固定缓存器的大小会导致本就倾向于拟合新类的增量学习加重对新类的过拟合。增强模型对旧类的泛化能力,同时缓解新类上的过拟合问题成为研究的关键。

对抗学习(adversarial learning,AL)[10]常作为一种以识别准确率换取鲁棒性的方法[11-12],但 Xie 等[13]所提出的 AdvProp(adversarial propagation)成功利用对抗样本提升了卷积神经网络(convolutional neural networks,CNNs)上的识别性能。Chen 等[14]进一步改进了 AdvProp,并验证了改进方法在深度神经网络(deep neural networks,DNNs)上的性能提升。换言之,对抗样本具有提升增量 IDS 识别能力的可能性。

基于此,为保持模型对旧类的泛化能力并缓解新类上的过拟合及灾难性遗忘问题,本文提出了对抗辅助增强的增量式 IDS(adversarial assistance enhanced incremental IDS,AAE-IIDS)。在采用对抗样本对原样本施加正则化约束的基础上,将双分布模拟缓存器和加权交叉熵损失引入检测系统。通过仿真实验,验证对抗样本对模型的识别性能、双分布模拟缓存器和加权交叉熵损失的效用。最后,在不同的神经网络模型下,将所提方法与最佳基线模型进行对比,以探究所提方法的适用性和有效性。

1 对抗样本生成及效用分析

1.1 对抗样本生成

在第 t 个增量任务中,给定原样本 x^t 及其编码后标签 y^t,S 为围绕 x^t 的 L_p 球空间内的一系列对抗扰动。利用投影梯度下降法(project gradient descent,PGD)[15-16]产生对抗攻击的方法表达式为

$$x^{t_{k+1}} = \Pi_{x^t+S}\{x^{t_k} + \sigma \mathrm{sgn}[\nabla_x L(\Theta^t, x^t, y^t)]\} \quad (1)$$

式中:$\Pi\{\cdot\}$ 为限制函数[17],可将扰动限制到 S;$L(\cdot)$ 为损失函数;Θ^t 为当前增量任务的分类模型;σ 为步长;x^{t_k} 为当前任务进行第 k 次迭代前所用的样本;$x^{t_{k+1}}$ 为当前任务进行第 k 次迭代后所得到的样本。在若干次迭代后,获得训练所需要的对抗样本,迭代停止。

1.2 辅助性对抗样本效用分析

研究表明,直接混合对抗样本与原样本训练神经网络会导致识别性能下降[18-19]。受 AdvProp 启发,本文分离训练两类样本,从而有效利用对抗样本的异性数据分布增强神经网络的表征能力。批量归一化(batch normalization, BN)极为适合分离训练这一任务。BN 利用每个 mini-batch 计算得到的均值和方差对输入进行标准化,并对输入进行缩放和移动以保持网络的非线性表达能力。这一方法成功解决了内部协变量偏移问题,被广泛应用于有关神经网络的研究中[20-21]。BN 以 mini-batch 为处理单位,只需确保同一 mini-batch 中输入的样本类别相同即可确保神经网络中的样本分离输入。

BN 的引入与否对训练数据分布的影响见图 1。横轴 X 表示数据,(μ_A,σ_A) 表示对抗样本的正态数据分布,(μ_C,σ_C) 表示原样本的正态数据分布,(μ_{A+C},σ_{A+C}) 表示混合数据分布。直接将对抗样本和原样本混合送入同一个 BN 中会导致整体训练数据分布完全异于原样本数据分布,这将导致神经网络识别性能下降。因此,除原样本 BN 外,引入辅助 BN 可以利用对抗样本对原样本施加正则化约束,能防止过拟合,从而提升神经网络的整体性能。

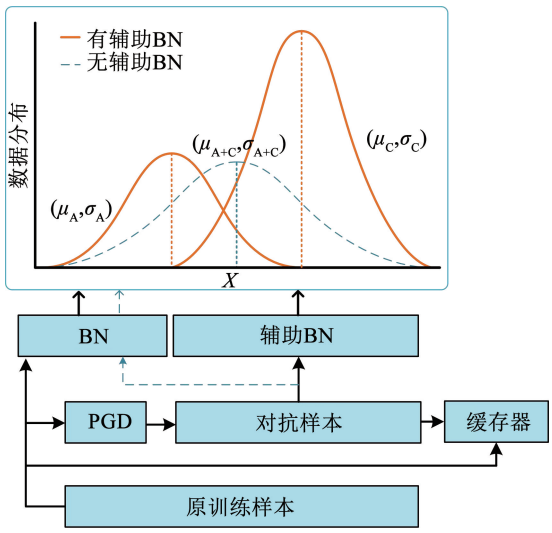

图 1 辅助 BN 影响下的不同数据分布

Fig. 1 Different data distribution on the effect of auxiliary BNs

2　双分布模拟缓存器

类增量学习赋予模型随入侵变化动态更新的检测能力，其中存储式方法将旧类存储起来以维持整体检测性能。在历史增量训练中，旧类对抗样本被用以增强对旧类原样本的学习，存储旧类对抗样本能够维持其所带来的识别性能增益。此外，在缓存器大小固定时，平均存储不同类型的数据会导致缓存器中的数据分布异于实际数据分布。因此，本文提出了双分布模拟缓存器。该缓存器动态调整其中原样本及对抗样本的占比。考虑到过大的缓存不符合实际应用，双分布模拟缓存器的存储大小保持固定不变。在每个增量任务结束时，存储当前任务中新类的原样本及对抗样本，并根据当前任务数据集在总历史任务数据集中的比例调整旧类别样本的比例。具体而言，假设每个历史增量任务中有 m 类数据，任务 T^t 中的旧类 C_j 的原样本大小为 S_j^t，那么在第 n 个增量任务 T^n 中各旧类在缓存器中的占比 m_j^n 以及对应对抗样本占比 $m_{j-\mathrm{adv}}^n$ 表示为

$$m_j^n = m_{j-\mathrm{adv}}^n = \frac{S_j^t}{2\sum_{t=1}^{n-1}\sum_{j=1}^{m} S_j^t} \tag{2}$$

3　加权交叉熵损失

为进一步清晰阐述增量过程中的加权交叉熵损失，将整个增量任务表示为 $\{T^t\}_{t=1}^N$，其中 N 表示任务总数；将第 t 个增量任务中由 b 对样本 x_i^t 以及其独热编码标签 y_i^t 所构成的一个 mini-batch 表示为 $\{X_b^t, Y_b^t\} = \{x_i^t, y_i^t\}_{i=1}^b$。当然，此处不区分对抗样本抑或原样本。给定当前任务缓存器 M 以及新类训练数据 D^t，此 mini-batch 的分类损失 L_{CE} 为

$$L_{\mathrm{CE}} = \frac{1}{b}\sum_{i=1}^{b} H[P^t(x_i^t, \boldsymbol{\Theta}^t), y_i^t] \tag{3}$$

式中：$H[\cdot]$ 为交叉熵损失，$P^t(\cdot)$ 为 $\boldsymbol{\Theta}^t$ 预测得到的概率分布。

L_{CE} 无法将 D^t 与 M 中数据量大小的不平衡纳入考量范围，因此，本文进一步采用了加权交叉熵损失函数[22]应对这一问题。简而言之，记样本 (x_i^t, y_i^t) 的标签值为 l，调节样本 (x_i^t, y_i^t) 的梯度测量 G_i^t 在其标签所对应输出神经元 N_l^t 上的大小以增强真实数据标签 y_i^t 与对应输出神经元 N_l^t 间的对应性，而其他输出神经元 $\{N_l^t\}_{j\neq l}^m$（m 为数据类别数）上的梯度测量保持不变。具体而言，第 l 个输出神经元 N_l^t 的梯度测量 G_i^t 为

$$G_i^t = \frac{\partial H[P^t(x_i^t, \boldsymbol{\Theta}^t), y_i^t]}{\partial N_l^t} = P^t(x_i^t, \boldsymbol{\Theta}^t)_l - 1 \tag{4}$$

式中 $P^t(x_i^t, \boldsymbol{\Theta}^t)_l$ 为样本 (x_i^t, y_i^t) 的第 l 个 Softmax 概率。由此，对 $\{x_i^t, y_i^t\}_{i=1}^b$，新类和旧类的权重定义为

$$G_{\mathrm{new}} = \frac{\sum_{i=1}^{b} | G_i^t | \varphi_{y_i^t \in \Omega^t}}{\sum_{i=1}^{b} \varphi_{y_i^t \in \Omega^t}} \tag{5}$$

$$G_{\mathrm{old}} = \frac{\sum_{i=1}^{b} | G_i^t | \varphi_{y_i^t \in \cup_{k=1}^{t-1}\Omega^k}}{\sum_{i=1}^{b} \varphi_{y_i^t \in \cup_{k=1}^{t-1}\Omega^k}} \tag{6}$$

式中：Ω^t 为新类的标签空间；$\cup_{k=1}^{t-1}\Omega^k$ 为旧类的标签空间；φ 为条件函数，当下标条件为真时其值为 1，为假则为 0。如此一来，模型对旧类与新类的学习速度产生了区别。样本越多的类的学习权重越小，样本越少的类的学习权重越大，这能够有效缓解已学习旧类知识的遗忘问题。加权交叉熵损失定义为

$$L_{\mathrm{WE}} = \frac{1}{b}\sum_{i=1}^{b} \frac{| G_i^t |}{\bar{G}_i} H[P^t(x_i^t, \boldsymbol{\Theta}^t), y_i^t] \tag{7}$$

式中 $\bar{G}_i = \varphi_{y_i^t \in \Omega^t} G_{\mathrm{new}} + \varphi_{y_i^t \in \cup_{k=1}^{t-1}\Omega^k} G_{\mathrm{old}}$。加权交叉熵损失下，随着新类的到来，旧类别攻击的训练权重不断增加，旧类的灾难性遗忘问题得到有效缓解。

4　对抗辅助增强的增量式 IDS

AAE-IIDS 的整体训练流程见图 2。为清晰表示不同任务间的数据走向，图 2 中主要展示了不同任务的模型情况。在每个增量任务中，对抗样本通过 PGD 在原样本的基础上生成，并随原样本一起输入模型。增量过程中，双分布模拟缓存器动态调整其内部样本占比以适应变化的网络环境，并同时维持旧类上的正则化约束。

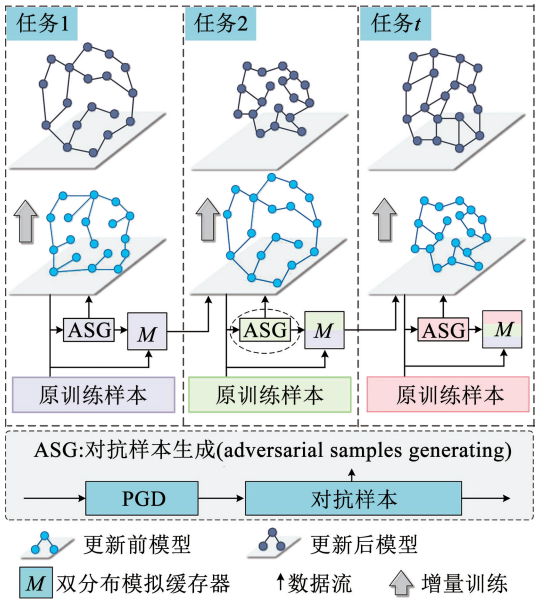

图 2　AAE-IIDS 训练流程

Fig. 2　Training process of AAE-IIDS

为清晰表述整个流程,本文给出 AAE-IIDS 的伪代码并做相关解释说明。首先,增量任务中,将原样本与所生成对抗样本一同输入神经网络;然后,利用总损失优化整个神经网络的参数。需要注意的是,在测试过程中只允许主 BN 起作用。伪代码如下:

输入:入侵数据集$\{X^t,Y^t\}_{i=1}^t$,对抗扰动 S,步长 σ
输出:增量学习模型 Θ^t
For $\{X^i,Y^i\}$ in $\{X^t,Y^t\}_{i=1}^t$ **do**
For 原 mini-batch in $\{X^i,Y^i\}$ **do**
使用 PGD(参数 S,σ)产生对抗样本
End for
Repeat
使用 G_{new} 重加权原样本 mini-batch 损失
使用 G_{new} 重加权对抗样本 mini-batch 损失
使用 G_{old} 重加权缓存器内样本损失
最小化总损失 L_{WE} 以更新网络参数
Until 当前任务训练结束
根据式(2)调整缓存器中数据分布
End for
Return 增量学习模型 Θ^t

5 仿 真

5.1 仿真数据集

本文使用目前被广泛应用于入侵检测研究的 CSE-CIC-IDS2018[23] 和 UNSW-NB15 数据集。对于 CSE-CIC-IDS2018 数据集,本仿真将原攻击类型归纳为 7 个大类[24-25]。考虑到 CSE-CIC-IDS2018 数据集的数据量比较庞大,在每类中随机抽取 1% 的样本进行增量测试。由于 Web 类样本量过小,仿真中并不使用。训练数据集与测试数据集的切分比例为 4∶1。采用官网公开的 UNSW-NB15-training 数据集和 UNSW-NB15-testing 数据集,因此不需要进一步划分训练集和测试集。此外,考虑到类别平衡,本仿真不使用数据集中标记为 worm 的攻击流量。本文主要在 CSE-CIC-IDS2018 数据集上对所提方法各模块机理进行消融分析,在 CSE-CIC-IDS2018 和 UNSW-NB15 数据集上对所提方法的适用性和有效性进行研究。针对本文类增量任务的具体数据集分布见表 1、表 2。

5.2 仿真环境与评价指标

考虑到类增量任务对类间差异的敏感性,所有仿真场景将限定于入侵的多重分类,而非二分类。使用 GeForce RTX 3090 GPU 在 PyTorch 1.12.0 环境进行仿真。双分布模拟缓存器固定存储 8 000 个样本。所有仿真的主要参数如下:batch size 设为 32,步长设为 1,local epoch 设为 5,优化器使用 SGD。

表 1 CSE-CIC-IDS2018 数据集样本分布

Tab. 1 Sample distribution of CSE-CIC-IDS2018 dataset

任务阶段	类别	训练集样本数/条	测试集样本数/条
任务 1	Normal	48 898	12 224
	DDos	5 502	1 375
	Infiltration	1 295	324
任务 2	Bot	2 290	572
	Brute Force	3 047	762
	DoS	5 234	1 309

表 2 UNSW-NB15 数据集样本分布

Tab. 2 Sample distribution of UNSW-NB15 dataset

任务阶段	类别	训练集样本数/条	测试集样本数/条
任务 1	Normal	56 000	36 988
	Fuzzers	18 184	6 062
	Backdoor	1 746	583
任务 2	Dos	12 264	4 089
	Exploits	33 393	11 132
	Analysis	2 000	677
任务 3	Generic	40 000	18 871
	Reconnaissance	10 491	3 496
	Shellcode	1 133	378

将预测结果正确的正常样本数记为 T_N(真阴性),预测结果正确的入侵样本数记为 T_P(真阳性),预测结果错误的正常样本数记为 F_P(假阳性),预测结果错误的入侵样本数记为 F_N(假阴性)。在此基础上,使用分类准确率(A)、精确率(D)和召回率(R)作为评价指标。A 能够体现模型整体分类能力。D 有效反映了入侵的识别准确率。R 有效反映了模型的入侵识别能力。评价指标的计算式如下:

$$A=\frac{T_P+T_N}{T_P+T_N+F_P+F_N} \tag{8}$$

$$D=\frac{T_P}{T_P+F_P} \tag{9}$$

$$R=\frac{T_P}{T_P+F_N} \tag{10}$$

5.3 对抗样本效用分析

为全面研究对抗样本在 AAE-IIDS 中的效用,本仿真同时在 DNN、CNN 和 RNN 上进行测试。在原样本参与的基础上,模型训练分别在以下 3 种条件下进行:1)无任何对抗样本(基线);2)有对抗样本但无辅助 BN(混合 BN);3)有对抗样本和辅助

BN(所提方法)。对应结果见表 3。

表 3　对抗样本效用多模型验证结果

Tab. 3　Effect of adversarial samples on multiple networks

测试网络	类别	*A*/%	
		任务 1	任务 2
DNN	基线	99.94	92.36
	混合 BN	99.93	84.45
	所提方法	99.96	93.06
CNN	基线	97.65	70.34
	混合 BN	91.19	55.66
	所提方法	97.82	77.38
RNN	基线	99.95	90.38
	混合 BN	99.83	80.21
	所提方法	99.96	92.50

由表 3 可知，直接混合对抗样本与原样本以训练检测模型将带来灾难性后果，与多数对抗学习相关研究相吻合。在 CNN 上，任务 2 混合 BN 条件所得准确率比基线所得降低了 14.68%。而与之相对应，所提方法的增强效果在任务 2 阶段极为显著。与基线相比，所提方法在 CNN 上取得了最显著的效果，提高了 7.04%。增量训练效果最好的是 DNN，从任务 1 到任务 2 只衰减了 6.90%。

为深入分析所提方法在不同类别上的检测能力，以热力图的形式展示了基线混淆矩阵与所提方法的混淆矩阵之间的成对差异，见图 3。总体而言，所提方法与基线的差异多体现在对旧类的记忆能力方面。在 DNN 和 CNN 上，所提方法与基线的主要区别体现在 DDoS 攻击的识别方面。部分 DDoS 攻击被基线错误识别为 Bot 攻击。Bot 攻击和 DDoS 攻击相同的分布式特性可能干扰了基线的识别准确性。缓存器中所存储的旧类对抗样本则辅助 DNN 和 CNN 有效地区分了这两种攻击。在 RNN 模型上，所提方法纠正了基线对部分正常样本和 Infiltration 攻击的错误识别。

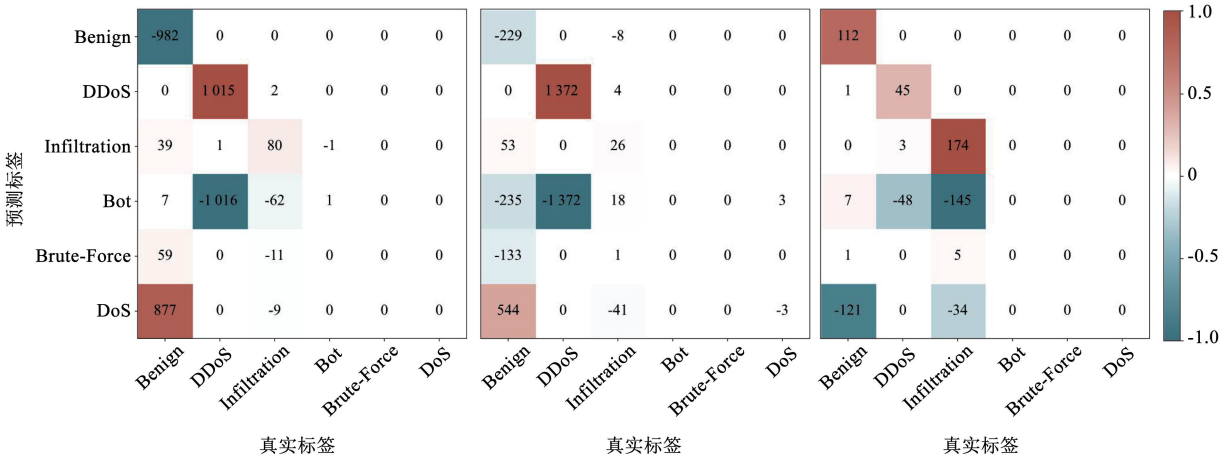

图 3　对抗样本变化下成对差异热力图

Fig. 3　Pairwise differential heatmap under adversarial smamples variation

5.4　双分布模拟缓存器仿真

为分析双分布模拟缓存器在 AAE-IIDS 中的具体效用，在 DNN、CNN、RNN 上开展仿真。数据重放是增量学习的前提，因此并不开展缓存器消融仿真。本仿真设置两个可变条件：一是缓存器中对抗样本是否存在；二是缓存器中各类的数量是否可动态调整(不可动态调整则平均存储所有类)。由此产生了 4 种相应仿真结果，见图 4。

由图 4 可知，DNN 上单一条件的消融都导致了准确率急剧下降。当消融掉所有条件时，准确率降至 53.08%。纵观整个实验结果，同一任务中单一条件消融都导致准确率下降，而准确率在消融掉所有条件时降至最低。当然也存在异常情况。在缓存器不可动态调整的条件下，RNN 上旧类对抗样本的

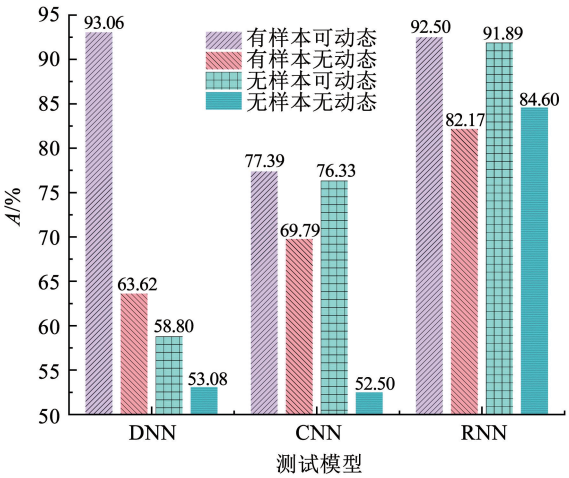

图 4　不同模型上缓存器条件变化的影响

Fig. 4　Effects due to memory conditional ablation on different models

加入使得准确率小幅下降。造成这种异常现象的可能原因如下:均匀存储由样本量较小的旧类所生成的对抗样本对旧类的原样本施加了过强的正则化约束,辅助 BN 难以完全缓冲掉这一约束,最终模型整体检测能力衰退。总之,动态调整缓存器极为必要,而且在动态缓存器中加入对抗样本能够提高模型的识别性能。

5.5 损失函数仿真

为测试不同的损失函数设置对 AAE-IIDS 性能的影响,仿真在 RNN 上测试了 3 种情况下的模型性能:1)仅使用交叉熵损失 L_{CE};2)仅使用加权交叉熵损失 L_{WE};3)使用混合损失($0.5L_{CE}+0.5L_{WE}$)。

不同损失函数对 AAE-IIDS 性能的影响结果见图 5。任务 1 中无旧类,因此 3 种情况下准确率完全一致,这符合加权交叉熵损失定义。在增量过程中,与使用交叉熵损失相比,加权交叉熵损失的检测准确率提升了 1.33%。结合混合损失的实验结果可知,加权交叉熵损失的性能增益与加权交叉熵损失的应用比例呈正相关。任务 2 中,由交叉熵损失到混合损失准确率提升了 1.02%,由混合损失到加权交叉熵损失提升了 0.31%。可见,加权交叉熵损失的应用程度愈深,模型提升效果愈明显。

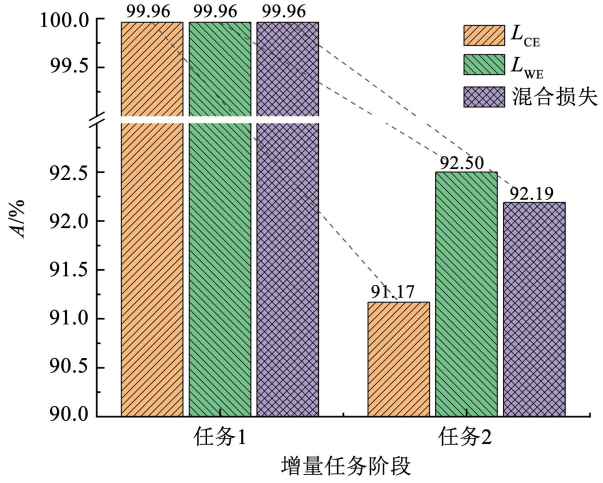

图 5 不同损失函数对模型性能的影响

Fig. 5 Impact of different losses on model performance

图 6 进一步展示了交叉熵损失与加权交叉熵损失之间混淆矩阵的成对差异热力图。由图 6 可知,加权交叉熵损失与交叉熵损失之间的差异主要体现在对旧类的识别能力方面。加权交叉熵损失在训练过程中增加了旧类的权重。被交叉熵损失错误判定为 DoS 攻击的 63 个样本中有 59 个被加权交叉熵损失正确识别为 Normal 样本。被判定为 Bot 攻击的 185 个样本中有 174 个被正确识别为 Infiltration 攻击。该纠正过程并未导致对新类别的错误识别,仿真现象完全符合理论分析结果。

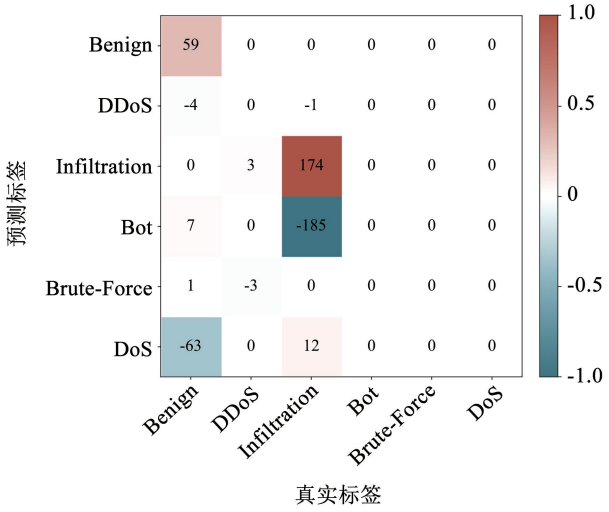

图 6 不同损失函数下成对差异热力图

Fig. 6 Pairwise differential heatmaps under different loss functions

5.6 不同模型效果对比

为探究所提方法的适用性和有效性,仿真构建了 DNN、CNN、RNN 和 CNN-GRU 网络,并分别在 CSE-CIC-IDS2018 和 UNSW-NB15 数据集上与最佳基线模型进行对比。对应结果见表 4、表 5。

在 CSE-CIC-IDS2018 数据集上,所提方法在 DNN 上取得了最佳结果,所有指标均高于基线。在 UNSW-NB15 数据集上,所提方法在 DNN 模型上取得了整体最佳结果。综合两个数据集仿真结果,所提方法在 DNN 上获得了较高的精确率和召回率,这意味着 DNN 对不同攻击的识别比其他网络更为准确。实际上,多分类任务中获取高精确率及高召回率本身较为困难,而类增量学习场景则进一步提高了识别难度。在类增量任务中,模型学习样本量较小的类的难度极大。而一旦模型无法识别这些

表 4 CSE-CIC-IDS2018 数据集 AAE-IIDS 多模型测试结果

Tab. 4 Multi-model testing results of AAE-IIDS on CSE-CIC-IDS2018 dataset

模型	任务阶段	指标		
		D/%	*R*/%	*A*/%
最佳基线	任务 1	99.61	99.41	99.94
	任务 2	86.97	79.59	92.37
AAE-IIDS(DNN)	任务 1	99.64	99.72	99.96
	任务 2	87.51	96.70	93.06
AAE-IIDS(CNN)	任务 1	85.38	70.23	97.83
	任务 2	71.30	79.90	77.38
AAE-IIDS(RNN)	任务 1	99.56	99.62	99.96
	任务 2	88.68	90.52	92.50
AAE-IIDS(CNN-GRU)	任务 1	95.92	76.23	98.28
	任务 2	59.21	73.02	84.30

表 5　UNSW-NB15 数据集 AAE-IIDS 多模型测试结果

Tab. 5　Multi-model testing results of AAE-IIDS on UNSW-NB15 dataset

模型	任务阶段	指标		
		D/%	*R*/%	*A*/%
最佳基线	任务 1	47.96	46.79	80.40
	任务 2	30.00	29.42	52.26
	任务 3	30.13	28.48	51.02
AAE-IIDS(DNN)	任务 1	51.95	49.59	81.04
	任务 2	30.31	29.51	55.12
	任务 3	30.12	31.45	51.23
AAE-IIDS(CNN)	任务 1	45.88	34.43	84.96
	任务 2	23.13	21.08	51.44
	任务 3	21.09	23.91	41.59
AAE-IIDS(RNN)	任务 1	42.18	43.08	82.06
	任务 2	23.52	25.84	52.76
	任务 3	18.23	26.75	51.59
AAE-IIDS(CNN-GRU)	任务 1	42.72	52.65	70.71
	任务 2	23.74	25.15	50.02
	任务 3	18.82	26.59	47.14

类，通过平均所有类别的精确率和召回率所得出的总精确率和召回率则急剧下降。这正是本仿真部分模型的精确率和召回率远低于相应准确率的原因。总体而言，所提方法通过提高对新类别的泛化能力及减轻灾难性遗忘问题，有效地增强了增量式 IDS 识别入侵的能力。

6　结　论

本文利用对抗学习增强检测模型的泛化能力以应对真实网络环境中不断到来的新攻击，结论如下：

1）对抗样本直接混合原样本进行训练导致检测模型性能下降，而以数据分布上隔离的形式参与训练则有效增强了模型对原样本的拟合能力，缓解了新类上的过拟合问题。

2）双分布模拟缓存器中对旧类对抗样本的存储成功保持了对旧类原样本的正则化约束，缓存器内样本的动态调整提高了模型对旧类的泛化能力。

3）加权交叉熵损失有效缓解了增量任务中新旧类样本不平衡所引发的灾难性遗忘问题，在不影响新类的识别能力的同时增强了对旧类的识别能力。

4）对抗样本能够提升入侵检测模型在 CNN、DNN、RNN 上的识别能力。

参考文献

[1] MOUSTAFA N, KORONIOTIS N, KESHK M, et al. Explainable intrusion detection for cyber defences in the internet of things: opportunities and solutions[J]. IEEE Communications Surveys & Tutorials, 2023, 25(3): 1775. DOI:10.1109/COMST.2023.3280465

[2] THAKKAR A, LOHIYA R. Fusion of statistical importance for feature selection in deep neural network-based intrusion detection system[J]. Information Fusion, 2023, 90: 353. DOI:10.1016/j.inffus.2022.09.026

[3] LOUK M H L, TAMA B A. Dual-IDS: a bagging-based gradient boosting decision tree model for network anomaly intrusion detection system[J]. Expert Systems with Applications, 2023, 213: 119030. DOI:10.1016/j.eswa.2022.119030

[4] PUTINA A, ROSSI D. Online anomaly detection leveraging stream-based clustering and real-time telemetry[J]. IEEE Transactions on Network and Service Management, 2021, 18(1): 839. DOI:10.1109/TNSM.2020.3037019

[5] WU Zhijun, GAO Pan, CUI Lei, et al. An incremental learning method based on dynamic ensemble RVM for intrusion detection[J]. IEEE Transactions on Network and Service Management, 2022, 19(1): 671. DOI:10.1109/TNSM.2021.3102388

[6] 付子燨，徐洋，吴招娣，等. 基于增量学习的 SVM-KNN 网络入侵检测方法[J]. 计算机工程，2020，46(4)：115
FU Zixi, XU Yang, WU Zhaodi, et al. SVM-KNN network intrusion detection method based on incremental learning[J]. Computer Engineering, 2020, 46(4): 115. DOI:10.19678/j.issn.1000-3428.0054701

[7] 刘强，张颖，周卫祥，等. 自适应类增量学习的物联网入侵检测系统[J]. 计算机工程，2023，49(2)：169
LIU Qiang, ZHANG Ying, ZHOU Weixiang, et al. Adaptive class incremental learning-based IoT intrusion detection system[J]. Computer Engineering, 2023, 49(2): 169. DOI:10.19678/j.issn.1000-3428.0063917

[8] 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

[9] LIN Huiwei, FENG Shanshan, LI Xutao, et al. Anchor assisted experience replay for online class-incremental learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(5): 2217. DOI:10.1109/TCSVT.2022.3219605

[10] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]// 3rd International Conference on Learning Representations. San Diego: ICLR, 2015

[11] HAN Sicong, LIN Chenhao, SHEN Chao, et al. Interpreting adversarial examples in deep learning: a review[J]. ACM Computing Surveys, 2023, 55(14): 1. DOI:10.1145/3594869

[12] HO C H, NVASCONCELOS N. Contrastive learning with adversarial examples[C]//34th Conference on Neural Information Processing Systems (NeurIPS 2020). Vancouver: Neural Information Processing Systems, 2020: 17081

[13] XIE Cihang, TAN Mingxing, GONG Boqing, et al. Adversarial examples improve image recognition[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Seattle: IEEE, 2020: 819. DOI:10.1109/CVPR42600.2020.00090

[14] CHEN Xiangning, XIE Cihang, TAN Mingxing, et al. Robust and accurate object detection via adversarial learning[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021: 16622. DOI: 10.1109/CVPR46437.2021.01635

(下转第 84 页)