计算机科学
COMPUTER SCIENCE

基于特征差分选择的集成模型流量对抗样本防御架构

何元康, 马海龙, 胡涛, 江逸茗

引用本文

何元康, 马海龙, 胡涛, 江逸茗. 基于特征差分选择的集成模型流量对抗样本防御架构[J]. 计算机科学, 2025, 52(4): 369-380.

HE Yuankang, MA Hailong, HU Tao, JIANG Yiming. Defense Architecture for Adversarial Examples of Ensemble Model Traffic Based on FeatureDifference Selection [J]. Computer Science, 2025, 52(4): 369-380.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240200092

基于特征差分选择的集成模型流量对抗样本防御架构

何元康[1]　马海龙[1,2]　胡　涛[1]　江逸茗[1,2]
1 解放军战略支援部队信息工程大学　郑州 450000
2 网络空间安全教育部重点实验室　郑州 450000
(yuankang_he@163.com)

摘　要　当前,基于深度学习的异常流量检测模型容易遭受流量对抗样本攻击。作为防御对抗攻击的有效方法,对抗训练虽然提升了模型鲁棒性,但也导致了模型检测精度下降。因此,如何有效平衡模型检测性能和鲁棒性是当前学术界研究的热点问题。针对该问题,基于集成学习思想构建多模型对抗防御框架,通过结合主动性特征差分选择和被动性对抗训练,来提升模型的对抗鲁棒性和检测性能。该框架由特征差分选择模块、检测体集成模块和投票裁决模块组成,用于解决单检测模型无法平衡检测性能与鲁棒性、防御滞后的问题。在模型训练方面,设计了基于特征差分选择的训练数据构造方法,通过有差异性地选择和组合流量特征,形成差异化流量样本数据,用于训练多个异构检测模型,以抵御单模型对抗攻击;在模型裁决方面,对多模型检测结果进行裁决输出,基于改进的启发式种群算法优化集成模型裁决策略,在提升检测精度的同时,增大了对抗样本生成的难度。实验效果显示,所提方法的性能相比单个模型对抗训练有较大提升,相较于现有的集成防御方法,其准确率和鲁棒性提升了近10%。

关键词:异常流量检测;对抗样本攻击;集成学习;多模裁决
中图分类号　TP309

Defense Architecture for Adversarial Examples of Ensemble Model Traffic Based on Feature Difference Selection

HE Yuankang[1],MA Hailong[1,2],HU Tao[1] and JIANG Yiming[1,2]
1 PLA Strategic Support Force Information Engineering University,Zhengzhou 450000,China
2 Key Laboratory of Cyberspace Security Ministry of Education,Zhengzhou 450000,China

Abstract　Currently,anomaly traffic detection models that leverage deep learning technologies are increasingly vulnerable to adversarial example attacks. Adversarial training has emerged as a potent defense mechanism against these adversarial attacks. By incorporating adversarial examples into the training process,it aims to enhance the model's robustness,making it more resistant to similar attacks in the future. However,this approach is not without its drawbacks. While it indeed increases the model's robustness,it also inadvertently leads to a decrease in the model's detection accuracy. This trade-off between robustness and accuracy has become a pivotal concern in the realm of deep learning-based anomaly detection,sparking intense debate and research within the academic community. Addressing this critical issue,this paper proposes a novel framework that seeks to balance the model's detection performance with its robustness against adversarial attacks. Drawing inspiration from ensemble learning,we construct a multi-model adversarial defense framework. This framework not only enhances the model's adversarial robustness but also aims to improve its detection performance. By integrating proactive feature differential selection with passive adversarial training,we develop a comprehensive strategy that fortifies the model against adversarial threats while maintaining high detection accuracy. The model consists of a feature differential selection module,a detection body integration module,and a voting decision module,to address the issue that a single detection model cannot balance detection performance and robustness,and the problem of defense lagging. In the aspect of model training,we introduce a sophisticated method for constructing training data based on feature differential selection. This method involves selectively combining traffic features that exhibit significant differences,thereby creating a set of differentiated traffic example data. These examples are then used to train multiple heterogeneous detection models. This approach is designed to bolster the models' resistance to adversarial attacks targeted at single models,presenting a more formidable challenge to attackers. Furthermore,the framework includes a novel adjudication mechanism for the detection results produced by the multiple models. Leveraging an improved heuristic population algorithm,we optimize the ensemble model's

到稿日期:2024-02-26　返修日期:2024-07-25
基金项目:雄安新区科技创新专项(2022XAGG0111)
This work was supported by the Xiong'an New Area Science and Technology Innovation Special Project(2022XAGG0111).
通信作者:马海龙(longmanclear@163.com)

adjudication strategy. This not only enhances the detection accuracy but also significantly increases the complexity and difficulty of generating effective adversarial examples, thereby providing an additional layer of defense. Experimental results underscore the efficacy of the proposed method. Compared to traditional single-model adversarial training approaches, the multi-model framework demonstrates a substantial improvement, with nearly a 10% increase in both accuracy and robustness.

Keywords Abnormal traffic detection, Adversarial example attack, Integrated learning, Multimode adjudication

1 引言

网络入侵检测系统(Network Intrusion Detection System, NIDS)作为不可或缺的网络防御技术之一[1],可以从流通网络的网络通信和数据包报头信息中导出输入特征,据此将流量分类为良性或是如拒绝服务(DDoS)、僵尸网络(Bot)之类的恶意网络攻击[2],以达到网络攻击态势感知的效果。

NIDS 作为能够检测网络攻击的技术,在与机器学习技术结合后大幅提升了检测能力,并且可以规模化地监测流量数据,对异常流量检测产生了深远影响。但由于机器学习技术面临模型鲁棒性问题,因此基于机器学习的 NIDS 也面临着对抗样本攻击的威胁。传统图像[3]和语音[4]对抗攻击能够对模型进行微小改变从而导致人工智能系统错误分类,但对于人类来说,模型的整体输入看起来并没有变化。流量对抗样本攻击与之相同,其根据对抗生成算法计算出特征中最小的扰动量,使得模型认为添加扰动的恶意流量是正常数据,从而导致 NIDS 将流量错误分类。

传统图像和语音对抗样本生成方法这两个领域的对抗样本生成是根据数据提取后的特征生成的,而流量对抗样本生成还需要保证流量的恶意性和可使用性[5-7],攻击者应确保修改后的恶意流量不会破坏通信协议规则,即该流量在性质上仍属于恶意流量并且能成功被接收方系统所接收。因此,常用于流量对抗样本生成的方法是在流量的样本空间上直接生成。故传统防御方法无法有效应用于流量对抗样本领域,流量对抗样本防御问题亟待研究。

目前,防御方法主要从增强检测模型鲁棒性和增加对抗样本生成难度这两种思路分别展开,其中以对抗样本训练方法为代表的增强鲁棒性思路可以部分防御黑盒攻击,该方法可增强检测模型对于对抗样本的识别泛化能力[2,8]。对抗训练的缺点则包括:1)防御者需要在模型鲁棒性和检测性能之间做一定的权衡,对抗训练会导致模型对原始数据的识别准确率降低;2)防御被动性,单个检测器无法做到对对抗样本空间的全覆盖[9],攻击可以随时变化,但防御需要根据攻击来进行改变,因此无法完全识别所有对抗样本。集成模型则是增加对抗样本生成难度的有效方法之一,但由于对抗样本具有可迁移性,同一对抗样本可使多个检测模型发生错误分类,因此集成学习的对抗防御能力增强效果也不佳。

为了解决该问题,本文提出了基于特征差分选择的集成模型(Ensemble Model Traffic Based on Feature Difference Selection, EFDS)对抗防御架构,结合集成学习的思想构建了多异构模型学习裁决框架,用于同时提升模型的对抗鲁棒性和检测性能。EFDS 由特征差分选择模块、检测体集成模块和投票裁决模块组成。特征差分选择模块和投票裁决模块的设计增加了对抗样本的生成难度,检测体集成模块则增强了模型鲁棒性,达到了两种防御思路结合的效果。

其中,在特征提取阶段设置了特征差分选择提取器,将输入的流量包头信息进行提取并转化为多维检测特征,对检测特征进行差分选择后将其送入检测模型。在模型训练方面结合了集成学习的防御思想,使用部分对抗样本对多个异构的检测模型进行对抗样本训练后再对检测模型进行集成,在提升模型鲁棒性的同时通过集成学习来弥补对抗训练导致的精度损失。最后,由于对抗样本具有迁移性,导致同一特征扰动的对抗样本可能攻破多个模型,因此本文在模型裁决方面融合了启发式种群算法,改进了集成模型的投票裁决策略。该方法从主被动两种防御角度进行了改进:1)从被动防御角度,减小模型对原始数据检测准确率的影响,同时提升模型鲁棒性,增强了整体模型的对抗防御能力;2)从主动防御[10]角度,增加了对抗样本的生成难度,提高了攻击门槛。

本文实验共使用了 3 种性质的流量,分别是:1)正常流量样本,即网络中正常行为所产生的流量,不会触发 NIDS 警报;2)恶意流量样本,即网络攻击行为所产生的流量,可被 NIDS 检测为恶意的流量样本;3)对抗流量样本,即对网络中恶意行为产生的流量进行微小改动,从而无法被 NIDS 检测为恶意的流量样本。本文实验采用 Sharon 等[11]提出的 TANTRA 方法对流持续时间[12]和流量包长度[13]生成的对抗样本进行测试,并使用 EFDS 方法对经典的自编码器(AutoEncoder, AE)[14]、多层感知机(Multilayer Perceptron, MLP)[15]和卷积神经网络(Convolutional Neural Networks, CNN)[16]进行异构训练,与 Sharon 使用的对抗训练方法以及 De 等[17]提出的集成模型并行方式进行对比。结果表明,本文使用的方法一定程度上保证了原始数据的识别准确率,并且防御者可以根据自己的实际情况平衡时间和空间成本,使得检测器拥有更强的检测性能。

综上所述,本文的主要贡献在于:

1)在集成学习思想的基础上构建对抗训练模型,结合主动的特征差分选择,使用了启发式种群算法优化的投票裁决策略。

2)减少了流量对抗样本攻击的攻击面,在提升模型鲁棒性的基础上,主动地提高了对抗样本的生成难度,增加了攻击成本和门槛。

3)所提方法较单一检测模型对抗训练方法在鲁棒性上平均提升了 10%,对原始数据的检测准确率提升了 50%以上;较集成学习方法在两项性能指标上均提升了 10%左右。

本文第 2 章介绍了研究的背景知识和相关工作;第 3 章介绍了 EFDS 模型框架及各个模块的功能及设计实施;第 4 章从防御能力和成本消耗上对实验数据进行了评估;最后总结全文并展望未来。

2 背景知识和相关工作

2.1 背景知识

目前,流量对抗样本生成方法可以分为两种:1)在特征空

间中使用图像作为媒介[18]；2）直接在问题空间中生成对抗样本。如图 1 所示，在图像对抗样本攻击领域，对抗样本可以在图像的任意位置发生[19]，此类对抗样本生成方法称为在特征空间生成对抗样本；但在流量对抗样本生成中，对抗样本还需要满足可执行性和攻击有效性的要求，并且由于流量特征具有高度异质性[7]，很难将对抗性示例映射到流量数据的问题空间[20]，因此目前更多使用的是在问题空间直接针对流量数据生成对抗样本[21]。通过结合流量领域的专业知识，可以提取包序列[22]、特征簇[23]、数据包大小[13]、时间[11-12,24]和空间[25]等特征，这些特征不会改变如 IP 地址等关键特征，从而可以在不破坏流量数据包内容的情况下直接生成对抗样本。

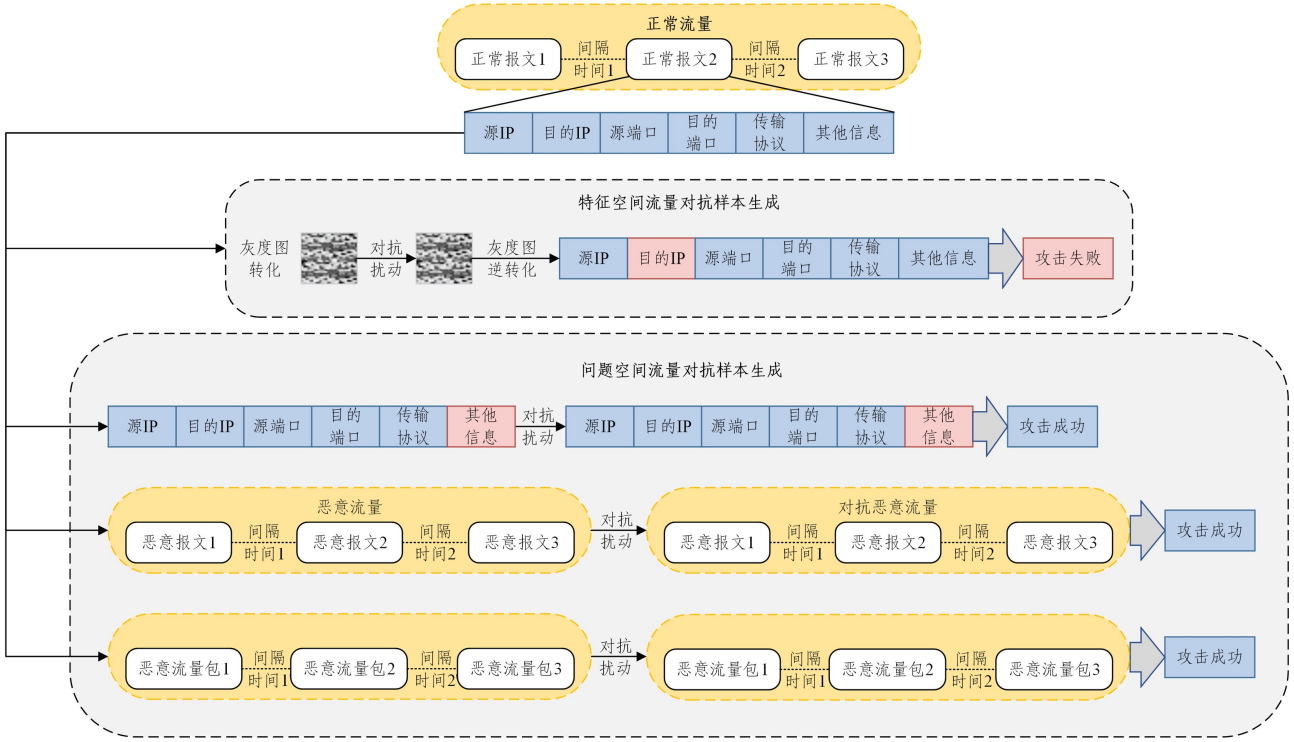


图 1　流量对抗样本生成方法

Fig. 1　Traffic adversarial example generation method

在对抗样本攻击领域，梯度正则化[26]和防御蒸馏[27]等方法运用了主动防御的思想，通过增加对抗样本生成难度来降低攻击的概率。而对抗样本训练[28]和分层集成[2,29-30]则采用了被动防御的思想，常用于增强模型鲁棒性。然而，仅仅增加对抗样本生成的难度只能暂时缓解对抗攻击，当攻击者有足够的次数去试探模型时，他们最终会获取到模型的信息，从而生成有效的对抗样本。对抗样本训练则会对原始数据的识别准确率造成损失，并且只能在已经训练的范围内有效地识别对抗样本，对于新型的对抗样本缺乏防御能力。因此，当面对新型对抗攻击时，防御者只能不断进行对抗训练。综上所述，目前的防御方法缺乏主被动防御思路的融合，对抗训练方法不能兼顾检测性能与模型鲁棒性。

2.2　相关工作

为了主动防御对抗样本，Ross 等[31]提出了梯度正则化的防御方法，通过隐藏训练过程中梯度的变化，使得攻击者无法有效获取模型信息，从而使对抗样本生成难度呈指数级增加。Papernot 等[27]提出了防御蒸馏来提供防御训练，使用从 DNN 中提取的知识来提高其自身对对抗攻击的弹性，从而达到隐藏底层教师模型信息的防御效果。Hashemi 等[32]提出了 RePO 方法，该方法在数据进入神经网络前添加了一个去噪自动编码器从而构建成新的 NIDS 系统。Beechey 等[33]提出利用证据分类方法来抵御 AML 攻击，该方法不需要对扰动数据集进行任何训练，其旨在降低扰动数据的误分类率，而不是提高准确性。大部分学者认为可以将对抗样本当作一种特殊扰动进行处理，只要能够屏蔽掉这些扰动便可以解决此问题。部分学者[2,34]也证明了将对抗样本作为训练集加入机器学习模型中进行训练，可以提高模型鲁棒性。Goodfellow 等[35]指出对抗训练可以使模型正则化以抵御对抗样本攻击，其效果比单纯正则化更好。Chen 等[34]通过对比对抗训练和正则化得出：使用攻击效果越强的对抗样本进行训练，防御效果会越好；但越强的攻击也代表需要耗费更多的时间和空间。Wang 等[28]针对异常流量检测领域对抗样本训练方法所存在的需要生成大量对抗样本、训练时间复杂度高等问题，提出了一种利用反向传播误差过程同时完成样本梯度和模型梯度计算的方法，在不影响对抗样本训练和增强模型鲁棒性的情况下加快了训练速度。

集成模型也是对抗样本防御常用的一种方法。Nhien 等[30]根据随机森林分类器确定的特征重要性对流量数据的像素进行重新排序，将更重要的特征所转换的像素分组映射到图像中心，还指出了真实流量特征是在实时网络流量上进行修改（如更改端口、数据包长度或间隔）的限制。Wang 等[28]利用入侵检测技术决策边界和流形之间的不一致性来进行检测，因为当添加小噪声时，对抗样本的分类结果比干净样本的分类结果更有可能发生变化，作者根据这个特点使用 IDS 分类结果的这种变化来检测流量对抗样本。McCarthy[2]提出了一种基于分层学习的新防御策略，以帮助减少对抗性

示例在预期攻击的参数空间约束下可以利用的攻击面。Debicha 等[17]设计实现了多个基于迁移学习的对抗性检测器，通过组合它们各自的决策结果，提高了流量对抗样本的可检测性。De 等[29]提出了一种新的网络安全分类器深度防御方法，该方法使用分类器的集成模型，每个分类器使用了不同的特征集来保护检测器免受对抗样本攻击。

3 基于特征差分选择的集成模型对抗防御架构

3.1 整体架构

本文设计特征差分选择的集成框架是为了从主被动两个角度去提高检测模型对流量对抗样本的防御能力，并在一定程度上减小异构体对抗训练后对原始数据识别准确率的影响。该框架如图 2 所示，主要由特征差分选择、检测模型集成和投票裁决 3 个功能模块组成。

其中特征差分选择模块负责从流量数据中提取流量特征，将数据从流量的问题空间映射到特征空间，再对特征进行差分选择，最后输出流量检测器可用的数据形式。

在异构的检测体集成模块中，本文使用了传统的流量检测模型 AE[14]，MLP[15] 和 CNN[16]，并使用了正常流量、恶意流量和流量对抗样本对各检测体进行训练。模型在能够检测正常和恶意流量的基础上具有检测恶意流量对抗样本的能力，提高了各异构体的鲁棒性。

检测体集成模块中的各检测体输出判定结果后，使用投票裁决模块来对该流量进行判定。在投票裁决模块中使用了启发式种群算法，通过检测情况对各模型判定结果进行权重赋值，增加投票结果的不确定性，弥补了由于对抗样本可迁移性和攻击者不断尝试而可能造成的防御失效。

通过主动地进行特征差分选择和投票裁决，增加了模型信息的不确定性，从主动防御的角度提高了攻击者的攻击难度。通过使用对抗训练对各异构体进行训练并集成，增强了整体架构的鲁棒性，从被动防御的角度提升了对于对抗样本的识别准确率。

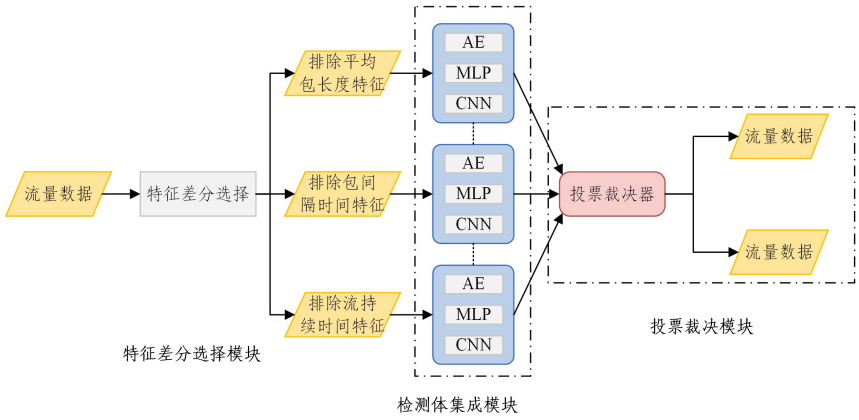


图 2 基于特征差分选择的集成模型对抗防御架构图

Fig. 2 Integrated model adversarial defense architecture based on feature difference selection

3.2 模块设计

3.2.1 特征差分选择模块

特征提取器会在流量经过时，根据流量包头信息及相应的时序统计提取流量(如协议、包时间间隔等)相关特征，因此特征差分选择模块还有对 IP 地址等数据的数据转换及数据归一化等操作，使数据转化为可供模型使用的标准模式。由于每个维度的特征之间属性值差异很大，在数据特征提取阶段还需要使用归一化的方法将每个维度特征值映射在[0,1]之间，数据归一化计算式如式(1)所示：

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

其中，x 是当前数据的数值，$x_{\min}$ 和 $x_{\max}$ 是该组数据中的最小值和最大值，x^* 则是归一化后的数据。

在问题空间上进行流量对抗样本生成常扰动的特征有流持续时间、包长度、包间隔时间[13]。由于流量对抗样本旨在添加最少的扰动来达到最大的攻击效果，因此常采用单特征对抗扰动，从而使特征差分选择可以有效避免对抗扰动对检测器的影响。因此，在本文实验中采用了上述 3 个特征作为对抗样本生成的差分选择特征。

在生成流量对抗样本的过程中，为保证其攻击效果，避免不满足条件的流量流入检测模型而引发警报，对抗样本生成需要符合以下条件：1)生成的对抗样本流量持续时间只能大于等于初始恶意流量持续时间；2)攻击算法预测所使用的时间，须小于流量包发送允许的最大时延。只有生成的流量对抗样本满足以上条件，方可判断其是否可用于实际网络的端到端环境。使用 TANTRA 对抗样本生成方式生成的有效率和均方根误差如表 1 所列，生成的流量对抗样本由于生成规则的限制，生成复杂度和误差会随着扰动特征的增加而增加。在实际研究中，学术界也常采用单特征对抗样本生成进行流量对抗样本生成[7,11-13,20-25]。

表 1 不同类型流量对抗样本生成有效率和均方根误差

Table 1 Different types of traffic against example generation efficiency and RMS error

特征	Bot 生成有效率/%	Bot-RMSE	PortScan 生成有效率/%	PortScan-RMSE	DDoS 生成有效率/%	DDoS-RMSE
流持续时间	83.73	2171.484	82.62	489.95	82.61	4763.67
平均包长度	80.63	2573.910	83.92	87.18	77.28	903.12
包间隔时间	83.93	3245.330	72.25	7664.61	78.72	4362.02

图 3 为特征差分选择模块功能示意图。由于采用了单个特征的对抗扰动生成，当异构数据对异构检测器进行测试时，

可以对 3 种扰动特征进行差分选择，使得部分单一特征的对抗样本扰动生成失效，加大了攻击者攻击成功的难度，对于对抗样本起到了主动筛选特征的防御作用，达到了提高模型鲁棒性的效果。

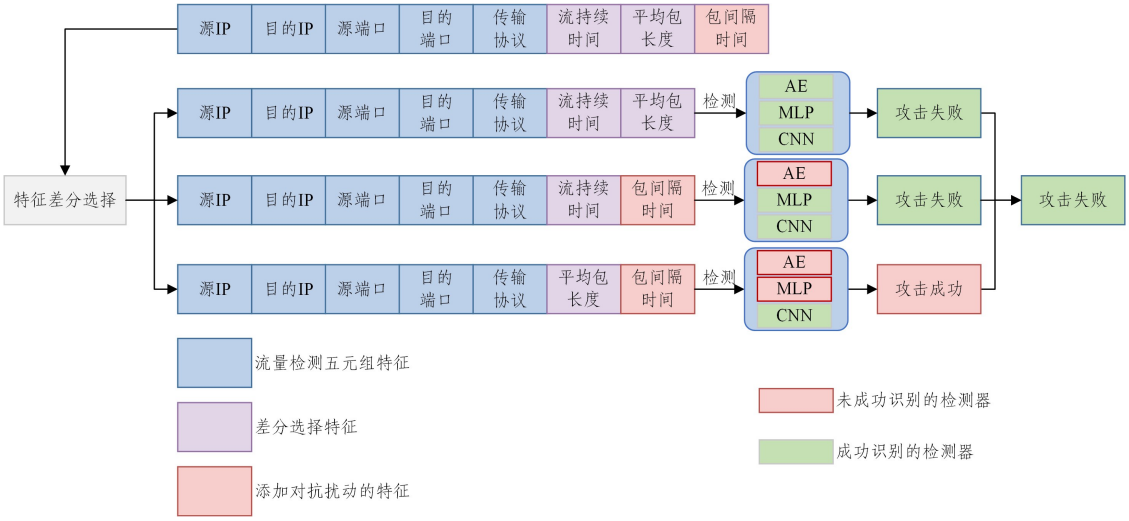
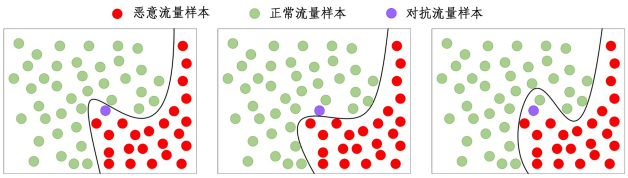


图 3　特征差分选择模块功能示意图

Fig. 3　Schematic diagram of feature difference selection module function

3.2.2　检测体集成模块

如图 4 所示，人工智能模型使用数据集进行训练的本质是使得检测模型能够区分不同类型样本空间的分布，通过划分样本边界来对边界内外的样本进行类型判定。



(a)数据样本空间分布　(b)检测器训练空间分布　(c)检测器对抗训练空间分布

图 4　数据样本空间分布图

Fig. 4　Spatial distribution map of data examples

正常流量和恶意流量分布在多维的样本空间中，相同类型的数据样本聚集在一起，可以使用分类边界将不同类别的数据分隔开来。图 4(a)给出了数据的准确空间分布，正常流量的特征聚集在分类边界的左上方，而分类边界的右下方则是恶意流量数据。人工智能检测模型则是在多维空间中找到数据的分类边界，当数据分布空间处于样本边界右下方时，检测器会将样本判定为恶意流量；相反，当处于样本边界左上方时则判定为正常流量。如图 4(b)所示，恶意样本经过测算模型样本边界的梯度方向并产生微小扰动生成对抗流量样本，从而跨越样本边界导致检测器错误分类。

对抗训练作为防御对抗样本的一种方式，通过将对抗样本加入模型训练，使得模型能够有效覆盖对抗样本范围，达到对其正确识别的效果。但如图 4(c)所示，这也会影响模型对于其他原始样本的识别准确率。模型梯度信息隐藏也是防御对抗样本的有效思路之一[10,26,31,36]，然而已有的方法属于被动式防御，只能达到缓解攻击的效果，在经过攻击者的多次试探之后终会掌握模型信息。为解决单个检测器高维空间分布不完整而导致对抗样本产生的问题，本文设置了检测体集成模块，以弥补单个检测器的不足。

采用了集成模块后，当有半数以上的检测器恶意边界范围覆盖对抗样本时，就可以有效地识别出对抗样本；而对抗训练也只有在超过半数检测器都受到相同影响时才会造成对原始数据类型的错误分类。若采用的异构检测体共有 N 个，检测过程中提取 n 维可扰动检测特征，则存在(N/n)个检测体可免疫对抗样本扰动，即对抗样本攻击需要在剩余的$[(n-1)\times N/n]$个检测体中攻破$(N/2)$个检测体，实际所需攻击检测体的成功比例由$(1/2)$提升到了$[n/2(n-1)]$。

流量数据在经过特征差分选择后会被划分为 3 种异构检测数据集，针对每一种异构数据集搭配一个检测体集成模块，每个检测体集成模块中分别设置了 3 种异构检测体：AE，MLP 和 CNN。现有的流量检测模型大多在 3 种经典模型上进行改进，通过改进使得模型复杂度得到了不同程度的提升。因此，使用无改进的简单模型更能体现出本文使用的防御方法在检测性能和鲁棒性上的提升效果[11,30,37]。

AE 模型由 Encoder 和 Decoder 组成，其中 Encoder 可以实现对数据的压缩，Decoder 则完成对数据的解压。AE 模型在降低数据维度的同时保证了相关信息的完整性，消除了对正确识别没有作用的噪声或信息量较小的特征，能够对有轻微扰动的数据进行降噪识别。MLP 模型则能够捕捉复杂的非线性依赖关系，在输入信息不完全的情况下也能提供一定的预测或决策能力，具有一定程度的模型鲁棒性。CNN 模型是深度学习常用的模型之一，能够适应不同层次的语义信息，过拟合的风险更低，泛化能力更强。

如图 5 所示，本文设置的 AE 模型将输入的 8 维检测数据拉伸至 128 维，再将其编码成 8 维的低维隐变量，从而使神经网络学习最有信息量的特征，之后再将低维隐变量还原到初始的 128 维，最后输出 2 维的判定结果。MLP 模型是一种高度并行的信息处理系统，具有很强的自适应学习能力。本文的 MLP 模型设置有 5 个层次，神经元个数分别为 128，

256,128,64 和 32,最后再输出 2 维判定结果。为了识别一维流量监测数据,在 CNN 模型中设置了内核沿一维进行推动,在 128×1 和 64×1 的卷积层后设置了池化层,再经过 16×1 的卷积层将数据进行拉伸,以适配判定结果的输出结果。由于对于流量的判定结果为二分类,因此各模型的输出为 0 或 1,0 表示正常流量,1 表示恶意流量或流量对抗样本。

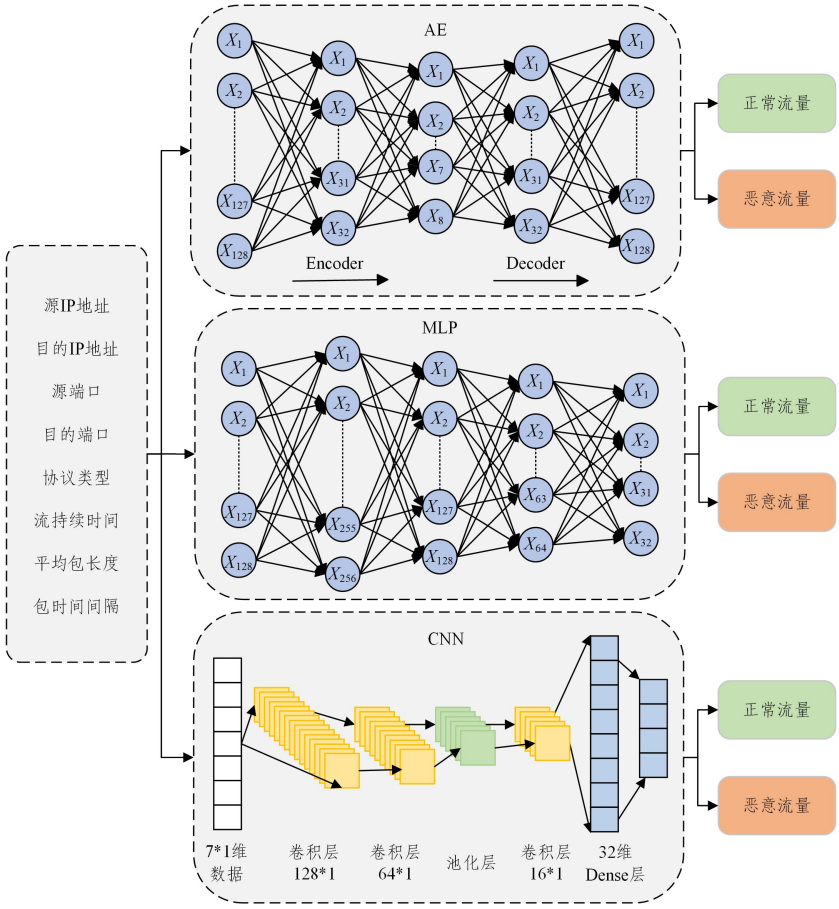


图 5 检测体集成模块示意图

Fig. 5 Detector integrated module diagram

3.2.3 投票裁决模块

在实际网络的动态和随机环境中面临着不确定性,这可能会使得实际处理结果与原始预期出现差异[38],如提高对抗样本识别率可能影响原始流量样本识别率;此外,由于对抗样本在多个模型间具有可迁移的性质,因此同一个对抗样本对于相似或相同的模型也具有攻击效果。故对抗训练后的模型有可能产生相同的分类错误,需要根据实际投票情况更改检测器投票权重,使得最终投票结果接近于实际样本分类。

为缓解传统投票模型受对抗样本可迁移性的影响,本文使用了遗传算法,在多个检测器之间进行权衡并构建一个有效的投票权重分配。投票权重分配的种群算法伪代码如算法 1 所示。

算法 1 种群投票权重分配算法

符号说明:检测体 D^n,$A^n_{恶意}$ 为第 n 个检测体对恶意流量的识别准确率,W^n 为第 n 个检测体的投票权重

最终输出:第 n 个检测体的投票权重 W^n

1. 设置检测体个数为 n
2. 设置迭代次数:Iteration
3. 输入各检测体对 3 种类型的流量识别准确率:$A^1_{恶意}$ $A^1_{正常}$ $A^1_{对抗}$,…, $A^n_{恶意}$ $A^n_{正常}$ $A^n_{对抗}$
4. 准确率矩阵:accuracy_matrix,accuracy_matrix=[$A^1_{恶意}$ $A^1_{正常}$ $A^1_{对抗}$ … $A^n_{恶意}$ $A^n_{正常}$ $A^n_{对抗}$]
5. 归一化权重:weights=weights / sum(weights)
6. 计算加权准确度:weighted_accuracy = accuracy_matrix. T * weights
7. 迭代更新权重至收敛:
8. while True:
9. new_weights=accuracy_matrix *(1/weighted_accuracy)
10. new_weights=new_weights/ sum(new_weights)
11. if max(abs(new_weights - weights)) <1×10^{-4}:
12. break
13. weights=new_weights
14. weighted_accuracy=accuracy_matrix. T * weights
15. iteration=iteration+1
16. 确定最佳权重和准确度:
17. W=weights
18. $W^*=W(D^*)$
19. best_accuracy=max(weighted_accuracy)

如图 6 所示,在投票裁决模块部分有两个阶段:权重更新阶段和流量识别阶段。首先需要进行权重更新,在权重更新

阶段使用含有 3 种性质的测试数据集对各检测模型进行测试，并将测试数据集的标签与各测试模型的判定结果进行对比，以计算识别准确率，再使用种群投票权重分配算法根据各模型的识别准确率进行权重更新。

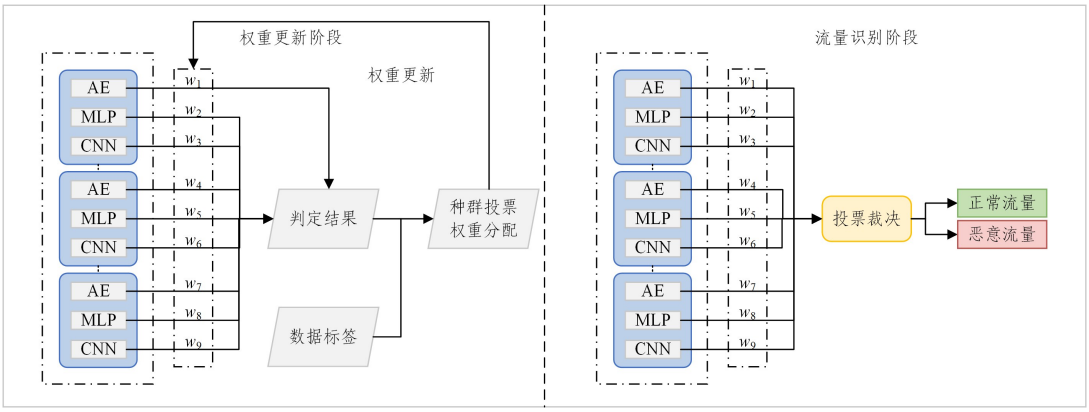


图 6 投票裁决模块示意图

Fig. 6 Schematic diagram of voting adjudication module

在流量识别阶段，待测流量通过检测模型后根据各检测器判定结果 $R_1,R_2,\cdots,R_9$ 及更新后的权重 $w_1,w_2,\cdots,w_9$ 对待测流量进行裁决：

$$S=w_1R_1+w_2R_2+\cdots+w_9R_9$$
$$\begin{cases}\text{恶意/对抗流量}, & S>0.5\\ \text{正常流量}, & S<0.5\end{cases} \tag{2}$$

其中，S 是各检测体投票裁决后的得分，当 S 值大于 0.5 时则该流量更可能为恶意流量或对抗样本，当 S 越接近 1 时概率越大；相反，当 S 小于 0.5 时该流量更可能为正常流量。

4 实验分析

4.1 实验设置

本文使用了 Sharon 等[11]提出的对抗样本训练方法来对 3 种经典检测模型 MLP，AE 和 CNN 进行对抗训练，使用 Debicha 等[17]的防御思想来进行对抗训练并构建集成模型。将所提 EFDS 方法与上述 4 种方法进行对比，以评估 EFDS 在平衡检测性能与鲁棒性上的效果。

4.1.1 实验环境

所有实验均在一台配备 32 GB 内存、Intel Core i9-10900 CPU 和 Intel UHD Graphics 630 的台式机上进行，使用 Python 3.7 编程语言。实验使用 CIC-IDS2017 数据集，并采用 Sharon 等[11]提出的方法生成所需的流量对抗样本，该方法属于对流量样本直接进行对抗样本生成攻击。

4.1.2 评估指标设置

本文使用准确率（Accuracy）、精确率（Precision）和由真阳性率（True Positive Rate，TPR）和假阳性率（False Positive Rate，FPR）组成的 ROC 曲线以及 F1 分数（F1Score）来对实验中使用的模型进行性能评估。

TPR，FPR，Accuracy，Precision，F1 Score 的计算式如下：

$$TPR=\frac{TP}{TP+FN}$$
$$FPR=\frac{FP}{FP+TN} \tag{3}$$
$$Accuracy=\frac{TP+TN}{TP+FP+TN+FN}$$
$$Precision=\frac{TP}{TP+FP} \tag{4}$$

其中，TP 代表预测和实际样本都为正样本；FP 代表预测为正样本，实际为负样本；TN 代表预测和实际都为负样本；FN 代表预测为负样本，实际为正样本。

$F1$ Score 是衡量二分类模型精确度的指标，兼顾了精确率和召回率，可表示为：

$$F1=\frac{2\times Precision\times TP}{Precision+TPR} \tag{5}$$

ROC 曲线又分为 Micro-ROC 曲线及 Macro-ROC 曲线。Micro-ROC 将所有类别的 F1 结果合并，计算整体的性能标准，适用于在不同类别上有明显不均衡样本分布的情况，更关注于整体性能。而 Macro-ROC 曲线对每个类别的 F1 性能指标分别计算平均值，并对这些平均值求平均，适用于每个类别的性能对整体性能重要性一致的情况。ROC 曲线的值介于[0，1]之间，越靠近 1 说明检测分类效果越好，在 ROC 曲线图上则表现为：若 ROC 曲线图越接近图的左上端，则说明该模型的检测分类效果越好。

4.1.3 数据集

实验使用了 CIC-IDS2017 数据集。如表 2 所列，CICIDS-2017[39]数据集由通信安全机构与加拿大网络安全研究所（Canadian Institute for Cybersecurity，CIC）在 2017 年收集，是一个具有复杂现代攻击类型的流量数据集。

表 2 CICIDS-2017 数据集的情况

Table 2 Situation of CICIDS-2017 dataset

攻击种类	恶意流量包数量	正常流量包数量	恶意占比
DDos	128027	97718	0.5671
PortScan	158930	127537	0.5547
Bot	1 966	189068	0.0102
Infiltration	36	288566	0.0001
Brute Force	1507	168186	0.0088
Sql Injection	21		0.0001
XSS	652		0.0038
FTP-Patator	7938	432075	0.0178
SSH-Patator	5897		0.0132

该数据集含有 78 维流量特征和 1 维类别标签，包含正常样本和 14 种攻击样本。实现了包括 PortScan、Bot、DoS、Heartbleed、Web 攻击、渗透等攻击。

在实验过程中将训练和测试的数据分割率设定为 60% 和 40%。由于部分类型网络攻击数据量在总数据量中占比较低，为避免数据不平衡的发生，本文使用了 DDos，PortScan 和 Bot 来分别生成恶意流量对抗样本加入防御模型训练。所有实验均使用异常流量检测常用的 8 维检测特征（源 IP、目的 IP、源端口、目的端口、协议类型、流持续时间、平均包长度、数据包间隔时间）进行检测。由于采用了特征差分选择方法，因此各流量检测体均为 7 维输入。

CIC-IDS2017 数据集的标签列标记了流量的类型，包括 Benign，DDOS，PortScan 等文本化标签，以及源 IP 地址和目的 IP 地址如 192.168.0.0 类型的数据，但机器学习算法不能直接对此类数据进行处理，因此在实验前对这些数据进行映射处理，将其转化为数值，例如将 Benign 转化为 0，将 DDoS 转化为 1。

4.1.4 实验参数设置

为保证正常流量和恶意流量的平衡，实验采用的检测模型统一使用 CICIDS-2017 数据集的 DDoS，PortScan 和 Bot 数据，按照训练集与测试集为 3∶2 的比例进行训练及测试。使用了窗口数为 3 的 TANTRA 方法进行对抗样本生成，所生成的对抗样本按照 1∶9 的比例划分为对抗训练的训练集和测试集。在超参数设置方面，各检测模型均设置学习率为 0.001，batch_size 设置为 2000，学习轮数为 10。

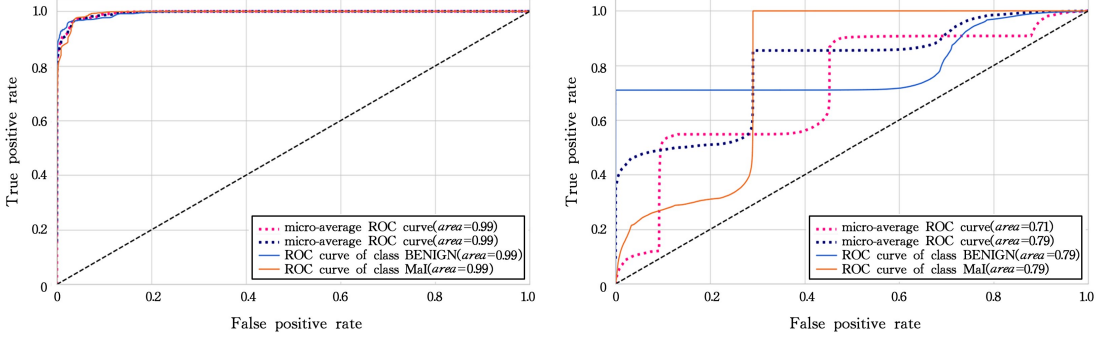
4.2 对抗防御效果评估

在对抗防御效果评估这一部分，本文将探究检测模型对对抗样本的防御能力，并比较对抗训练对原始检测数据集准确度的损失以及集成检测所需消耗的时间成本。本文将通过实验来评估模型在防御对抗样本方面的表现，并分析其对原始检测数据集的影响以及异构检测的时间开销，这将有助于更好地了解模型在面对对抗攻击时的鲁棒性和实用性。

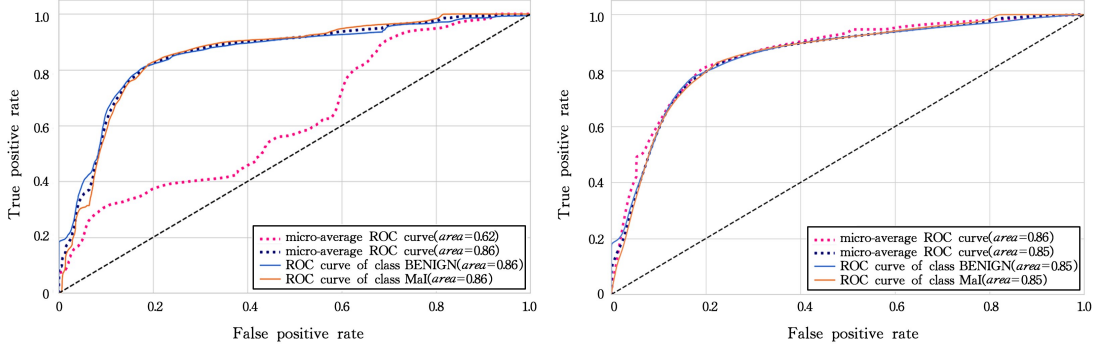
4.2.1 对抗样本防御能力

实验部分共使用了 3 种类型的流量数据，其中将正常和恶意流量组成的数据称为原始数据，在恶意流量基础上进行对抗扰动生成的流量样本则称为对抗样本。为了验证 EFDS 的效果，实验采用了 MLP，AE 和 CNN 这 3 种未经过任何性能提升的传统流量检测模型，通过在对抗训练前和对抗训练后分别使用原始数据和对抗样本对模型进行测试，形象地展示对抗训练对模型的原始数据集识别准确率的影响和鲁棒性提升。

图 7—图 9 分别是 3 种检测模型在对抗训练前后对于 3 种流量识别准确率的 ROC 曲线图，其中粉红的点线代表 Micro-Roc 曲线、深蓝色点线代表 Macro-Roc 曲线、蓝色和黄色实线是流量二分类为正常和恶意的 ROC 曲线。图 7(a)、图 8(a)、图 9(a) 为对抗训练前后检测模型对原始流量的 ROC 曲线图，从图中可以明显发现，在对抗训练后 4 条曲线均往右下方移动，说明对抗训练使模型对原始数据的识别准确率造成了不利影响，其中 CNN 模型下降最多，ROC 覆盖面积只略微超过了图形面积的一半。而图 7(b)、图 8(b)、图 9(b) 为对抗训练后 3 种模型对于对抗样本的检测率，可以看到 ROC 曲线明显得到了提高，证明对抗训练方法能够有效提高模型的鲁棒性，但会影响检测器的检测性能。



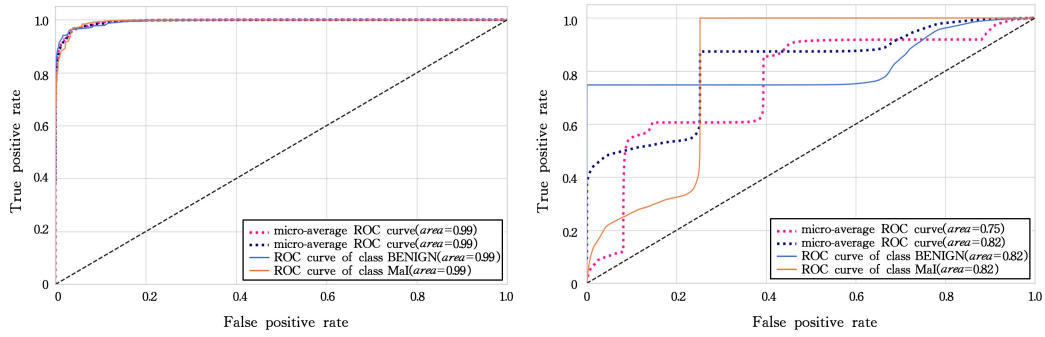
(a) MLP 对抗训练前后原始流量准确率



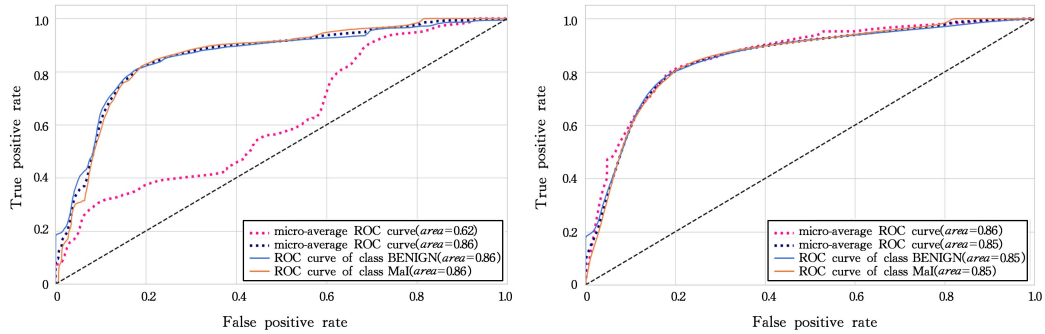
(b) MLP 对抗训练前后对抗样本准确率

图 7 MLP 检测模型对抗训练前后检测性能与鲁棒性对比（电子版为彩图）

Fig. 7 Comparison of detection performance and robustness of MLP detection model before and after adversarial training



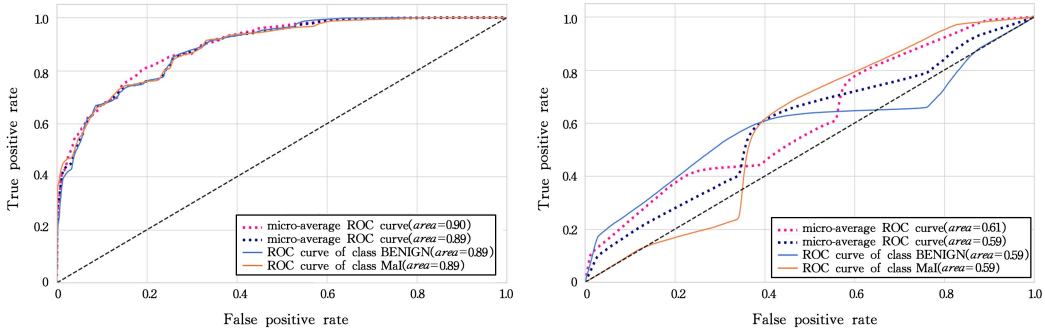
(a)AE 对抗训练前后原始流量准确率



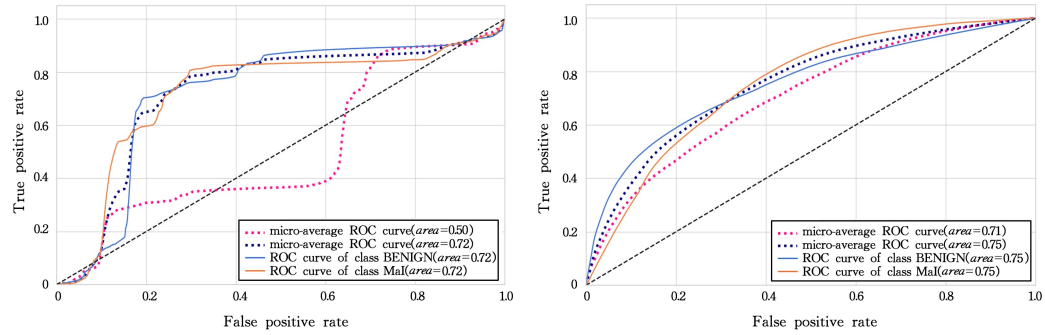
(b)AE 对抗训练前后对抗样本准确率

图 8 AE 检测模型对抗训练前后检测性能与鲁棒性对比(电子版为彩图)

Fig. 8 Comparison of detection performance and robustness of AE detection model before and after adversarial training



(a)CNN 对抗训练前后原始流量准确率



(b)CNN 对抗训练前后对抗流量准确率

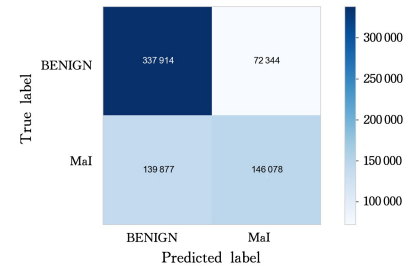
图 9 CNN 检测模型对抗训练前后检测性能与鲁棒性对比(电子版为彩图)

Fig. 9 Comparison of detection performance and robustness of CNN detection model before and after adversarial training

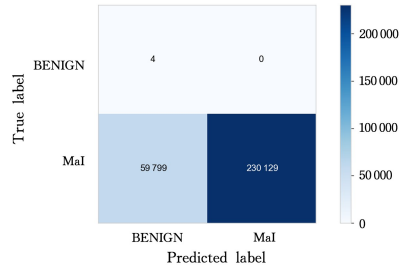
图 10 给出了集成模型和 EFDS 方法对抗训练后对原始数据和对抗样本的识别准确度。集成模型使用了 3 种检测器进行堆叠,从结果可知该方法大幅度地提高了对抗训练后原始数据的识别准确率,并保证了对抗样本的防御能力。

由于使用的 Bot 类型恶意流量与 PortScan 和 DDoS 类型流量的数量比例差别明显,属于网络攻击中数据量占比较小的部分。因此,还需对 3 种不同类型恶意流量及其生成的对抗样本进行分类研究,以分析不同占比的数据在对抗样本训练后对模型产生的影响。表 3 列出了不同类型数据对抗训练

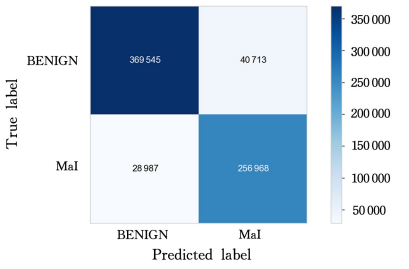
前后的识别准确率对比结果，其中 PortScan 和 DDoS 的恶意占比都接近于 50%，EFDS 与集成模型在这两类数据上的识别准确率都比经典检测模型高；而 Bot 的恶意占比为1.02%。在使用经典检测模型进行对抗训练后，3 种模型的识别效果也存在着较大差异，但是较集成模型与所提方法的识别准确率更高。



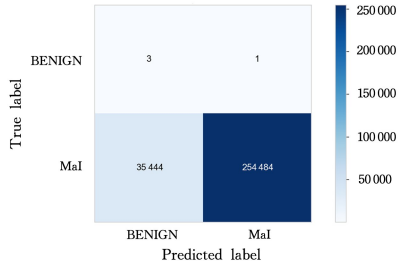
(a)集成模型对抗训练后流量准确率



(b)集成模型对抗训练后对抗样本准确率



(c)EFDS 对抗训练后流量准确率



(d)EFDS 对抗训练后对抗样本准确率

图 10 集成模型与 EFDS 对抗训练前后数据识别准确率

Fig. 10 Recognition accuracy of integrated method and EFDS before and after training

表 3 不同类型数据对抗训练前后识别准确率的对比

Table 3 Comparison of identification accuracy before and after adversarial training with different types of data

(%)

识别模型	Bot 恶意类型流量（恶意占比 1.02%）				PortScan 恶意类型流量（恶意占比 55.47%）				DDoS 恶意类型流量（恶意占比 56.71%）			
	训练前流量识别准确率	训练前对抗样本识别率	训练后流量识别准确率	训练后对抗样本识别率	训练前流量识别准确率	训练前对抗样本识别率	训练后流量识别准确率	训练后对抗样本识别率	训练前流量识别准确率	训练前对抗样本识别率	训练后流量识别准确率	训练后对抗样本识别率
MLP	93.60	54.80	56.12	80.76	90.85	54.80	78.56	80.37	86.59	52.80	76.12	80.63
AE	92.87	59.25	71.10	77.86	96.42	59.25	85.45	76.80	93.45	59.25	82.40	77.64
CNN	93.17	39.40	66.99	66.84	92.36	39.41	34.74	64.49	74.47	39.40	50.83	64.87
集成模型	—	—	81.63	41.71	—	—	63.79	93.40	—	—	67.89	66.68
EFDS	—	—	87.47	62.02	—	—	91.02	93.76	—	—	72.22	81.51

通过分析可知，出现上述情况是由于 Bot 恶意类型流量的占比过小，导致模型出现了小样本问题。虽然 MLP 和 AE 在对抗训练后有较高的识别准确率，但由 3 个模型组成的集成模型却表现不佳，说明在某些数据上只存在一个检测模型正确地识别出数据类型，投票权重无法过半而使最终结果正确，这也证明了模型无法有效地学习到恶意流量和对抗样本特征的提取知识。

表 4 列出了所使用的各类型数据混合后，各模型在对抗训练前后识别准确率的对比数据。从中可知，EFDS 采用了特征差分选择的方法，较集成模型来看同时提高了模型性能和鲁棒性，两类数据的识别准确率都逼近 90%，有效地避免了对抗训练对模型造成的影响。相较于 3 种单一的经典检测模型，集成模型和 EFDS 在对抗训练后对原始流量数据的识别准确率提高了 40%和 50%左右，对抗样本识别率也较单一检测模型有一定提高。EFDS 与集成模型相比，对抗训练后对 3 种数据的识别准确率提高了近 10%。

表 4 对抗训练前后识别准确率的对比

Table 4 Comparison of accuracy before and after adversarial training

(%)

采用模型	训练前流量识别准确率	训练前对抗样本识别率	训练后流量识别准确率	训练后对抗样本识别率
MLP	96.51	34.80	47.82	80.75
AE	96.32	39.24	41.70	77.28
CNN	79.81	19.40	33.30	64.40
集成模型	—	—	66.51	79.37
EFDS	—	—	89.99	87.77

4.2.2 检测成本对比

为降低检测成本，本文对 EFDS 采用了多进程同步检测的方法，这在一定程度上增加了空间成本，但换来了极大成度的检测准确度的提升。在针对两个多模型检测方法进行实验时设置了 3 个进程池，其余单个对抗训练模型只设置了 1 个进程池。EFDS 的时间成本随着排除特征和异构检测模型种类的增加而增加，随进程池的增加而减小，因此该方法需要根据防御者的实际情况实现时间成本和空间成本的平衡。

表 5 列出了检测体个数和异构模型类别对对抗训练后流量识别准确率的影响情况。从中可以得出 EFDS 有效提高了模型鲁棒性，并且随着检测体个数和模型类别的增加，模型鲁棒性也随之增强。在检测体个数为 3 时，使用该方法较集成模型对原始流量识别准确率的影响更小，本文推测是由于集成类方法虽然弥补了单个检测器的检测空白，但也可能存在多个模型具有同一漏洞的概率，使得整个模型也产生一定的误差。

表 5 EFDS 不同模型类别与检测体个数识别准确率的对比

Table 5 Comparison of recognition accuracy of different EFDS model categories and number of detectors

异构模型类别	检测体个数	识别准确率	
		训练后流量识别准确率/%	训练后对抗样本识别率/%
MLP	3	90.11	79.79
AE	3	90.12	78.80
CNN	3	81.93	79.39
MLP+AE	6	87.86	83.34
MLP+CNN	6	87.31	82.37
AE+CNN	6	85.65	81.72
MLP+AE+CNN	9	89.99	87.77

结束语 由于流量数据与图像和语音数据具有较大差异，因此学术界对于流量对抗样本防御方法的研究较少，而对抗样本训练则是可以使用且防御效果良好的方法之一，但对抗样本防御需要防御者在模型检测性能和鲁棒性之间做出抉择。本文提出的 EFDS 有效地解决了对抗样本训练的不足，在防御理念上将被动的防御思想与主动防御的理念融合，通过特征差分选择构建异构训练集，从而达到主动排除对抗扰动信息的效果；通过集成异构检测模型并设置投票权重分配算法来隐藏梯度模型信息，提高了攻击者的攻击门槛，在提升模型鲁棒性的同时也尽可能减小了对抗训练对检测性能的影响。在与 3 种经典检测模型、集成模型进行对比后可知，EFDS 在保证对抗训练后模型对原始数据识别性能的基础上，也一定程度提升了对流量对抗样本的识别能力，识别性能和鲁棒性较单一模型提高了约 50%和 10%，在检测体数量相同的情况下较集成模型也有一定程度的提升。

虽然本文提出的 EFDS 具有良好的性能提升，但与集成模型一样需要消耗更多的成本，需要防御者根据实际情况对时间和空间成本进行评估选择。未来的研究可以针对减少检测空间成本消耗展开。此外，EFDS 对于网络攻击比例分布较低的数据效果不明显。因此，未来可以结合解决小样本的研究和解决样本泛化方面的研究来共同开展，在更高维度空间中实现识别范围的最大覆盖，从而从根源上解决对抗样本问题。

参 考 文 献

[1] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-Based Learning Applied to Document Recognition[J]. The IEEE, 1998, 86(11): 2278-2324.

[2] MCCARTHY A, GHADAFI E, ANDRIOTIS P, et al. Defending against adversarial machine learning attacks using hierarchical learning: A case study on network traffic attack classification [J]. Journal of Information Security and Applications, 2023, 72: 103398.

[3] BONNET B. Understanding, taming, and defending from adversarial examples[D]. Université de Rennes, 2023.

[4] KO K, KIM S H, KWON H. Multi-targeted audio adversarial example for use against speech recognition systems[J]. Computers & Security, 2023, 128: 103168.

[5] MACAS M, WU C, FUERTES W. Adversarial examples: A survey of attacks and defenses in deep learning-enabled cybersecurity systems [J]. Expert Systems with Applications, 2023: 122223.

[6] FAN H, WANG R, HUANG X, et al. Deep joint adversarial learning for anomaly detection on attribute networks[J]. Information Sciences, 2024, 654: 119840.

[7] WANG K, WANG Z, HAN D, et al. BARS: Local Robustness Certification for Deep Learning based Traffic Analysis Systems [C]// NDSS. 2023.

[8] ANTHI E, WILLIAMS L, RHODE M, et al. Adversarial attacks on machine learning cybersecurity defences in industrial control systems[J]. Journal of Information Security and Applications, 2021, 58: 102717.

[9] HORCHULHACK P, VIEGAS E K, LOPEZ M A. A Stream Learning Intrusion Detection System for Concept Drifting Network Traffic[C]// 2022 6th Cyber Security in Networking Conference (CSNet). IEEE, 2022: 1-7.

[10] HU Y J, GUO Y B, MA J, et al. Method to generate cyber deception traffic based on adversarial example[J]. Journal on Communications, 2020, 41(9): 59-70.

[11] SHARON Y, BEREND D, LIU Y, et al. Tantra: timing-based adversarial network traffic reshaping attack[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 3225-3237.

[12] NOVO C, MORLA R. Flow-based detection and proxy-based evasion of encrypted malware c2 traffic[C]// Proceedings of the 13th ACM Workshop on Artificial Intelligence and Security. 2020: 83-91.

[13] SADEGHZADEH A M, SHIRAVI S, JALILI R. Adversarial network traffic: Towards evaluating the robustness of deep-learning-based network traffic classification[J]. IEEE Transactions on Network and Service Management, 2021, 18(2): 1962-1976.

[14] XIANG Y, HØJVANG J L, RASMUSSEN M H, et al. A two-stage deep representation learning-based speech enhancement method using variational autoencoder and adversarial training [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 32: 164-177.

[15] YIN Y, JANG-JACCARD J, XU W, et al. IGRF-RFE: a hybrid feature selection method for MLP-based network intrusion detection on UNSW-NB15 dataset[J]. Journal of Big data, 2023, 10(1): 15.

[16] ARIVAZHAGAN S, RUSSEL N S, SARANYAA M. CNN-based Approach for Robust Detection of Copy-Move Forgery in Images[J]. Inteligencia Artificial, 2024, 27(73): 80-91.

[17] DEBICHA I, BAUWENS R, DEBATTY T, et al. TAD: Trans-

fer learning-based multi-adversarial detection of evasion attacks against network intrusion detection systems[J]. Future Generation Computer Systems, 2023, 138: 185-197.
[18] SHU D, LESLIE N O, KAMHOUA C A, et al. Generative adversarial attacks against intrusion detection systems using active learning[C]// Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning. 2020: 1-6.
[19] MACHADO G R, SILVA E, GOLDSCHMIDT R R. Adversarial machine learning in image classification: A survey toward the defender's perspective[J]. ACM Computing Surveys, 2021, 55(1): 1-38.
[20] SUN P, LI S, XIE J, et al. GPMT: Generating practical malicious traffic based on adversarial attacks with little prior knowledge [J]. Computers & Security, 2023, 130: 103257.
[21] RUST-NGUYEN N, SHARMA S, STAMP M. Darknet Traffic Classification and Adversarial Attacks Using Machine Learning [J]. Computers & Security, 2023, 127: 103098.
[22] CHENG Q, ZHOU S, SHEN Y, et al. Packet-level adversarial network traffic crafting using sequence generative adversarial networks[J]. arXiv: 2103. 04794, 2021.
[23] CHERNIKOVA A, OPREA A. Fence: Feasible evasion attacks on neural networks in constrained environments [J]. ACM Transactions on Privacy and Security, 2022, 25(4): 1-34.
[24] WANG N, CHEN Y, XIAO Y, et al. Manda: On adversarial example detection for network intrusion detection system [J]. IEEE Transactions on Dependable and Secure Computing, 2022, 20(2): 1139-1153.
[25] HUANG W, PENG X, SHI Z, et al. Adversarial attack against LSTM-based DDoS intrusion detection system[C]// 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence(ICTAI). IEEE, 2020: 686-693.
[26] CHEN J Y, WU C A, ZHENG H B. Novel defense based on softmax activation transformation[J]. Chinese Journal of Network and Information Security, 2022, 8(2): 48-63.
[27] PAPERNOT N, MCDANIEL P, WU X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]// 2016 IEEE Symposium on Security and Privacy (SP). IEEE, 2016: 582-597.
[28] WANG B, GUO Y K, QIAN Y G, et al. Defense of Traffic Classifiers based on Convolutional Networks against Adversarial Examples[J]. Journal of Cyber Security, 2022, 7(1): 145-156.
[29] DE LUCIA M J, COTTON C. A network security classifier defense: against adversarial machine learning attacks[C]// Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning. 2020: 67-73.
[30] RUST-NGUYEN N, SHARMA S, STAMP M. Darknet traffic classification and adversarial attacks using machine learning[J]. Computers & Security, 2023, 127: 103098.
[31] ROSS A, MACHADO G R, SILVA E, et al. Adversarial machine learning in image classification: A survey toward the defender's perspective[J]. ACM Computing Surveys(CSUR), 2021, 55(1): 1-38.
[32] HASHEMI M J, KELLER E. Enhancing robustness against adversarial examples in network intrusion detection systems[C]// 2020 IEEE Conference on Network Function Virtualization and Software Defined Networks(NFV-SDN). IEEE, 2020: 37-43.
[33] BEECHEY M, LAMBOTHARAN S, KYRIAKOPOULOS K G. Evidential classification for defending against adversarial attacks on network traffic[J]. Information Fusion, 2023, 92: 115-126.
[34] CHEN S H, SHEN H J, WANG R, et al. Relationship Between Prediction Uncertainty and Adversarial Robustness[J]. Journal of Software, 2022, 33(2): 524-538.
[35] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv: 1412. 6572, 2014.
[36] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. arXiv: 1503. 02531, 2015.
[37] MACAS M, WU C, FUERTES W. Adversarial examples: A survey of attacks and defenses in deep learning-enabled cybersecurity systems[J]. Expert Systems with Applications, 2023, 238: 122223.
[38] BORGONJON T, MAENHOUT B. A genetic algorithm for the personnel task rescheduling problem with time preemption[J]. Expert Systems with Applications, 2024, 238: 121868.
[39] SHARAFALDIN I, LASHKARI A H, GHORBANI A A. Toward generating a new intrusion detection dataset and intrusion traffic characterization[J]. ICISSP, 2018, 1: 108-116.

HE Yuankang, born in 1999, master. His main research interests include network security and cyberspace security, machine learning and adversarial example.

MA Hailong, born in 1980, Ph. D, professor, Ph. D supervisor. His main research interests include endogenous security in cyberspace, intelligent awareness of cyber threats, and innovative cyber systems.

(责任编辑:何杨)