计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210400021

DRL-IDS:基于深度强化学习的工业物联网入侵检测系统

李贝贝　宋佳芮　杜卿芸　何俊江
四川大学网络空间安全学院　成都 610041
(libeibei@scu.edu.cn)

摘　要　近年来,工业物联网迅猛发展,在实现工业数字化、自动化、智能化的同时也带来了大量的网络威胁,且复杂、多样的工业物联网环境为网络入侵者创造了全新的攻击面。传统的入侵检测技术已无法满足当前工业物联网环境下的网络威胁发现需求。对此,文中提出了一种基于深度强化学习算法近端策略优化(Proximal Policy Optimization 2.0,PPO2)的工业物联网入侵检测系统。该系统将深度学习的感知能力和强化学习的决策能力相结合,以实现对工业物联网多种类型网络攻击的有效检测。首先,运用基于 LightGBM 的特征选择算法筛选出工业物联网数据中最有效的特征集合;然后,结合深度学习算法将多层感知器网络的隐藏层作为 PPO2 算法中的价值网络和策略网络的共享网络结构;最后,基于 PPO2 算法构建入侵检测模型,并使用 ReLU(Rectified Linear Unit)进行分类输出。在美国能源部橡树岭国家实验室公开发布的工业物联网真实数据集上开展的大量实验表明,所提出的入侵检测系统在检测对工业物联网的多种类型网络攻击时,获得了 99.09%的准确率,且在准确率、精密度、召回率、F1 评分等指标上均优于目前基于 LSTM,CNN,RNN 等深度学习模型和 DDQN,DQN 等深度强化学习模型的入侵检测系统。

关键词:工业物联网;网络安全;入侵检测系统;深度强化学习;PPO2 算法
中图法分类号　TP393

DRL-IDS:Deep Reinforcement Learning Based Intrusion Detection System for Industrial Internet of Things

LI Bei-bei,SONG Jia-rui,DU Qing-yun and HE Jun-jiang
School of Cyber Science and Engineering,Sichuan University,Chengdu 610041,China

Abstract　In recent years,the Industrial Internet of Things (IIoT) has developed rapidly. While realizing industrial digitization, automation,and intelligence,the IIoT has introduced tremendous cyber threats. Further,the complex,heterogeneous,and distributed IIoT environment has created a brand-new attack surface for cyber intruders. Traditional intrusion detection techniques no longer fulfill the needs of intrusion detection for the current IIoT environment. This paper proposes a deep reinforcement learning algorithm (i.e.,Proximal Policy Optimization 2.0,PPO2) based intrusion detection system for the IIoT. The proposed intrusion detection system combines the perceptual ability of deep learning with the decision-making ability of reinforcement learning, which can effectively detect multiple types of cyber attacks for the IIoT. First,a LightGBM-based feature selection algorithm is used to filter the most effective feature sets in IIoT data. Then,the hidden layer of the multilayer perceptron network is used as the shared network structure of the value network and policy network in the PPO2 algorithm. At last,the PPO2 algorithm is used to construct the intrusion detection model and ReLU (Rectified Linear Unit) is employed for classification output. Extensive experiments conducted on a real IIoT dataset released by the Oak Ridge National Laboratory,sponsored by the U.S. Department of Energy,show that the proposed intrusion detection system achieves 99.09% accuracy in detecting multiple types of network attacks for the IIoT,and it outperforms state-of-the-art deep learning models (e.g.,LSTM,CNN,RNN) based and deep reinforcement learning models (e.g.,DDQN and DQN) based intrusion detection systems,in terms of the accuracy,precision,recall,and F1 score.

Keywords　Industrial internet of things,Cyber security,Intrusion detection system,Deep reinforcement learning,PPO2 algorithm

---

到稿日期:2021-03-31　返修日期:2021-04-28
基金项目:国家重点研发计划项目(2020YFB1805400);国家自然科学基金(U19A2068,62002248);中国博士后科学基金(2019TQ0217,2020M673277);四川省重点研发项目(20ZDYF3145);中央高校基本科研业务经费(YJ201933)
This work was supported by the National Key Research and Development Program of China(2020YFB1805400), National Natural Science Foundation of China(U19A2068,62002248),China Postdoctoral Science Foundation(2019TQ0217,2020M673277),Provincial Key Research and Development Program of Sichuan(20ZDYF3145) and Fundamental Research Funds for the Central Universities(YJ201933).
通信作者:何俊江(hejunjiang@stu.scu.edu.cn)

1 引言

工业物联网是物联网技术在工业领域的应用，其本质是工业自动化和物联网技术的融合发展。工业物联网实现了生产、监控、管理等子系统的空前结合，不同的系统在控制中心的统一管理下可以更加高效地完成对各类工业数据的处理，其高复杂性、开放性的特点增加了工业物联网面临的网络安全风险。国家互联网应急中心 CNCERT 发布的《2019 年我国互联网网络安全态势综述》指出，存在高危漏洞隐患的工业物联网设备占比约 41%。通过监测发现，电力、石油天然气、城市轨道交通等重点行业暴露的联网监控管理系统有 2 249 套，其中电力有 653 套、石油天然气有 584 套、城市轨道交通有 100 套[1]。

工业物联网是一个复杂的网络，系统的一部分出现任何故障或异常都可能在短时间内对整个系统造成巨大损害。因此，尽早发现网络攻击对于及时有效的网络响应至关重要。入侵检测系统(Intrusion Detection System，IDS)是网络安全防护的重要组成部分，能够帮助系统有效发现网络入侵行为。然而，近年来，由于工业物联网的运行环境和结构持续变化，传统的入侵检测模型(如基于简单机器学习的入侵检测模型等)往往不具备对网络威胁的自适应调整能力，不能在工业物联网的网络风险环境发生变化时动态地调整自身的辨识策略，进而无法针对复杂网络攻击提供自适应的检测、响应和防御等。

深度强化学习 (Deep Reinforcement Learning，DRL)算法可有效解决工业物联网环境中存在的不确定性等问题，它通过智能体这一强化学习的载体探索，利用未知环境，并结合自身的经验进行学习[2-3]。深度强化学习将强化学习的决策能力和深度学习的感知能力相结合，利用深度学习将大规模原始输入数据进行简单且非线性的变换，转换为更高层次的抽象表达，从而进一步发现数据的内在规律；利用强化学习，基于反馈信号学习和基于试错学习的模式，在没有指导信息的情况下，从与环境的交互过程中找到可能的最佳方案[4-5]。

针对上述问题，本文提出了一种新的基于深度强化算法 PPO2 的面向工业物联网环境的入侵检测系统 DRL-IDS。该系统运用基于 LightGBM 的特征选择算法提取出最有效的特征集合，可有效减小模型的计算复杂度；用包含 3 层隐藏层的多层感知机作为该入侵检测系统中价值网络和策略网络共享的深度神经网络结构，构建基于深度强化学习 PPO2 算法的入侵检测系统，该系统最后用 ReUL 函数在减少过拟合的情况下作为分类输出。本文提出的入侵检测系统在美国能源部橡树岭国家实验室公开发布的工业物联网真实数据集上进行了大量实验，结果表明，该入侵检测系统在检测对工业物联网的多种类型的网络攻击时，获得了 99.09%的准确率，且在准确率、精密度、召回率、F1 评分等指标上均优于目前基于长短期记忆网络(Long Short-Term Memory，LSTM)、卷积神经网络(Convolutional Neural Networks，CNN)、循环神经网络(Recurrent Neural Network，RNN)等深度学习模型和深度双 Q 网络(Double Deep Q-Learning Network，DDQN)、深度 Q 网络(Deep Q-Learning Network，DQN)等深度强化学习模型的入侵检测系统。

2 相关工作

2.1 面向工业物联网的入侵检测系统

入侵检测系统在传统工业控制系统以及现代工业物联网中均有着广泛的应用[6]。2018 年，AL-Hawawreh 等[7]提出了一种基于深度学习模型的 IICS 异常检测技术，该模型可以使用从 TCP/IP 数据包收集的信息进行学习和验证。同年，Roy 等[8]使用 BiLSTM-RNN 检测了工业物联网攻击，并采用新型 UNSWNB15 数据集对多层深度神经网络进行了训练。2019 年，Yang 等[9]设计了一种基于 CNN 面向数据采集与监视控制系统 (Supervisory Control And Data Acquisition，SCADA)的网络入侵检测系统，以保护工业物联网免受 DDoS 等常规网络攻击以及针对 SCADA 的特定网络攻击。2020 年，Ismail 等[10]研究了智能电网中的电力盗窃攻击，并针对这种网络攻击提出了一种基于深度学习的入侵检测系统。同年，Li 等[11]提出了一种名为 DeepFed 的联邦深度学习方案，用于检测和缓解对分布式工业物联网的网络威胁。上述方法在面对当前高实时、大容量以及复杂多维的工业物联网数据时，往往需要复杂的训练过程，且准确率有待提高。

2.2 基于深度强化学习的入侵检测技术

深度强化学习不仅能够从已知网络环境中获取最大化的奖励，而且其具备探索功能，能够自动在网络环境中挖掘更多有价值的信息，模型收敛速度快。2018 年，Kurt 等[12]提出了一种基于强化学习的入侵检测系统，用于监视和分析传感器网络，并将其提出的系统与基于自适应机器学习的入侵检测系统和集群混合入侵检测系统进行了比较。2019 年，Sethi 等[13]提出了一种上下文自适应的入侵检测系统，该系统使用分布在网络上的多个独立的深度强化学习智能体来增强针对新型复杂网络攻击的检测准确性。同年，Otoum 等[14]提出了一种基于无模型强化学习的部分可观测马尔可夫决策过程(Partially Observable Markov Decision Processes，POMDPs)在线网络攻击检测算法。2019 年 5 月，Caminero 等[15]提出了一种用于入侵检测系统的最新多主体对抗增强学习模型。基于深度强化学习的入侵检测系统在训练效率、准确率等方面还存在进一步提升和优化的空间。2020 年，Hsu 等[5]提出了一种基于深度强化学习的异常网络入侵检测系统，该系统具有自我更新的能力，可有效检测新型恶意网络流量行为。

3 基于深度强化学习的入侵检测系统

3.1 入侵检测系统总体框架

本文提出的基于 PPO2 的入侵检测系统 DRL-IDS 主要由 3 个部分组成，分别为数据处理模块、入侵检测智能体 agent)构造模块以及入侵检测智能体训练模块(见图 1)。其中，数据处理模块主要包括特征选择和数据预处理；入侵检测智能体构造模块主要包括确定强化学习的环境状态模型、价值函数构造、训练策略定义。环境状态模型是环境的私有呈现，包括了环境用来决定奖惩的规则，对入侵检测智能体不可见。训练策略通过评估价值函数，来进一步优化入侵检测智能体的动作决策策略，用环境状态模型反馈的奖惩(损失函

数）来更新训练策略中的参数，入侵检测智能体的训练模块通过损失函数不断地优化策略，直到模型收敛或者完成指定的训练步长时得到最终的 DRL-IDS 模型，最后用混淆矩阵测试来评估该模型的性能表现。

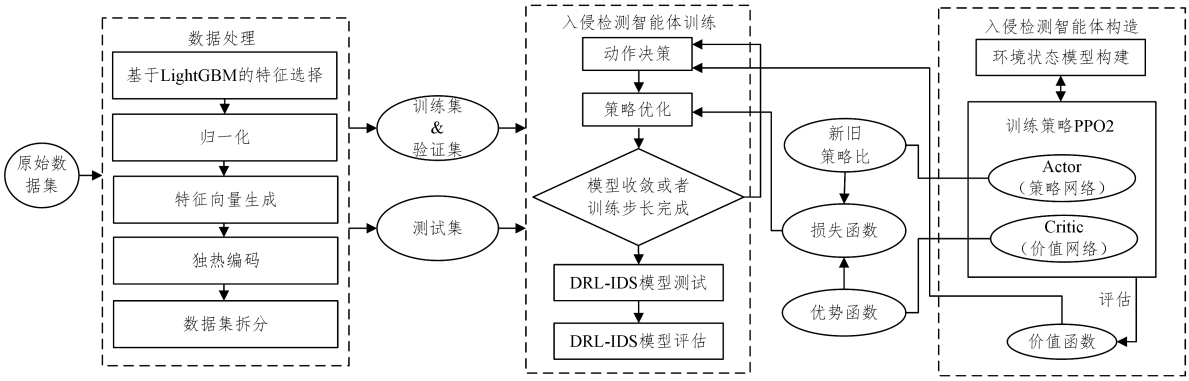


图 1 入侵检测系统 DRL-IDS 框架图

Fig. 1 Framework of the proposed DRL-IDS

3.2 数据处理模块

为了减少原始数据的噪声冗余，提高模型的多分类检测精度，本文的入侵检测系统首先进行特征选择，在保证入侵检测性能的前提下有效降低数据的冗余维度。该入侵检测系统用基于嵌入式的特征选择算法 LightGBM 进行特征的筛选[16]，具体策略如下：

1）删去缺失值大于阈值 60%的特征。根据经验得出，当一个特征的缺失率大于 60%时，这个特征对于入侵检测智能体的训练意义甚微。

2）删去唯一值的特征。

3）删除每个强相关特征对中的任意一个特征。定义强相关特征对的具体阈值（绝对值）为 Pearson 相关系数[17] 0.99。

4）删去 LightGBM 算法得出特征的重要性排名较低的特征。最终特征数量按照特征重要性评分的高低顺序逐步增加特征，直到模型性能不再提高为止。

5）将所有在不同区间的变量归一化，使用最简单的 min－max 函数缩放范围特征值到[0,1]区间，具体公式如下：

$$x'=\frac{x-\min(x)}{\max(x)-\min(x)} \tag{1}$$

其中，x 为原始值，x'为规范化值。

6）生成特征向量，最后进行独热编码。

3.3 入侵检测智能体构造

本节描述构成入侵检测智能体的组成部分：环境状态模型、价值函数以及训练策略。

3.3.1 环境状态模型

使用真实的工业物联网数据集来模拟网络流量环境，构成入侵检测系统的智能体构造所需要的环境，其包括 3 个部分：

1）用来接收入侵检测系统输出的管理员，主要体现在环境对智能体的反馈过程中。

2）普通网络用户，网络用户的流量是通过现有数据集的流量来模拟的。

3）攻击者，制造多种恶意攻击。

入侵检测智能体能够感知环境的状态，并且根据环境提供的反馈信号 r_t（时间步 t），通过学习选择一个合适的动作，来最大化未来积累奖励，即从当前时间步 t 开始，直到最终状态的奖励 $r_{t,n}$ 的累积和为 $R_t=r_{t,1}+r_{t,2}+\cdots+r_{t,n}$。由于入侵检测环境是随机的或者未知的，这意味着下一个状态可能也是随机的，从而使得入侵检测智能体采取的动作具有随机性，随着步数的增加，不确定性会随之增大，获得相同奖励的可能性降低。为了降低不确定性和随机性，使用折扣因子来减小步骤之间的强关联性，用折扣未来累积奖励 G_t 来代替未来奖励。时间步长 t 的总折扣未来累积奖励表达式为：

$$G_t=R_{t+1}+\gamma R_{t+2}+\cdots+\gamma^k R_{t+k+1}=\sum_{k=0}^{\infty}\gamma^k R_{t+k+1} \tag{2}$$

其中，$\gamma\in[0,1]$是折扣系数，用于提升即时奖励而不是延迟奖励。

1）γ 接近 0，代表更加重视当下的回报。

2）γ 接近 1，代表更加重视未来的回报。

由于入侵检测的网络流量之间是离散的、相互独立的[18]，因此在实验中 γ 应该尽可能接近 0，使得网络流量之间的连续性减弱。

入侵检测智能体与环境交互反馈的规则如下：

1）当入侵检测系统成功检测出攻击，并且成功分类此攻击的类型时，给予正反馈 r_t+1 。

2）当错失了一个攻击或者成功检测出攻击，但是错误地分类了该攻击的类型时，给予负反馈 r_t-1。

3）当流量是正常流量并且没有发出警报时，没有反馈。

本文用实数的加减 1 来定义奖惩的规则，从而达到入侵检测智能体训练的目的。所有能够影响环境从而产生奖惩的数据都被认为是环境状态的一部分，DRL-IDS 的智能体与环境交互过程中基于网络流量环境的反馈规则，构成了环境状态模型。

3.3.2 价值函数构造

价值函数是奖励的期望，主要用于评价不同状态的好坏，指导智能体动作的选择。能够影响智能体做出下一个动作决策的数据是智能体状态的一部分。用价值函数来评估入侵检测智能体在某时间 t、状态 s 的好坏程度。定义 $Q_\pi(s,a)$ 为动作值函数，$V_\pi(s)$ 为状态值函数，前者用于评估当前智能体从状态 s 开始，执行动作 a 且服从策略 π 的期望返回（见式(3)），后者表示智能体在状态 s 下，执行动作 a 得到的奖励期望（见式(4)）。

$$Q_\pi(s,a)=\mathbb{E}_\pi[G_t \mid S_t=s, A_t=a] \tag{3}$$

$$V_\pi(s)=\mathbb{E}_\pi[G_t \mid S_t=s] \tag{4}$$

该入侵检测系统框架中入侵检测智能体的动作空间为正数离散值，“0”表示预判为正常流量，“1,2,…,n”表示 n 种类型的攻击。马尔可夫决策过程定义了入侵检测智能体在进行动作决策过程中的状态值函数和动作值函数，之后，通过贝尔曼方程对状态值函数或动作值函数进行形式化表示，完成入侵检测智能体的动作决策过程。

马尔可夫决策过程具有马尔可夫性质，即在时间步 $t+1$ 时，环境的反馈仅取决于上一时间步 t 的状态和动作 a，与时间步 $t-1$ 以及 $t-1$ 步之前的时间没有相关性，而系统的下一个状态只与当前状态有关，因此可以简化入侵检测智能体的决策过程。该系统的马尔可夫决策过程是一个五元组，S 为状态空间集，A 为动作空间集，P_{sa} 表示状态转移概率（在状态 s 下执行动作 a 后，转移到另一个状态 s' 的概率分布并带有动作奖励，记作 $P(s',r|s,a)$），R 为奖励函数，γ 为折扣因子。

$$\begin{aligned} &MDP=(S,A,P_{sa},R,\gamma)\\ &S=\{S_1,S_2,\cdots,S_n\}\\ &A=\{A_1,A_2,\cdots,A_n\} \end{aligned} \tag{5}$$

贝尔曼方程将及时奖励 Rt 和未来状态的折扣值 γ、时间步 $t+1$ 的状态值 $V(S_{t+1})$ 相加，反映了当前状态下的状态值函数 $V(S_t)$ 和下一时刻状态值函数 $V(S_{t+1})$ 之间的关系。

$$\begin{aligned} V(s)&=\mathbb{E}[G_t \mid S_t=s]\\ &=\mathbb{E}[R_{t+1}+\gamma G_{t+1} \mid S_t=s]\\ &=\mathbb{E}[R_{t+1}+\gamma V(S_{t+1}) \mid S_t=s] \end{aligned} \tag{6}$$

类似地，可以得出动作值函数 $Q(s,a)$：

$$\begin{aligned} Q(s,a)&=\mathbb{E}[R_{t+1}+\gamma V(S_{t+1}) \mid S_t=s, A_t=a]\\ &=\mathbb{E}[R_{t+1}+\gamma\,\mathbb{E}_{a\sim\pi}Q(S_{t+1},a) \mid S_t=s, A_t=a] \end{aligned} \tag{7}$$

贝尔曼方程的递归更新，可分解为动作值函数和状态值函数。当进行下一个动作时，两个值函数分别遵循策略 π 更新值函数，其中 $P^a_{ss'}$ 代表状态转换概率。贝尔曼方程的递归更新公式如下：

$$V_\pi(s)=\sum_{a\in\mathcal{A}}\pi(a \mid s)(R(s,a)+\gamma\sum_{s'\in S}P^a_{ss'}V_\pi(s')) \tag{8}$$

$$Q_\pi(s,a)=R(s,a)+\gamma\sum_{s'\in S}P^a_{ss'}\sum_{a\in\mathcal{A}}\pi(a' \mid s')Q_\pi(s',a') \tag{9}$$

入侵检测智能体用状态值函数计算当前策略下每个状态的价值，用动作值函数计算每个状态下分别做出不同动作的值。选取当前状态下使价值函数取得最大值的动作，进行策略优化。

3.3.3 训练策略定义

入侵检测智能体的训练策略是状态到动作的映射。DRL-IDS 运用的算法 PPO2 由 DeepMind[1] 和 OpenAI[2] 提出。该算法源于区域最佳算法 TRPO，相比 TRPO，PPO2 更简单、更通用，复杂度也更低。PPO2 的主要贡献是简化了 Kull-back-Leibler 惩罚系数的数学运算过程。作为一种基于策略梯度的方法，其特性是直接训练随机模型或者是神经网络，相比于另一种深度强化学习算法 A3C[20] 的误差函数需要针对目标策略取对数后进行优化，PPO2 的误差函数则主要是评估新旧策略的比例；与 Q-Learning 不同，PPO2 不使用经验缓冲区来存储过去的经验，而是在线地从环境中学习。

策略梯度方法是计算策略梯度的估计量，并将其插入到随机梯度提升算法中，通过对策略参数 θ 进行随机梯度上升，计算策略梯度损失来更新策略网络的参数。

$$L^{PG}(\theta)=\hat{\mathbb{E}}_t[\log\pi_\theta(a_t \mid s_t)\hat{A}_t] \tag{10}$$

$$A^\pi(s,a)=Q^\pi(s,a)-V^\pi(s) \tag{11}$$

其中，$\hat{A}_t$ 是时间步长 t 的优势函数估计量。当 $\hat{A}_t$ 是正数、梯度为正时，应当增加这些动作的概率，反之，应当降低这些动作的概率。期望 $\hat{\mathbb{E}}_t$ 表明了有限批次样本的经验平均，当采用策略 π_θ 时，一般来说，神经网络把从环境中观察到的状态作为输入，采取的行动作为输出，$\log\pi_\theta$ 是策略网络输出的概率对数。

为了防止入侵检测智能体训练时震荡幅度过大，PPO2 引入了目标函数（clipped surrogate function）来约束新旧策略的更新比例，实现在多个步骤中小批量更新。定义 $r_t(\theta)=\dfrac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}$ 为新旧策略比例，保守政策迭代（CPI）损失的表达式为：

$$L^{CPI}(\theta)=\hat{\mathbb{E}}_t[r_t(\theta)\hat{A}_t] \tag{12}$$

如果没有约束，CPI 的最大化会导致梯度爆炸，使用 Schulman 等[21] 提出的目标可以惩罚那些比例远离 1 的改变。剪切函数如下：

$$L^{CLIP}(\theta)=\hat{\mathbb{E}}_t[\min(r_t(\theta)\hat{A}_t,\text{clip}(r_t(\theta),1-\varepsilon,1+\varepsilon)\hat{A}_t)] \tag{13}$$

$$L^{CLIP}_{\theta_k}(\theta)=\underset{\tau\sim\pi_k}{E}[\sum_{t=0}^{T}\min(r_t(\theta)\hat{A}^{\pi_k}_t,\text{clip}(r_t(\theta),1-\varepsilon,1+\varepsilon)\hat{A}^{\pi_k}_t))] \tag{14}$$

通过 $\theta_{k+1}=\arg\max_\theta L^{CLIP}_{\theta_k}(\theta)$ 对策略进行更新，其中 $\varepsilon=0.2$ 是超参数，最小值内的第一项是 CPI，第二项通过剪切比例来修改替代目标，这将保证 r_t 在对应刺激区间 $(1-\varepsilon,1+\varepsilon)$。min 函数使得最终目标是 CPI 的下界。当且仅当目标改善时忽略比例，当使得目标变差时考虑比例。

3.4 入侵检测智能体训练

图 2 给出了入侵检测智能体与环境交互的过程，$t=0,1,2,3,\cdots$，智能体观察到环境中的部分或者是全部状态，$s_t\in S$，然后做出动作的决策 $a_t\in A(s)$，在每一次动作完成后，会更新环境的状态 s_{t+1} 和得到一个刚才完成动作的反馈结果 $r_{t+1}\in R\subset\mathbb{R}$，反馈的好坏会间接影响以后的决策。在这个入侵检测框架中，用数据集中的每条网络流量数据来模拟真实工业物联网中产生的网络流量数据，用价值函数模拟现实生活中的网络管理员，评判入侵检测智能体做出的动作的好与坏。

入侵检测智能体训练过程分为 3 个阶段。第一阶段，入侵检测智能体从环境中得到策略，并学习得到环境状态模型；第二阶段，入侵检测智能体根据策略评估相应的价值函数；第三阶段，入侵检测智能体根据价值函数判断当前状态应该做出的动作，并根据优势函数计算回报，最后进一步优化策略直

1) http://deepmind.com/

2) http://openai.com/

到神经网络的训练收敛，训练结束后得到成熟的入侵检测智能体。

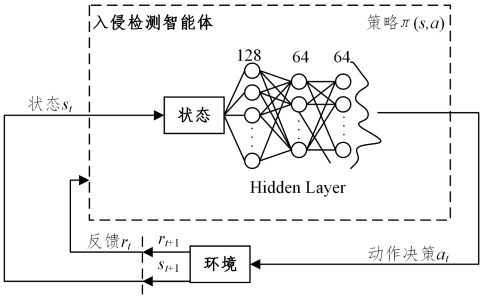


图 2 DRL-IDS 入侵检测强化学习过程

Fig. 2 Reinforcement learning process of DRL-IDS

本文使用的 PPO2 算法基于演员-评论家风格（Actor-Critic Style[21]），入侵检测智能体作为演员（actor）训练策略网络。以最大化未来累积奖励 G 为目标，预测出目前应当执行的策略。价值函数通过价值网络（critic）来评价策略网络的输出。为了使神经网络更新及时，需要融合策略网络和价值网络，让两个网络的权重进行共享，同时进行更新。PPO2 使用定长轨迹段，轨迹段指一组状态和动作的历史记录形式化的表示。在每次迭代中，N 个并行智能体收集 T 个步长的数据。在 NT 步上构造损失，并使用小批量梯度下降或者 Adam 优化器对其进行优化，使用在策略网络和价值网络之间共享 3 个隐含层的 MLP 网络，第 1 层有 128 个神经元，第 2 层有 64 个神经元，第 3 层有 64 个神经元，并在每个隐含层后增加一个线性单元 ReLU 激活函数。该框架的训练过程如算法 1 所示。

算法 1 基于 PPO2 的入侵检测系统构建

输入：(ε,θ,F)

1. S=SelectFeature(F)；//特征选择算法

2. 初始化入侵检测环境：A；O；Reset(S)；Step()；

3. for k=1,2,…,do

for actor=1,2,…,N do

用旧策略$\pi_{\theta_{old}}$在入侵检测环境中运行 T 个时间步长；

用优势函数估计$\hat{A}_1,\cdots,\hat{A}_t$；

计算策略更新$\theta_{k+1}=\arg\max_{\theta} L_{\theta_k}^{CLIP}(\theta)$；

Step()；

Reset(S)；

end for

经过 K 个时间步长用小批量 SGD(Adam)更新，其中：

$$L_{\theta_k}^{CLIP}(\theta)=\underset{\tau\sim\pi_k}{E}\left[\sum_{t=0}^{T}\min\left(r_t(\theta)\hat{A}_t^{\pi_k},\text{clip}\left(r_t(\theta),1-\varepsilon,1+\varepsilon\right)\hat{A}_t^{\pi_k}\right)\right];$$

end for

算法 1 中，F 表示原始数据集；S 是特征选择以及数据处理后的数据集；A 为动作空间，1 表示发出警报，0 表示不做任何事情；O 为观察空间，维度为 S 的特征个数，0 是正常流量，其他代表攻击流量；Reset(S)为重置函数，随机地从数据集 S 中选择训练开始的位置；Step( )为比较动作空间和观察空间的标签函数，错失了攻击则会终止，否则继续选择下一个状态进行动作。PPO2 算法以 Actor-Critic Style 呈现[21]。

4 实验与结果

4.1 实验环境

本文提出的 DRL-IDS 中使用的强化学习的框架基于 Stable baseline[1)] (2.10.0)[22]，是一套基于 OpenAI 基线的强化学习算法的改进实现。依赖 Tensorflow (1.14.0) 这一端到端的机器学习开源平台进行神经网络的搭建，使用 OpenAI Gym[2)] (0.17.2) 库来辅助完成强化学习中自定义的环境，并使用 4 个指标来评估模型的性能，即准确率、精确率、召回率和 F1 评分。对于硬件，本次实验在 Ubuntu 18.04.3 LTS 系统的机器上完成，实验所采用的硬件具体如下：CPU Model：Intel Xeon E5-2618L v3，GPU：NVIDIA GeForce RTX 2080TI，RAM：64 GB。

4.2 数据集

本文采用了美国能源部橡树岭国家实验室公开发布的天然气管道运输网络的真实数据集[23]，对所提的 DRL-IDS 入侵检测模型进行性能评估等实验。在该数据资源中，分别收集了正常网络流量数据和 7 种不同类型的攻击数据。该数据集有 26 个特征和 1 个标签。在实验中，我们将数据集划分为 3 个部分，60％用于训练，20％用于测试，20％用于验证。对于本文所有实验，均在相同数据集上进行测试。每种类别的数量情况如表 1 所列。同时，为了尽可能避免实验结果产生误差，本文所有实验均重复进行 10 次，并对所有计算结果取平均值，从而产生最终的实验结果。

表 1 数据集描述

Table 1 Dataset description

攻击种类	数量	攻击类型描述
Nomal	61156	正常网络流量
NMRI	2763	简单的恶意响应注入
CMRI	15466	复杂的恶意响应注入
MSCI	782	恶意状态命令注入
MPCI	7637	恶意参数命令注入
MFCI	573	恶意函数命令注入
DoS	1837	拒绝服务攻击
Reconnaissance	6805	侦察攻击

4.3 数据预处理

该天然气管道运输网络的数据集总共有 26 个特征，我们使用 3.2 节提到的特征选择方案去除无用的特征，在不降低性能的前提下减小运算的复杂度。第一步，删除 0 个缺失率大于 0.6 的特征。第二步，删除 8 个只有单一值的特征。第三步，删去 4 对强相关联的特征中的任意一个。第四步，用 LightGBM 进行特征的重要性排序，选取前 12 个特征进行实验。第五步，将变量归一化，生成特征向量，最后进行独热编码。

4.4 基于 PPO2 的 DRL-IDS 性能评估

实验使用 Stable baseline 的 PPO2 接口实现模型训练。

1) http://stable-baselines.readthedocs.io/

2) http://gym.openai.com/

训练过程中的主要参数如表 2 所列。

表 2 PPO2 超参数表

Table 2 PPO2 hyperparameters table

超参数名称	含义	取值
n_env	并行运行的环境副本数	15 int
n_steps	每个环境每次更新运行的步骤数(例如,批大小为 n_steps×n_env)	512 int
ent_coef	损失熵系数	0.00001 float
learning_rate	学习速率(可以是一个函数)	linear schedule from 0.0021 to 0.0
max_grad_norm	梯度剪切的最大值	0.8 float
minibatches	每次更新的培训小批数量	16 int

DRL-IDS 入侵检测智能体分别在测试集、训练集和验证集上进行测试,结果如表 3 所列,各项指标均在 97%以上。

表 3 DRL-IDS 入侵检测模型性能表现

Table 3 Performance of DRL-IDS

数据子集	准确率	召回率	精密度	F1 评分
Test	0.9909	0.9717	0.9857	0.9785
Train	0.9909	0.9762	0.9840	0.9799
Validation	0.9891	0.9717	0.9830	0.9773

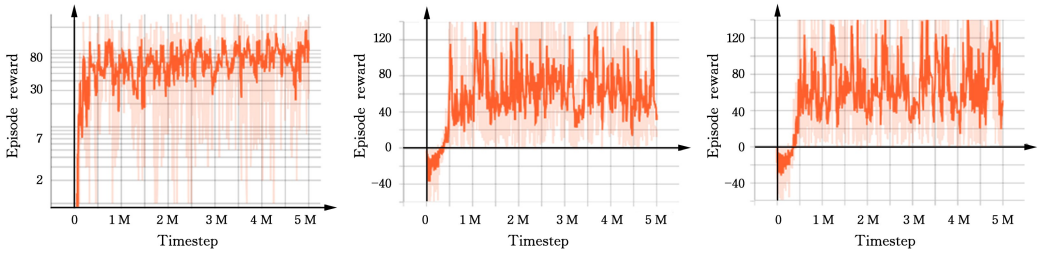
表 4 列出了 DRL-IDS 的各类攻击的检测结果,实验结果表明,每种攻击检测的精密度、F1 评分、召回率都较为理想。所有实验使用宏平均来综合评价模型的综合表现,得到 DRL-IDS 的准确率为 99.09%。

表 4 DRL-IDS 入侵检测模型每种类型的表现

Table 4 Classification report of DRL-IDS

攻击种类	F1 评分	召回率	精密度	准确率
Nomal	0.9930	0.9924	0.9936	0.9909
NMRI	0.9512	0.9457	0.9568	0.9909
CMRI	0.9943	0.9987	0.9899	0.9909
MSCI	0.9647	0.9588	0.9706	0.9909
MPCI	0.9767	0.9793	0.9741	0.9909
MFCI	0.9737	0.9487	1.0000	0.9909
DoS	0.9855	0.9755	0.9957	0.9909
Reconnaissance	1.0000	1.0000	1.0000	0.9909

此框架基于 PPO2 的入侵检测系统 DRL-IDS 在训练中用 TensorBoard(TensorFlow 可视化工具包)跟踪不同的变量,图 3 中比较了 PPO2,DQN,DDQN 的“Episode 奖励”,图 4—图 6 分别展现了 PPO2 的“折现回报”、线性学习率以及“Loss 函数”。从“Episode 奖励”可以看出,相比其他两种深度强化学习算法,基于 PPO2 算法的入侵检测在环境中得到的奖励在开始训练时就得到稳步提升。从图 4—图 6 可以看出,基于 PPO2 算法的入侵检测框架收敛快且稳定。



(a)PPO2 Episode 奖励 (b)DDQN Episode 奖励 (c)DQN Episode 奖励

图 3 Episode 奖励

Fig. 3 Episode reward

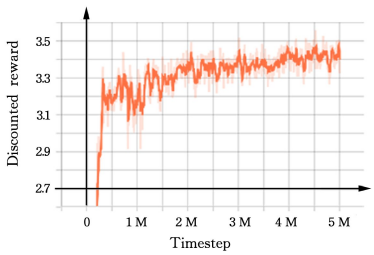


图 4 折现回报

Fig. 4 Discounted reward

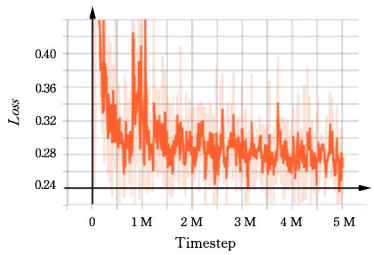


图 6 Loss 函数

Fig. 6 Loss functio

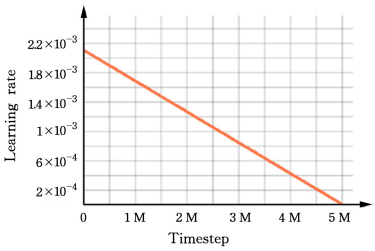


图 5 线性学习率

Fig. 5 Learning rate

4.5 与主流入侵检测系统的性能对比

将此基于 PPO2 算法的入侵检测系统在相同的神经网络结构下与强化学习领域的另一种算法 DDQN[24]进行对比,同时还与基于深度学习的主流学习算法 LSTM[25],CNN[26],RNN[27]的入侵检测系统进行对比。本文对比算法的参数设置分别参考文献[25-27]。实验结果表明(如表 5 和图 7 所示),DRL-IDS 在准确率、精密度、召回率、F1 评分等方面均优于其他基准系统(所有实验均使用统一数据集)。同时,将基于深度学习的检测方法和基于深度强化学习的检测方法进行

对比可以看出，基于深度强化学习的检测方法整体表现更好。

表 5 入侵检测模型对比

Table 5 Performance comparison with other detection models

模型	准确率	召回率	精密度	*F*1 评分
CNN	0.9884	0.9769	0.9695	0.9730
RNN	0.9854	0.9817	0.9602	0.9705
LSTM	0.9884	0.9829	0.9604	0.9712
DDQN	0.9905	0.9841	0.9708	0.9773
DQN	0.9904	0.9848	0.9691	0.9768
PPO2	**0.9910**	**0.9856**	**0.9717**	**0.9785**

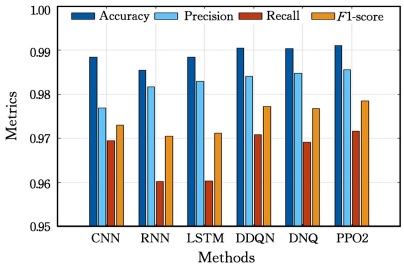


图 7 与其他模型的对比

Fig. 7 Comparison with other model based IDSs

实验进一步比较了本文提出的基于 PPO2 算法的入侵检测系统 DRL-IDS 和基于 DDQN 的入侵检测系统、基于 DQN 的入侵检测系统在相同数据量的情况下所需要的训练时间，实验结果表明，DRL-IDS 需要 7467.60 s，基于 DDQN 的入侵检测系统需要 19290.87 s，基于 DQN 的入侵检测系统需要 18945.10 s。结果表明，基于 PPO2 的工业物联网场景下的入侵检测系统训练的时间成本低，更适用于真实的入侵检测场景。

结束语 本文针对复杂多样的工业物联网环境，提出了一种新的基于深度强化学习算法 PPO2 的入侵检测系统 DRL-IDS。该系统运用了基于 LightGBM 的特征选择算法，有效提取了工业物联网数据中最有效的特征集合；结合深度学习算法将多层感知器网络的隐藏层作为 PPO2 算法中的价值网络和策略网络的共享网络结构；基于 PPO2 算法构建了工业物联网入侵检测模型。在美国能源部橡树岭国家实验室公开发布的工业物联网真实数据集上开展的大量实验结果表明，本文提出的入侵检测系统 DRL-IDS 在检测对工业物联网的多种类型网络攻击时性能表现良好，相比现有的基于深度学习或深度强化学习的入侵检测系统，在准确率、精密度、召回率、F1 评分方面均更胜一筹，且显著降低了入侵检测模型的训练时间。在今后的工作中，我们将探索基于分布式架构的工业物联网入侵检测系统。

参 考 文 献

[1] ZHOU W G. Analysis of Hidden Dangers of Industrial Internet of Things and Exploration of Protection Strategies[J]. Electronics World, 2019(21): 13-18.

[2] LING M H, YAU K L A, QADIR J, et al. Application of reinforcement learning for security enhancement in cognitive radio networks[J]. Applied Soft Computing, 2015, 37: 809-829.

[3] LU X, XIAO L, XU T, et al. Reinforcement Learning Based PHY Authentication for VANETs[J]. IEEE Transactions on Vehicular Technology, 2020, 69(3): 3068-3079.

[4] LOPEZ-MARTIN M, CARRO B, SANCHEZ-ESGUEVILLAS A. Application of deep reinforcement learning to intrusion detection for supervised problems[J]. Expert Systems with Applications, 2020, 141: 112963.

[5] HSU Y F, MATSUOKA M. A Deep Reinforcement Learning Approach for Anomaly Network Intrusion Detection System [C]// 2020 IEEE 9th International Conference on Cloud Networking (CloudNet). 2020: 1-6.

[6] PENG A N, ZHOU W, JIA Y, et al. Overview of Research on Security of Internet of Things Operating System[J]. Journal on Communications, 2018, 39(3): 22-34.

[7] AL-HAWAWREH M, MOUSTAFA N, SITNIKOVA E. Identification of malicious activities in industrial internet of things based on deep learning models[J]. Journal of Information Security and Applications, 2018, 41: 1-11.

[8] ROY B, CHEUNG H. A Deep Learning Approach for Intrusion Detection in Internet of Things using Bi-Directional Long Short-Term Memory Recurrent Neural Network[C]// 28th International Telecommunication Networks and Applications Conference (ITNAC). 2018: 1-6.

[9] YANG H, CHENG L, CHUAH M C. Deep-Learning-Based Network Intrusion Detection for SCADA Systems[C]// 2019 IEEE Conference on Communications and Network Security (CNS). Washington, DC, USA: IEEE, 2019: 3-5.

[10] ISMAIL M, SHAABAN M, NAIDU M, et al. Deep Learning Detection of Electricity Theft Cyber-Attacks in Renewable Distributed Generation[C]// IEEE Transactions on Smart Grid, 2020: 3428-3431.

[11] LI B, WU Y, SONG J, et al. DeepFed: Federated Deep Learning for Intrusion Detection in Industrial Cyber-Physical Systems [J]. IEEE Transactions on Industrial Informatics, 2021, 17(8): 5615-5624.

[12] KURT M N, OGUNDIJO O, LI C, et al. Online Cyber-Attack Detection in Smart Grid: A Reinforcement Learning Approach [J]. IEEE Transactions on Smart Grid, 2019, 10(5): 5174-5185.

[13] SETHI K, EDUPUGANTI S, KUMAR R, et al. A context-aware robust intrusion detection system: a reinforcement learning-based approach[J]. International Journal of Information Security, 2020, 19: 657-678.

[14] OTOUM S, KANTARCI B, MOUFTAH H. Empowering Reinforcement Learning on Big Sensed Data for Intrusion Detection [C]// 2019 IEEE International Conference on Communications (ICC 2019). 2019: 1-7.

[15] CAMINERO G, LOPEZ-MARTIN M, CARRO B. Adversarial environment reinforcement learning algorithm for intrusion detection[J]. Computer Networks, 2019, 159: 96-109.

[16] SONG J, LI B, WU Y, et al. ReAL: A New ResNet-ALSTM Based Intrusion Detection System for the Internet of Energy [C]// 2020 IEEE 45th Conference on Local Computer Networks (LCN). 2020: 491-496.

[17] NAHLER G. Pearson correlation coefficient[J]. Dictionary of Pharmaceutical Medicine, 2009, 1025: 132-132.

[18] WANG H, CHEN H Y, LIU S F. Intrusion Detection System Based on Improved Naive Bayes Algorithm[J]. Computer Science, 2014, 41(4): 111-115, 119.

[19] WU Y, MANSIMOV E, LIAO S. Scalable Trust-Region Method for Deep Reinforcement Learning Using Kronecker-Factored Approximation[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. California: Curran Associates Inc, 2017: 5285-5294.

[20] MNIH V, BADIA A P, MIRZA M, et al. Asynchronous Methods for Deep Reinforcement Learning[C]// International Conference on Machine Learning (PMLR 2016). 2016: 1928-1937.

[21] SCHULMAN J, WOLSKI F, DHARIWAL P. Proximal Policy Optimization Algorithms[EB/OL]. http://arxiv.org/abs/1707.06347.

[22] HILL A. Stable-baselines[EB/OL]. (2021). https://stable-baselines.readthedocs.io/en/master/.

[23] MORRIS T, GAO W. Industrial Control System Traffic Data Sets for Intrusion Detection Research[C]// International Conference on Critical Infrastructure Protection. Berlin, Heidelberg: Springer, 2014: 65-78.

[24] VAN HASSELT H, GUEZ A, SILVER D. Deep Reinforcement Learning with Double Q-learning[EB/OL]. http://arxiv.org/abs/1509.06461v2.

[25] MIRZA A, COSAN S. Computer network intrusion detection using sequential LSTM Neural Networks autoencoders[C]// 2018 26th Signal Processing and Communications Applications Conference (SIU). Izmir, Turkey: IEEE, 2018: 2-5.

[26] MELIBOYEV A, ALIKHANOV J, KIM W. 1D CNN Based Network Intrusion Detection with Normalization on Imbalanced Data[EB/OL]. http://arxiv.org/abs/2003.00476v2.

[27] YIN C L, ZHU Y F, FEI J L, et al. A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks[J]. IEEE Access, 2017, 5: 21954-21961.

LI Bei-bei, born in 1992, Ph.D, associate professor, is a member of China Computer Federation. His main research interests include cyber-physical system security, industrial control system security, big data & privacy preservation, and applied cryptography.

HE Jun-jiang, born in 1993, Ph.D, assistant professor. His main research interests include cyber security, artificial immune system, data mining, machine learning, and evolutionary computing.