交互博弈引导的网络流量异常检测建模方法研究

张文哲, 杨　栋, 魏松杰

南京理工大学 计算机科学与工程学院 南京 中国 210094

摘要　基于网络流量的系统入侵会带来严重破坏, 因此寻找能够准确识别和分类异常流量的方法具有重要的研究价值。数据作为基于机器学习模型的检测算法的唯一依据, 训练过程对于外界是一个黑盒过程, 整个模型在训练和使用过程中缺乏用户交互。这导致在网络运维场景中, 专业运维人员不能根据当前模型检测结果, 实时将指导信息反馈到系统中, 进而削弱了系统的场景适应能力和检测纠错能力。本文基于强化学习过程, 设计了一种基于动态贝叶斯博弈的交互引导式的网络流量异常检测方法。通过检测模型和运维人员交互的方式, 在训练过程中让运维人员提供专业反馈使得模型获得外界针对当前检测效果的奖惩信号, 从而对自身特征聚焦方向和收敛过程起到引导的作用。将运维人员和检测模型视为博弈的双方, 建立博弈模型, 使双方之间的交互引导行为达到动态平衡状态。通过博弈对于模型交互频次和内容反馈给出指导, 从而使得模型具有动态适应当前场景的能力, 有效控制了人机交互反馈所带来的系统开销。实验部分验证了交互式博弈的流量检测方法中, 双方博弈指导交互行为的可行性与有效性, 证明了该方法在动态场景中具有良好的适应能力。相较于传统的机器学习方法, 交互引导式模型提高了模型整体的检测性能。性能对比测试结果表明交互频次每增加 0.02%, 系统整体检测性能随之提升 0.01%。

关键词　动态贝叶斯博弈; 强化学习; 网络流量; 异常检测

中图分类号　TN915.08　DOI 号 10.19363/J.cnki.cn10-1380/tn.2024.03.03

Interactive-Gaming Guided Modeling and Detection for Network Traffic Anomaly Detection

ZHANG Wenzhe, YANG Dong, WEI Songjie

School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

Abstract　Since system intrusion through network traffic may cause serious damages, it is of great value to research for more accurate methods for network traffic recognition and anomaly classification. Traditional machine-learning based detection methods rely only on data, with the model training and application procedures lack interaction with domain users, which makes the mode just mystery running in a black box. The domain experts in network anomaly detection scenarios cannot provide instant feedback about the model detection results to the system, and thus the detection system is short of adaptability and self-correction capability in these scenarios. This paper proposes an interaction guided network traffic anomaly detection based on the improved reinforcement learning procedure with the dynamic Bayesian gaming. The new model training and detection procedure enables system administrators and domain experts to return feedbacks about the model behaviors into the system as incentive signals for feature focusing and model convergence. System administrators and detection models are interacting with each other following the gaming theory to approximate a dynamic equilibrium state. We design the interactive gaming strategy to control the interaction frequency and content, which optimize the detection model to achieve dynamic adaptability to the current network traffic scenarios, with constrained interaction overhead. We have conduct experiments with public dataset for traffic anomaly detection to verify the interactive gaming performance, detection improvement and effectiveness. The experimental results effectively prove that the interaction-guided model has good adaptability and usability in dynamic scenarios. It can make the interaction frequency controllable by adjusting parameters. It can achieve a balance between performance and interaction frequency on data sets of different types and scenarios. Compared with traditional machine learning methods, the interactive guided model improves the overall detection performance of the model. Results show that the detection performance is improve by 0.01% for every 0.02% more interaction frequency.

通讯作者：魏松杰, 副教授, Email: swei@njust.edu.cn。

本课题得到国家重点研发计划子课题内生安全交换机关键技术研究(No. 2020YFB1804604)、工业互联网创新发展工程项目工业企业网络安全综合防护平台(No. TC200H01V)资助。

收稿日期: 2022-05-25; 修改日期: 2022-07-05; 定稿日期: 2023-11-02

Key words dynamic Bayesian gaming; reinforcement learning; network traffic; anomaly detection

1 引言

由于网络中的任何入侵和异常都会严重影响许多领域, 如私人和社会数据的安全、国家安全、社会和金融问题等, 因此, 科学界对信息和通信的安全性越来越关注。自 21 世纪初以来, 研究人员在异常检测领域做出大量研究, 并提出了众多异常检测方法, 主要分为基于规则的异常检测和基于机器学习的异常检测两大类。其中基于机器学习的异常检测框架又分为无监督学习和有监督学习。在有监督学习中, 利用已标记数据集的有用信息进行流量的分类, 使用此类有监督学习算法会达到高预测精度的效果。然而, 手动标记所有数据这一过程任务量较大。在无监督学习中, 模型通过寻找数据间的共同点去做相应的分类, 整个过程不依赖于标签, 但是其训练效果相比有监督学习较差, 寻找能够检测恶意流量的新的快速而强大的算法模式对于应对不断变化的威胁和增加的检测难度至关重要。

任何的网络流量异常检测模型都要具备适应不同网络场景的能力, 随着网络环境的变化, 模型要做出相应变化。在传统的网络流量异常检测中, 将某个场景中的数据进行定量采集后, 将其作为预先设计的机器学习算法模型的输入, 进行模型的训练, 当模型训练完备后, 将当前模型发布到正式生产环境中使用, 在使用过程中模型不具备针对当前网络的自适应能力。对于新型的攻击模型无法识别时, 发出告警将对于当前数据的检测和排查工作交由网络运维人员处理。面对当下比以往规模更大、结构更复杂的网络基础设施、基础服务和线上业务环境, 该过程加大了网络运维人员的工作量, 已经无法满足当前监控和管理任务的需要。智能运维(Artificial intelligence for IT operations, AIOps)的提出基于已有的运维数据, 通过机器学习算法和运维场景结合的方式来进一步解决自动化运维所未能解决的问题, 以提高监测系统的分析预判能力、准确率和稳定性, 并有效地降低运维成本。但是 AIOps 中的算法模型使用已有的运维数据(日志、监控信息、应用信息等)进行驱动, 往往训练过程对于外界是一个黑盒子, 已有的历史数据成为模型的唯一依赖, 整个模型在训练和使用过程中不具有与用户交互的能力和机制, 进而导致在运维环境中运维人员不能根据模型判断结果实时将信息反馈到 AIOps 系统中, 从而削弱了系统的适应能力和纠错能力。

针对上述问题, 本文结合强化学习算法的 Reward 回馈机制, 提出一种交互引导式的模型进行网络流量异常检测。相较于传统的 AIOps 系统, 在该场景中采用交互式博弈的方式, 运维人员通过可视化交互界面, 对模型的分析与决策进行反馈和评估, 进而使得模型不断的在运维人员提供的外部信息训练过程中获得更加精确的训练结果。为了更好的使得运维人员和模型之间进行合作, 建立模型合理的交互策略以及模型合理的反馈策略, 本文使用动态贝叶斯博弈模型建立了一个二者间的信誉评估和更新机制, 通过博弈建立起运维人员和模型之间的均衡态, 从而指导模型和运维人员的行为。

本文结构如下所示: 第一节为引言, 第二节介绍网络流量异常检测领域相关工作, 第三节介绍基于动态贝叶斯博弈的交互式引导异常检测方法以及相关理论, 第四节详细描述系统中人机博弈模型的构建流程, 第五节对实验结果进行整理和分析, 第六节为总结和展望。

2 相关工作

针对网络异常流量检测, 机器学习作为一种基于统计的分析工具, 已经在各个领域得到了广泛的讨论和应用。针对网络异常检测所制定的系统为入侵检测系统(Intrusion detection system,IDS), 是用来保护网络的一种典型的对抗手段。当前机器学习领域应用于 IDS 的算法多为传统的机器学习算法, 例如支持向量机(Support vector machines, SVM), K 近邻算法(K-nearest neighbors, KNN), 人工神经网络(Artificial neural networks, ANN)和决策树(Decision tree, DT)以及深度神经网络(Deep neural network, DNN)。

在网络流量的历史数据中, 每条数据在空间和时间上相互关联依赖, 所以网络流量的历史数据之间具有高度非线性和复杂性的特征。2012 年 K. Sethi 等人[1]提出了一种将日志关联和强化学习相结合的入侵检测系统。算法通过奖励回馈机制来识别已知和未知的攻击。2015 年 Kumar 等人[2]使用季节性差分自回归滑动平均模型(Seasonal autoregressive integrated moving average, SARIMA)模型, 通过分析流量分布进而设置流量分类阈值的方法进行流量的时间序列预测, 但是该方法设置的阈值不具有普适性, 难以

适应各种网络场景。同年 Basant Subba 等人[3]提出线性判别分析(Linear discriminant analysis, LDA)和逻辑回归(Logistic regression, LR)应用于网络异常检测中, 相对于 SVM 具有更小的整体系统开销, 并且便于实际应用中的部署。2017 年 Zhang Xiaofeng 等人[4]提出了一种改进的半监督学习网络入侵检测算法。该算法首先使用改进的 K-means 将要检测的数据划分为不同的聚类, 然后使用多级 SVM 对标记为异常的簇进行分类, 以达到提高检测效率的目的。2018 年 R. Blanco 等人[5]在多层感知器(MLP)上使用深度 Q 网络(Deep Q network, DQN)体系结构构建基于 RL 的 IDS 系统。Daochen Zha 等人[6]提出了一种采用元策略的主动异常检测新框架, 并命名为 Meta-AAD。Meta-AAD 框架通过近端策略梯度(Proximal policy optimization, PPO)进行实例化, 利用深度强化学习(Deep reinforcement learning, DRL)来训练元策略以选择最合适的实例, 从而在整个查询过程中显式优化发现的异常数量。随后 G. Apruzzese 等人[7]提出通过双深 Q 网络(Double deep Q network, DDQN)算法进行僵尸网络检测程序的逃避攻击, 自动生成能够逃避检测的攻击样本, 并使用此类样本生成用于生成强化检测器的增强训练集, 然后利用深度强化学习算法来生成保留其恶意逻辑并能够逃避检测的对抗样本。

在此类数据驱动的算法模型中, 数据为算法模型的唯一依赖, 整个模型不具有与用户交互的能力, 导致模型缺乏对不同真实场景的适应能力。为了增强模型在网络流量异常检测系统中的动态适应性, 本文结合强化学习算法, 提出了一种基于动态贝叶斯博弈的交互引导式模型用于网络流量异常检测, 通过模型和运维人员交互的方式增加模型的可维护性和扩展性, 并且在其基础上建立动态贝叶斯博弈模型, 寻找二者交互的平衡点。

3 交互式引导检测方法

基于贝叶斯博弈的交互式引导检测方法整合了强化学习算法的奖励回馈机制以及动态贝叶斯博弈。在异常检测系统中运维人员通过给予模型反馈使得模型快速适应动态环境。

系统整体结构如图 1 所示。该架构主要分为 2 个部分: 第 1 部分为基于强化学习的交互式引导异常检测方法, 第 2 部分为交互式博弈过程中的博弈系统。本节对于基于强化学习的交互式引导方法进行介绍, 第 4 节对人机博弈系统进行阐述。

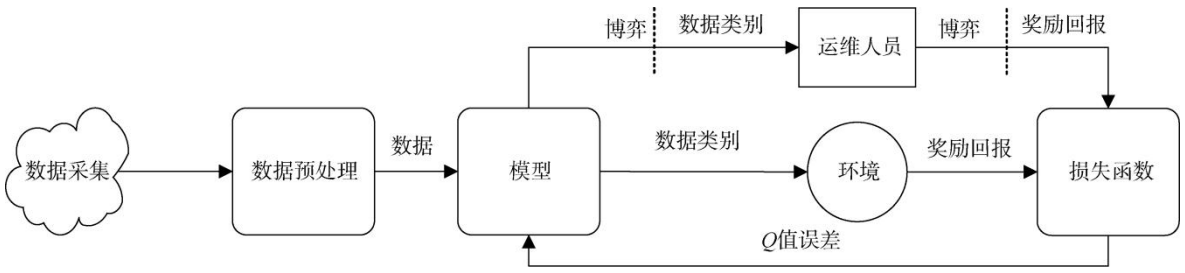

图 1 系统结构图

Figure 1 System structure diagram

3.1 算法概述

强化学习(Reinforcement learning, RL)是一种机器学习范式, 具有通过在动态环境中的模拟试错来自我学习的能力[8]。在 RL 框架中, 代理、状态、动作、环境和奖励是学习循环的组成部分, 即 agent, environment, state, reward, action。其中模型为强化学习的本体, 作为决策者; 环境为强化学习智能体以外的组成部分, 主要由状态集合组成; 状态用来表示环境的数据, 状态集则是环境中所有可能的状态; 动作是智能体可以做出的动作; 奖励为智能体执行一个动作后从环境中所获得的正/负反馈信号[9]。在当前异常检测场景下, 强化学习范式各个模块定义如表 1 所示。强化学习通过在状态 s 处获取期望值 $E(R_t|s_t=s)$来引入价值函数$V^{\pi}(s)$, 表明状态 s 的价值。价值函数取决于 agent 选择 action 的策略π。在所有可能的函数中, 存在最优值函数, 表示为 $V^{*}(s)=max_{\pi}V^{\pi}(s)$, 并且最大化可实现的动作值的最优策略为$\pi^{*}$, 计算式表示为$\pi^{*}=argmax_{\pi}V^{\pi}(s)=argmax_{a}Q^{*}(s,a)$。其中$Q$ 函数的输入为状态和动作, 输出为奖励的值, Q^{*}表示 Q 函数的最优值。根据 Bellman 方程可知, 最优 Q 函数的递归定义如公式(1)所示。

$$Q^{*}(s,a)=R(s,a)+\gamma E_{s'}[V^{*}(s')] \tag{1}$$

其中 $R(s,a)$表示在状态 s 执行动作 a 后的即时奖励, $\gamma E_{s'}\left[V^{*}\left(s'\right)\right]$表示转换到下一个状态$s'$后的预期奖励。

表 1 模块定义表

Table 1 Module definition

模块	定义
Agents	代理
Environment	训练集
State	流量数据
Action	当前预测标签
Reward	当前预测所获奖励

在 Q-learning 算法中使用 Q 表格的形式记录 Q 值, 表的大小为 $m*n$, 其中 m, n 分别表示状态和动作的数量, 通过查表的方式完成一次决策过程。Q 值更新方程如公式(2)所示。

$$Q^{new}(s_t,a_t)=(1-\alpha)Q(s_t,a_t)+\alpha(r_t+\gamma max_a Q(s_{t+1},a_{t+1})) \quad (2)$$

由于在 Q-learning 算法中状态空间过大时, 会使得 Q 表格过大从而导致维度灾难。本文使用 DQN 进行优化, 在DQN算法中, 使用神经网络代替 Q 表格。该神经网络将当前状态作为输入, 每个动作的 Q 值估计作为输出。在 Q-learning 中, 目标 Q 值为:$\gamma_j+\gamma_{a'}Q(\text{ø}_{j+1}, a';\theta)$。其中$\emptyset$表示状态 s, θ表示当前神经网络中的参数。在该神经网络中目标为最小化目标 Q 值与神经网络输出的 Q 值之间的误差, 其用公式表示为: 。$Loss=(y_j-Q(\text{ø}_j, a_j;\theta))^2$ 通过梯度下降寻找最小误差, 进而训练神经网络。模型 Agent 中为一个 5 层的神经网络, 其中隐藏层共 3 层, 每层之间均采用全连接的方式。输入层共 12 个单元, 通过 PCA 主成分分析的方式将数据集中的所有特征放缩为 12, 其中输出层为数据的类别。经过 3 层隐藏层后模型的输出为四个 action 所对应的 Q 值。

3.2 交互机制

在本文基于强化学习算法的交互式引导异常检测模型中, 模型分别接收来自环境和运维人员的反馈奖励。奖励方程如公式(3)所示。

$$R_{all}(s_t,a_t)=\delta R_A(s_t,a_t)+(1-\delta)R_E(s_t,a_t) \quad (3)$$

其中 $R_E(s_t,a_t)$ 表示模型接收的来自环境的奖励, $R_A(s_t,a_t)$表示模型接收到的来自运维人员的反馈奖励, δ表示环境奖励和反馈奖励之间的权重比值。

根据模型判断结果的正确率等指标, 动态调节δ数值, 从而调节运维人员反馈奖励的占比, 当模型在异常检测过程中表现较好时, 需要适当减小δ数值, 从而降低运维人员反馈奖励对于模型总奖励的占比; 当模型表现较差时, 需要适当增加δ数值, 提升运维人员反馈奖励对于模型总奖励的占比, 从而增大运维人员的引导强度。训练的过程中, 通过运维人员的引导可以帮助异常检测模型收敛速度加快, 并且在交互引导的过程中可以使得模型动态改变其对于异常数据的认知, 从而加强模型推断运维人员意图的能力。

4 博弈模型和均衡解

4.1 博弈要素提取

博弈论作为对抗研究的数学理论, 参与博弈的个体必须根据对手的选择得到自身选择的最佳方案, 以获得最大利益[10]。一个完整的博弈过程应该由参与者、策略、收益和收益函数组成[11]。博弈过程中各个变量通过数学符号表示, 各符号含义如表 2 所示。在运维人员和模型博弈的过程中, 博弈模型中的主要要素组成如下。

表 2 符号含义表

Table 2 Symbol meaning

符号	含义
M	模型
A	运维人员
m_j	模型策略
a_k	运维人员策略
p_j	模型选择策略m_j的概率
q_k	运维人员选择策略a_k的概率
c_m^j	模型选择交互策略 j 的成本
c_a^k	运维人员选择反馈策略 k 的成本
P	失败方所受惩罚
R	整个系统的总资源
J	模型策略数量
K	运维员策略数量
U_M	模型的效用函数
U_A	运维人员的效用函数
η	运维人员对模型交互的先验信念
η_{new}	先验信念的更新

(1) 参与者: 由运维人员和模型组成。他们根据对手的行动理性地决定自己的下一步策略, 从而将自身收益最大化[12]。

(2) 策略: 参与者在博弈过程中可选择的行动。模型 M 的策略集为$[m_0,m_1,m_2]$, 其中m_0表示不发起交互, m_1表示发起交互且自身将该数据判断为异常数据, m_2表示发起交互且自身将该数据判断为正常数据; $[p_0,p_1,p_2]$分别表示模型策略的概率分布。其中$p_0+p_1+p_2=1$。运维人员 A 的策略集是$[a_0,a_1,a_2]$, 其中a_0表示不反馈, a_1表示反馈且反馈内容为数据类别为正常数据, a_2表示反馈并且反馈内容为数据类别为异常数据。$[q_0,q_1,q_2]$分别表示运维人员策略

的概率分布。其中$q_0+q_1+q_2=1$。

(3) 收益: 参与者博弈过程中可获得的收益。当模型选择交互, 运维人员选择给予反馈并且模型所判断的数据类别和运维人员反馈结果中的数据类别不一致时, 记作模型在单步博弈过程中失败, 模型的收益等于交互的成本c_m^j加上给予模型相应的惩罚P。运维人员的收益为系统总资源R减去给予模型反馈所需要的成本c_a^k; 当模型选择交互, 运维人员选择给予反馈并且模型所判断的数据类别和运维人员反馈结果中的数据类别一致时, 记作模型在单步博弈过程中胜出, 模型的收益等于系统总资源R减去交互的成本c_m^j, 运维人员的收益等于相应的惩罚P加上反馈所需要的成本c_a^k; 当模型选择交互并且运维人员选择不给予反馈, 模型胜出, 其收益为系统总资源R减去交互所需要的成本c_m^j, 运维人员的收益为相应的惩罚P。

当模型选择不进行交互, 运维人员也没有主动向模型反馈时, 模型和运维人员的收益都为 0; 运维人员认为模型对于某条流量预测结果不正确, 主动反馈但其反馈结果不正确时, 运维人员的收益为$-c_a^k$, 模型的收益为 0; 运维人员认为模型对于某条流量预测结果不正确, 主动反馈并且反馈结果正确时, 运维人员的收益为$R-c_a^k$, 模型的收益为 0。收益矩阵如表 3 所示。

表 3 运维人员-模型收益矩阵

Table 3 Admin-model payoff matrix

运维人员 \ 模型	发起交互	不发起交互
反馈	$R-c_a^k$; $-P-c_m^j(j=k)$	$R-c_a^k;0(j=k)$
	$-P-c_a^k$; $R-c_m^j(j\neq k)$	$-c_a^k;0(j\neq k)$
不反馈	$-P;R-c_m^j$	0;0

(4) 收益函数。在博弈过程中, 博弈参与者的策略受根据博弈要素构造的收益函数所支配。模型的收益函数。当模型选择不发起交互时, 模型的收益一定为 0。即$U_M(m_0)=0$。当模型选择交互时($1\leqslant j\leqslant N$), 其收益函数如等式(4)所示。

$$U_M(m_j)=\eta\begin{bmatrix}-q_{k=j}(p+c_m^j)+\\(1-q_{k=j})(R-c_m^j)\end{bmatrix}+(1-\eta)*0 \tag{4}$$

综上所述, 模型总收益如等式(5)所示。

$$\begin{aligned}U_M&=U_M(m_0)+\sum_{j=1}^{N}U_M(m_j)=\\&\eta\sum_{j=1}^{N}\left[-q_{k=j}(P+c_m^j)+(1-q_{k=j})(R-c_m^j)\right]\end{aligned} \tag{5}$$

● 运维人员的收益函数。当运维人员选择不反馈时, 其收益函数如等式(6)所示。

$$\begin{aligned}&U_A(a_0)=\eta\left[\sum_{j=1}^{N}U_A(a_0,m_j)+U_A(a_0,m_0)\right]+\\&(1-\eta)U_A(a_0,m_0)=\eta\left[\sum_{j=1}^{N}p_j(-P)\right]\end{aligned} \tag{6}$$

当运维人员选择反馈时($1\leqslant k\leqslant N$), 其收益函数如等式(7)。

$$\begin{aligned}U_A(a_k)=\eta&\left[\sum_{j=1}^{N}U_A(a_k,m_j)+U_A(a_k,m_0)\right]+\\&(1-\eta)U_A(a_k,m_0)\end{aligned} \tag{7}$$

将表 3 的数据代入等式(7)可得等式(8)。

$$\begin{aligned}U_A(a_k)=\eta&\begin{Bmatrix}p_{j=k}(R-c_a^k)+\sum_{k\neq j}p_j(-P-c_a^k)\\+p_0\left[q_{k=j}(R-c_a^k)-(1-q_{k=j})c_a^k\right]\end{Bmatrix}+\\&(1-\eta)\left[q_{k=j}(R-c_a^k)-(1-q_{k=j})c_a^k\right]\end{aligned} \tag{8}$$

4.2 确定均衡态

纳什均衡是指由所有参与者的最优策略组成的策略组合。在这种策略组合中给定其他参与者的策略没有任何单个参与者有积极性选择其他策略打破当前状态下的均衡[13]。

定理 1 在完全信息博弈中, 如果在每个给定信息下, 只能选择一种特定策略, 这个策略为纯策略(Pure strategy)[14]。对于参与者i的策略集$A_i=\{a_{i1},a_{i2},...,a_{in}\}$, 则纯策略$a_i\in A_i$。

定理 2 混合策略是指参与者可以在给定信息下, 在一定概率分布中随机选择不同动作的策略, 是其策略空间内的一种概率分布[15]。混合策略给每个纯策略分配一个概率, 参与者的策略集就是一个“样本空间”。用ΔA_i表示A_i上的概率分布, 即: $\Delta A_i=\{p_i=\{p_{i1},p_{i2},\cdots,p_{in}\},p_{ij}\geqslant 0,\sum_j p_{ij}=1\}$

那么, 混合策略$p_i=(p_{i1},p_{i2},\cdots,p_{in})\in\Delta A_i$。

定理 3 纳什均衡存在性定理, 即有限的策略式博弈一定存在混合策略纳什均衡。

因为模型是否发起交互是不确定的, 所以当前博弈模型中没有纯纳什均衡策略, 只有混合纳什均衡策略。根据定理 3 可知, 上述博弈模型中一定存在一个纳什均衡态。双方的混合纳什均衡策略集是$\{[p_0,p_1,p_2];[q_0,q_1,q_2]\}$, 即运维人员和模型都会在博弈的某个阶段随机选择概率不同的行动。

1. 运维人员的混合策略。根据模型的收益函数可知, 运维人员的混合策略$[q_0,q_1,q_2]$还有待确定。在混合策略纳什均衡中, 运维人员选择策略概率分布使模型不会偏好于任何行动, 即选择每一个策略都会得到相同

的收益。则由$U_M(m_j)=U_M(m_0)$可得等式(9)。

$$\eta\left[-q_{k=j}(P+c_m^j)+(1-q_{k=j})(R-c_m^j)\right]=0 \tag{9}$$

$$q_{k=j}=\frac{R-c_m^j}{R+P}(1\leqslant j\leqslant N) \tag{10}$$

$$q_0=1-\sum_{j=1}^{N}q_j \tag{11}$$

由等式(10)可知, 运维人员选择某种行为的概率与模型某种行为的成本和失败者所受的惩罚有关。由于行为成本越高, 惩罚力度越大, 模型选择该行为的可能性越小, 所以运维人员选择该行为的可能性也越小。

(2) 模型的混合策略。根据运维人员的收益函数可知, 模型的混合策略$[p_0,p_1,p_2]$还有待确定。在混合策略纳什均衡中, 模型选择策略概率分布使运维人员不会偏好于任何行动, 即选择每一个策略都会得到相同的收益。则由等式$U_A(a_k)=U_A(a_0)$得等式(12)。结合等式(10)进而可得$p_{k=j}$如等式(13)所示。

$$\begin{aligned}&\eta\left[\begin{array}{l}p_{j=k}(R-c_a^k)+\sum_{k\neq j}p_j(-P-c_a^k)+\\\left(1-\sum_{j=1}^{N}p_j\right)\left[q_{k=j}(R-c_a^k)-(1-q_{k=j})c_a^k\right]\end{array}\right]\\&+(1-\eta)\left[q_{k=j}(R-c_a^k)-(1-q_{k=j})c_a^k\right]\\&=\eta\left[\sum_{j=1}^{N}p_j(-P)\right]\end{aligned} \tag{12}$$

$$p_{k=j}=\frac{\left(\frac{R-c_m^j}{R+P}\right)+c_a^k}{(1+\eta)(P+R)}(1\leqslant k\leqslant N) \tag{13}$$

$$p_0=1-\sum_{j=1}^{N}p_j \tag{14}$$

从等式(10)可以看出, 模型选择是否交互的概率与相应交互成本和运维人员关于模型交互的信念η等因素有关。模型发起交互的交互成本越大, 交互的可能性就越小; 运维人员反馈所需要的成本越大, 交互的可能性越大; 其关于模型交互的信念η越大, 交互的可能性就越小。

$$\left[\begin{array}{l}\left(p_0,p_{k=j}=\frac{\left(\frac{R-c_m^j}{R+P}\right)+c_a^k}{(1+\eta)(P+R)}(1\leqslant k\leqslant N)\right);\\\left(q_0,q_{k=j}=\frac{R-c_m^j}{R+P}\right)(1\leqslant j\leqslant N)\end{array}\right] \tag{15}$$

综上所述, 运维人员和模型的混合纳什均衡策略集如等式(15)所示。即博弈参与者在每个回合的博弈中会随机选择概率不同的行动。

4.3 先验信念更新

修正先验信念η_{new}在动态贝叶斯博弈中至关重要。在下一博弈回合开始之前, 运维人员根据当前观察到的的行为来修改对于模型的先验信念, 并相应地改变下一阶段的反馈策略。假设模型交互状态为$S_M=\{0,1\}$, 其中$S_M=0$表示模型没有发起交互, $S_M=1$表示模型发起交互。运维人员始终处在待反馈状态, 即$S_A=1$。运维人员检测当前交互状态结果的正确率为P_D, 虚警率为P_F。运维人员会根据检测到的交互状态判断模型是否发起交互。模型 M 历史动作集为:$h_M(t_t)=\{m_M(t_0), m_M(t_1), m_M(t_2),\cdots,m_M(t_{t-1})\}$其中$m_M(t_t)$代表$t_t$时刻运维人员检测到的交互状态; $m_M(t_t)=\{0,1\}$, $\{0\}$表示模型在t_t时刻没有发起交互, $\{1\}$表示模型在t_t时刻发起交互。不同状态下模型在t_n时刻的可选行为的概率分布如下:

$$P(m_M(t_t)=1|S_M=1)=P_D*(1-p_0)+P_F*p_0$$

$$P(m_M(t_t)=0|S_M=1)=(1-P_D)(1-p_0)+(1-P_F)*p_0$$

$$P(m_M(t_t)=1|S_M=0)=P_F$$

$$P(m_M(t_t)=0|S_M=0)=1-P_F$$

其中 $P(m_M(t_t)|S_M)$表示当实际交互状态为S_M时, 运维人员检测到交互状态$m_M(t_t)$的概率。t_{t+1}时刻的先验信念是时刻t_t的先验信念更新。由贝叶斯公式可知, 运维人员对是否存在交互的先验信念更新η_{new}为:

$$\eta_A(S_M\mid m_M(t_t),h_M(t_t))=\frac{\eta_A(S_M\mid h_M(t_t))P(m_M(t_t)\mid S_M,h_M(t_t))}{\sum\eta_A(S_M\mid h_M(t_t))P(m_M(t_t)\mid S_M,h_M(t_t))}$$

其中η_{new}是t_{t+1}时刻运维人员的先验信念, 将指导运维人员下一阶段的策略选择。整个博弈流程如下:

1) 输入初始参数。先验信念η、模型交互成本c_m、运维人员反馈成本c_a、博弈失败时所受惩罚 P、整个交互系统的总资源 R、检测率P_D和虚警率P_F;

2) 双方根据当前先验信念计算当前阶段的最大收益, 然后输出下一阶段的策略空间A^*和D^*以及随机选择模型和运维人员的动作;

3) 根据检测到的交互情况(0 表示检测到模型没有发起交互, 1 表示检测到模型发起交互), 运维人员利用贝叶斯公式更新先验信念;

4) 双方循环进行下一回合的博弈。

5 实验评估

5.1 数据预处理

由于不同的数据集中异常数据的特征和占比等条件都有所不同, 所以仅使用一种数据集不足以说明该算法的性能表现具有普适性。为了证明该方法的通用性, 本文分别使用公共网络异常数据集和模

拟异常环境进行模型的训练和性能对比。

5.1.1 公共数据集

本文使用加拿大网络安全研究所发布的CIC-IDS2017 数据集作为公共数据集对检测模型进行训练。该数据集包含良性和最新的常见攻击, 其对网络中的数据进行长达 5 天的采集, 其中周二到周五 4 天包含了所有的异常数据[16]。以下是对于实验中涉及的攻击的简要说明:

1) 拒绝服务攻击(Denial of Service, DoS), 指通过发送不重要的信息, 以影响合法用户的使用, 占用工作机器的内存空间, 使工作机器的计算资源超载的行为。

2) 分布式拒绝服务攻击(Distributed Denial of Service, DDoS), 指处于不同位置的多个攻击者同时向一个或数个目标发动攻击, 或者一个攻击者控制位于不同位置的多台机器并利用多台机器对受害者同时实施攻击[17]。

3) 端口扫描攻击(PortScan 攻击), 是黑客用来发现网络中开放端口的常用技术。端口扫描攻击可帮助攻击者找到开放端口并确定端口是否在发送数据。它还可以揭示组织是否正在使用防火墙等主动安全设备。

本文通过合并 CIC-IDS-2017 数据集中周三以及周五下午的 Port Scan 和 DDoS 数据进行数据的集中整合, 对上述整合后的数据中每类数据进行同比例缩放后所产生的数据中各类数据占比如表 4 所示, 并采取以下方式进行数据清洗。

表 4 公共数据集类型占比

Table 4 Proportion of public dataset types

攻击类型	占比/%
Benign	61.8
DDoS	9.1
DoS	13.8
Port Scan	15.3

1) 处理缺失值和无穷值。数据集 CIC-IDS- 2017 中的缺失值仅存在于 Flow Bytes/s 特征中, 由于缺失值占比较小所以采用删除的方式对缺失值进行处理。其中无穷值存在于 Flow Bytes/s 和 Flow Packets/s 特征中, 由于无穷值无法进行正常计算, 因此对于无穷值同样进行删除处理。

2) 数据标准化。本文通过 *min-max* 标准化方法将特征按比例缩放至[0,1]区间, 从而去除数据的单位限制, 将其转化为无量纲的纯数值。转换函数公式为 $x^* = ((x - min)/(max - min))$, 其中 min 和 max 分别为该属性中的最小值和最大值。

3) 特征提取。采用 PCA 主成分分析进行特征提取, 将原始数据特征由 78 维降至 12 维。

5.1.2 模拟数据

本文在局域网内建立模拟网络拓扑, 通过虚拟机在特定的网络环境中模拟客户机和服务器之间发送数据包的过程, 将整个过程中发送的数据包作为采集对象。其中流量包中正常数据和异常数据占比如表 5 所示。

表 5 模拟数据集类型占比

Table 5 Proportion of simulated dataset types

攻击类型	占比/%
Benign	61.8
DDoS	9.1
DoS	13.8

在当前模拟环境下, 本文模拟出 DDoS 攻击, 其中模拟客户机共 9 台, 服务器共 1 台。其中 9 台客户机全部作为攻击者, 服务器作为攻击的受害者出现在网络拓扑中。在攻击者机器中, 通过 hping3 模拟 DoS 攻击, 每 10 ms 同时向服务器发送一个数据包帧, 在服务器端使用 tcpdump 对流量包数据进行采集。其中提取的数据共包含两部分, 分别是 TCP 连接相关的特征和基于时间的网络流量统计特征, 其中主要特征如表 6 所示。

表 6 数据采集的主要特征

Table 6 Main features of data collection

特征	含义
duration	连接持续时间
service	目标主机的网络服务类型 (本实验中只涉及 ftp 和 http)
flag	连接状态: 正常或错误
src_bytes	目标主机到源主机数据的字节数
wrong_fragment	错误分段的数量
rmem_defaul	默认的 TCP 数据接收窗口大小
rmem_max	最大的 TCP 数据接收窗口
count	具有相同的目标主机的连接数
serror_rate	“SYN”错误连接的百分比
rerror_rate	“REJ”错误连接的百分比
same_srv_rate	与当前连接具有相同服务的百分比
diff_srv_rate	与当前连接具有不同服务的百分比

5.2 初始参数配置

5.2.1 博弈要素参数

假设运维人员和模型每个博弈回合中, 双方资源各为 1, 总资源数为 2, 对于博弈失败方的惩罚为

0.8。数据类别不影响模型发起交互所需要的成本，所以在本文的仿真实验中，假定数据类别为正常和异常时，模型发起交互所需要的成本均为 0.8。根据等式(10)可知，$q_1 = q_2 = 0.42$ 即运维人员反馈且反馈数据类别是正常的概率和数据类别是异常为 0.42，故$q_0 = 0.16$，即当模型发起交互之后运维人员不给予反馈的概率为 0.16，可见当模型发起交互之后，运维人员为获取更多的收益，其选择反馈的概率更大。反馈中的数据类别则需要结合自身经验以及对于异常数据的把控能力进行自行判断。

对于运维人员，不同的数据类别所消耗的时间和人力成本不同，由于大多异常数据具有明显的异常特征，运维人员对其进行判断所消耗的时间和人力成本相对于正常数据较低。鉴于不同运维人员对于异常数据的把控能力以及其自身的熟练度都有所不同，所以运维人员对于不同类型数据的反馈所花费的成本，正确率和虚警率无法通过严苛的公式或者理论去推导而得。由于对于不同数据类别的反馈所消耗的成本之间的差别不会影响到最终实验结果要证明的内容，故，在本文的仿真实验中，假定当前运维人员反馈结果的正确率为 0.9，虚警率为 0.05，反馈一次正常数据所花费的成本为 0.6，反馈一次异常数据所花费的成本为 0.8，初始先验信念η设置为中间值0.5，则由等式(13)可知，$p_1 = 0.29$，即模型将自身检测类别为异常的数据发起交互的概率为 0.29，$p_2 = 0.24$，即模型将自身检测类别为正常的数据发起交互的概率为 0.24。故$p_0 = 0.47$。

5.2.2 强化学习模型参数

(1) 经验池大小。经验池目的是保证正样本被循环使用从而加速模型收敛速度，其设置过小不利于模型的收敛和训练，但是由于在本文使用的 DQN 算法中，经验池均匀采集数据样本用于模型的训练，新旧样本的采集概率相等，如果旧样本在经验池中存留时间过长，反而会阻碍模型的进一步优化。为了确保模型的收敛速度以及准确率，本文中将经验池的大小设置为 100000。

(2) 反馈奖励分配。本文中对于运维人员反馈奖励机制设计如表 7 所示。在模型中通过 Reward 奖励回馈机制将分类目标数值化和具体化，引导 Agent 中的网络模型提取数据之间的关联性并用作异常的判断，为了降低模型误判率和假阳性率，在本文实验中加大运维人员对于被模型预测为正类的正样本和被模型预测为负类的正样本的反馈奖励和惩罚力度。在加大相应奖励和惩罚力度的同时又要保证 Agent 中的网络模型最终能顺利学习到数据特征，确保算法正常收敛。

表 7 反馈奖励设计表

Table 7 Feedback reward design

预测 \ 真实	正例	反例
正例	3	–1
反例	–3	1

5.3 实验结果

5.3.1 可行性分析

为证明运维人员和模型之间建立动态贝叶斯博弈模型可以控制交互和反馈的频率，本文对两者的交互和反馈行为进行建模，图 2 反映了交互过程中模型交互策略和运维人员对于模型的先验信念变化。为使得先验信念和静默概率之间变化趋势在图上突显的更加明显，只展示了交互过程中前 200 个时间步内静默概率和先验信念的变化趋势。

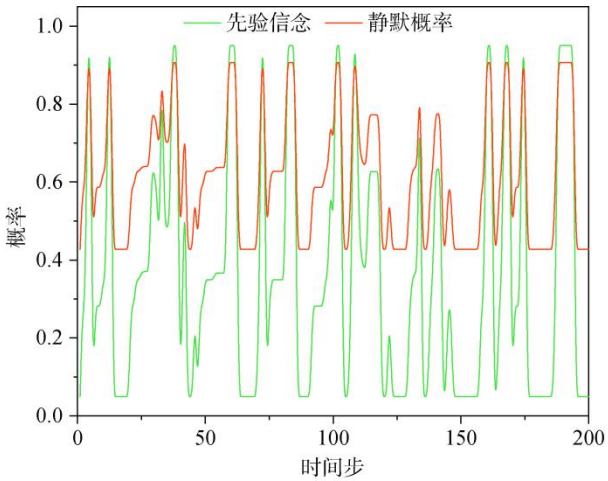

图 2 先验信念和模型策略变化图

Figure 2 Prior beliefs and interaction policy changes

其中曲线分别代表运维人员对于系统中模型的先验信念和模型不发起交互的概率，即静默概率，以及运维人员根据模型交互的历史行为对先验信念做出相应的改变。最初，运维人员认为模型发起交互的概率较低，所以先验信念较低。当模型发起交互时，先验信念的值随之增大，此时运维人员认为未来模型发起交互的可能性增加，随后维持一段时间高先验信念的状态，以用来防止模型交互时无法及时给与反馈，从而导致收益减小。在长时间未检测到模型的交互请求后，先验信念逐渐降低并保持在较低水平，在模型发起交互后再次上升。从实验结果可知，先验信念修正灵敏度较高，能够有效应对真实场景

的变化, 从而对模型和运维人员的决策产生引导的作用。从本文推出的等式(13)和(14)可知, 模型选择不发起交互的概率与先验信念η有关。

随着运维人员先验信念η的变化, 模型为保证自身的收益, 在先验信念高的状态下会增大不发起交互的概率; 在先验信念低的状态下会结合当前数据的实际情况进行交互的选择。进而使得双方的整个博弈过程中达到动态平衡。

本文在相同硬件环境, 模型参数前提下, 在 5.1 部分所合并的用于模型训练的公共数据集中随机抽出不同数量的数据集通过 8 组对比实验记录了模型发起交互的次数和运维人员反馈次数结果, 如表 8 所示。

表 8　交互和反馈次数对比

Table 8　Number of interactions and feedback

数据总数	交互次数	反馈次数	交互比例/%	反馈比例/%
28479	6317	1067	22.18	16.89
25370	5249	861	20.69	16.42
23697	5099	846	21.51	16.60
23689	5646	948	23.83	16.78
15213	3341	536	21.96	16.04
13972	2962	489	21.20	16.53
12645	2840	472	22.46	16.63
9887	2013	322	20.37	15.96

其中交互次数和数据总数之比表示为交互比例, 反馈次数和交互次数之比表示为反馈比例。结果证明了通过建立运维人员和智能体之间的博弈模型进行交互和反馈与否的控制和指导是切实可行的。由表可知, 模型发起交互的次数占比以及每次发起交互后运维人员被动反馈占比分别恒定在 0.22 和 0.16 左右, 在兼顾人力开销的同时也可以保证可以通过交互的方式提升模型的整体性能。

5.3.2　方法性能

为证明交互引导有助于模型收敛速度的加快和性能的提升, 实验在训练数据规模和迭代次数均相同的前提下, 将不同反馈比例下的 Loss 收敛结果进行对比, 对应的 4 条 Loss 曲线在前 2000 个时间步的数据通过 Y 偏移堆积制图, 如图 3 所示。其中反馈比例 $[0.08, 0.06, 0.04, 0.02]$ 对应的反馈次数为 $[6224, 4668, 3112, 1556]$, 与上文的反馈比例定义不同, 此处反馈比例指反馈次数与参与训练的数据总数之比。从图中可以得出, 在模型训练的过程中运维人员可以通过反馈给予模型一个正向引导从而使得模型更快收敛。模型收敛速度和反馈次数呈正相关关系。

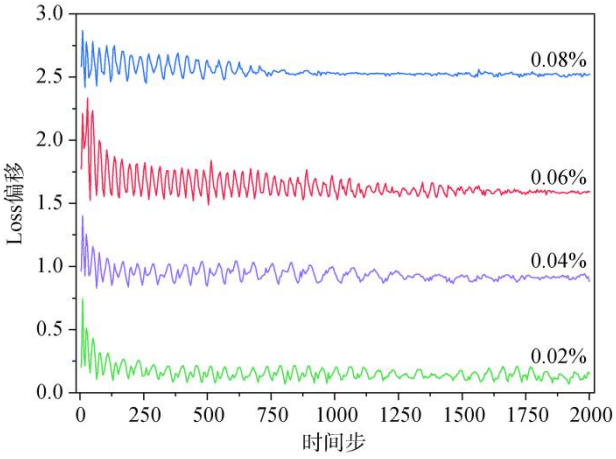

图 3　损失曲线和反馈比例

Figure 3　Loss curve and feedback ratio

运维人员反馈可以使得模型收敛速度加快的同时, 还可以提升模型整体性能。对于不同数据集下, 不同模型和不同反馈比例下当前模型的性能对比, 如图 4 所示。

其中左图为所有模型在公共数据集上的实验结果, 右图为所有模型在模拟数据集上的实验结果。由图可知, 各算法在公共数据集上的表现要优于自行在局域网内模拟攻击产生的数据集, 但是在两种数据集的同一纬度上, 各个算法的整体表现以及本文所提出的交互引导模型对于算法准确度的提升趋势是大致相同的。结合图中不同反馈次数对应结果可知, 模型的整体性能表现和反馈次数呈正相关, 运维人员的反馈在模型训练的过程中发挥了正向引导的作用。

由实验结果可知, 在传统机器学习算法中, 由于强化学习算法采样率相对较低, 所以 LSTM 算法在性能上优于 DDQN 和不引入交互机制时的 DQN 算法; 其中 AE-RL[18]算法引入动态采样的思想, 从模拟环境中通过两个 Agent 代理之间的合作, 使用 DQN 算法进行训练, 该算法获得了较好的性能; SD-CNN[19]算法将网络流转换成光谱图像后使用 CNN 进行检测, 同样取得了较好的结果。相较于传统的机器学习方法, AE-RL 和 SD-CNN 两种算法虽然都取得了较好的性能, 但是在 AE-RL 算法中, 引入两个 Agent 代理对数据进行并行训练, 其中分类器代理进行数据样本类型的预测, 环境代理决定分类器代理下一条数据样本的攻击类别, 两者在一个对抗模型中并行工作, 环境代理对分类器代理的训练样本进行类别上的筛选, 使得分类器代理更专注于

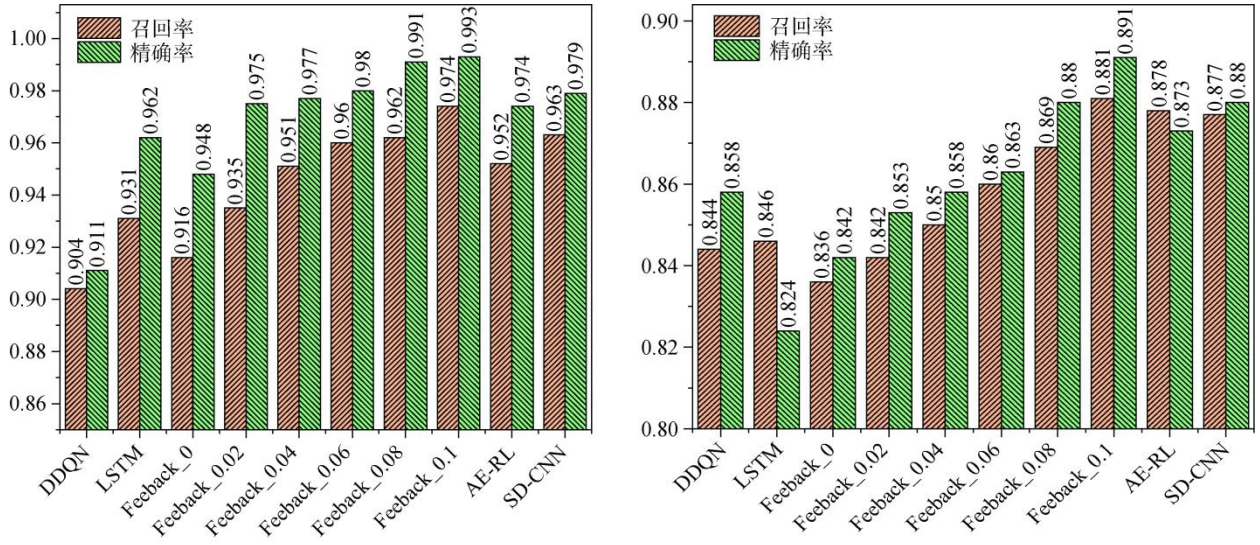

图 4　性能对比

Figure 4　Performance comparison

训练占较少, 训练较为困难的样本。由于对于不同的数据样本, 分类器代理所训练的数据的类别由环境代理所决定, 而环境代理模型要经过数轮迭代后才会产生对于某种攻击类别的提示并交由分类器代理, 所以该算法在事先标注好类别的的静态数据训练中, 可以获得较好的实验效果, 但是在全新类别的数据样本的训练和识别中, 其灵敏度较低, 并且在实验过程中需要并行训练两个算法模型, 所以整个实验过程对于硬件性能要求较高; SD-CNN 算法中, 使用短时傅立叶变换将网络流转换为网络谱图图像, 然后将该图像应用于深度卷积神经网络进行训练, 在模型训练的前期需要生成网络谱图图像, 该过程由于数据维度的转换, 所以硬件设备要求较高。并且上述方法相较于本文提出的基于交互式博弈的网络流量异常检测方法, 缺乏训练过程中的人工参与度, 整个训练过程类似一个“黑盒子”, 缺少匹配不同运维人员的专家知识和自适应动态环境的能力。

6　总结与展望

传统的以数据作为唯一驱动和依赖的机器学习模型不具备自适应动态环境的能力, 本文基于动态贝叶斯博弈理论建立运维人员和模型之间的博弈模型, 提出了一种基于博弈论的交互引导式网络流量异常检测模型。将运维人员的反馈作为奖惩信号引导模型的学习, 从而使其具备一个动态适应场景中变化的能力以及整体系统可用性。相较于传统的机器学习方法, 交互引导式模型提高了模型整体的性能以及收敛速度。

在未来的研究中, 我们希望在强化学习奖励回馈机制的基础上, 进一步优化强化学习算法模型, 提高实时网络流量环境中算法的适应能力, 提高攻击预测的准确性。

参考文献

[1] DEOKAR B, HAZARNIS A. Intrusion detection system using log files and reinforcement learning [J]. *International Journal of Computer Applications*, 2012, 45(19): 28-35.

[2] Kumar S V, Vanajakshi L. Short-Term Traffic Flow Prediction Using Seasonal ARIMA Model with Limited Input Data[J]. *European Transport Research Review*, 2015, 7(3): 1-9.

[3] Subba B, Biswas S, Karmakar S. Intrusion Detection Systems Using Linear Discriminant Analysis and Logistic Regression[C]. *2015 Annual IEEE India Conference*, 2016: 1-6.

[4] Zhang X F, Hao X H. Research on Intrusion Detection Based on Improved Combination of K-Means and Multi-Level SVM[C]. *2017 IEEE 17th International Conference on Communication Technology*, 2018: 2042-2045.

[5] Blanco R, Cilla J J, Briongos S, et al. Applying Cost-Sensitive Classifiers with Reinforcement Learning to IDS[C]. *International Conference on Intelligent Data Engineering and Automated Learning*, 2018: 531-538.

[6] Zha D C, Lai K H, Wan M Y, et al. Meta-AAD: Active Anomaly Detection with Deep Reinforcement Learning[C]. *2020 IEEE International Conference on Data Mining*, 2021: 771-780.

[7] Apruzzese G, Andreolini M, Marchetti M, et al. Deep Reinforcement Adversarial Learning Against Botnet Evasion Attacks[J]. *IEEE Transactions on Network and Service Management*, 2020, 17(4): 1975-1987.

[8] Sharafaldin I, Habibi Lashkari A, Ghorbani A A. Toward Generating a New Intrusion Detection Dataset and Intrusion

Traffic Characterization[C]. *The 4th International Conference on Information Systems Security and Privacy*, 2018: 108-116.

[9] Ma X Y, Shi W. AESMOTE: Adversarial Reinforcement Learning with SMOTE for Anomaly Detection[J]. *IEEE Transactions on Network Science and Engineering*, 2021, 8(2): 943-956.

[10] Sha K W, Gehlot J, Greve R. Multipath Routing Techniques in Wireless Sensor Networks: A Survey[J]. *Wireless Personal Communications*, 2013, 70(2): 807-829.

[11] Govindaraj L, Sundan B S, Thangasamy A. An Intrusion Detection and Prevention System for DDoS Attacks Using a 2-Player Bayesian Game Theoretic Approach[C]. *2021 4th International Conference on Computing and Communications Technologies*, 2022: 319-324.

[12] Cui H Y, Zhang Z. A Cooperative Multi-Agent Reinforcement Learning Method Based on Coordination Degree[J]. *IEEE Access*, 2021, 9: 123805-123814.

[13] Wu H, Wang W. A Game Theory Based Collaborative Security Detection Method for Internet of Things Systems[J]. *IEEE Transactions on Information Forensics and Security*, 2018, 13(6): 1432-1445.

[14] BAI F, LIU X Y, ZHANG Y L, et al. 2019. Research on Game Model of Wireless Sensor Network Intrusion Detection [M], Proceedings of the 2019 International Conference on Embedded Wireless Systems and Networks. Junction Publishing; Beijing, China: 373-378.

[15] Wang B Y, Xia Y, Zhao S P. Clustering Routing Algorithm for Wireless Sensor Network Based on Mixed Strategy Game Theory[J]. *Sensors and Materials*, 2022, 34(2): 885.

[16] Sharafaldin I, Habibi Lashkari A, Ghorbani A A. A Detailed Analysis of the CICIDS2017 Data Set[C]. *International Conference on Information Systems Security and Privacy*, 2019: 172-188.

[17] Mittal M, Kumar K, Behal S. Deep Learning Approaches for Detecting DDoS Attacks: A Systematic Review[J]. *Soft Computing*, 2023, 27(18): 13039-13075.

[18] Caminero G, Lopez-Martin M, Carro B. Adversarial Environment Reinforcement Learning Algorithm for Intrusion Detection[J]. *Computer Networks*, 2019, 159: 96-109.

[19] KHAN A S, AHMAD Z, ABDULLAH J, et al. A Spectrogram Image-Based Network Anomaly Detection System Using Deep Convolutional Neural Network [J]. *IEEE Access,* 2021, 9: 87079-87093.

张文哲 于 2020 年在湖南工业大学计算机科学与技术专业获得学士学位。现在南京理工大学计算机技术专业攻读硕士学位。研究领域为网络异常检测、强化学习。Email: wenzhe.zhang@njust.edu.cn

杨栋 于 2006 年在北京理工大学，控制理论与控制工程专业获得硕士学位，现在南京理工大学攻读博士学位。Email: yangdong21014@163.com

魏宋杰 于 2009 年在特拉华大学计算机科学与技术专业获得博士学位。现为南京理工大学计算机科学与工程学院网络与通信技术系副教授。研究领域为网络流量分析与检测，分布式系统。Email: swei@njust.edu.cn