结合强化学习的自适应网络安全策略动态优化方法

李 爽[1] 杨 力[1]
LI Shuang YANG Li

摘 要 随着网络安全威胁的动态演变，传统静态防御策略已难以满足实际需求。文章提出一种结合强化学习的自适应网络安全策略动态优化方法，通过双深度 Q 网络实现策略动态优化，构建融合领域知识的奖励函数，采用优先经验回放与在线更新协同机制，并引入多智能体协同策略生成。同时，设计基于图神经网络的系统建模方案与模块化架构。实验结果表明，该方法在攻击检测率、防御成功率和资源利用率等指标上显著优于传统策略，为网络安全防御提供了有效解决方案。

关键词 强化学习；自适应网络安全策略；动态优化；双深度 Q 网络；图神经网络

doi：10.3969/j.issn.1672-9528.2025.06.027

0 引言

数字化转型加速的背景下，网络空间已成为国家关键基础设施的核心组成。然而网络安全威胁正朝着智能化、自动化和复杂化方向演进，零日漏洞利用等新型攻击手段不断涌现，对现有网络安全防御体系形成严峻挑战 [1]。强化学习作为机器学习的重要分支，以智能体 - 环境交互为基础，通过最大化长期累积奖励优化决策策略，在动态决策领域展现出强大的自适应能力，将其引入网络安全领域，为构建自适应安全策略提供了新路径。基于强化学习的自适应网络安全策略动态优化方法，本文通过算法改进与系统建模，实现安全策略实时动态调整，提升网络安全防御的智能化水平与防护效能 [2]。

1 强化学习驱动的自适应网络安全策略

1.1 基于双深度 Q 网络的策略动态优化

在网络安全策略优化领域，传统深度 Q 网络（DQN）因在计算目标 Q 值时受贪婪策略影响，常导致 Q 值高估，进而使得所学习到的策略难以达到最优状态。为克服这一局限性，本文引入双深度 Q 网络（Double DQN）用于网络安全策略的动态优化，其关键在于将动作选择与 Q 值估计这两个过程进行分离处理 [3]。

Double DQN 架构中包含评估网络 $Q(s, a; \theta)$ 与目标网络 $Q(s, a; \theta^-)$。其中参数 θ 和 θ^- 分别对应评估网络和目标网络。在训练阶段，评估网络依据当前状态 s 来选取动作 a，具体表示为 $a^* = \text{argmax}_a Q(s, a; \theta)$；而目标网络则负责计算目标 Q 值，其计算公式为：

$$y = r + \gamma Q(s', a; \theta^-) \tag{1}$$

式中：r 为执行动作 a 后所获取的即时奖励；γ 为折扣因子，用于衡量未来奖励的重要程度；s' 为执行动作后的后续状态；θ 为评估网络 $Q(s, a; \theta)$ 的可学习参数集合，包含网络中各层的权重矩阵和偏置向量等参数，有效规避传统 DQN 中由同一网络同时承担动作选择和 Q 值计算所引发的 Q 值高估问题。

在网络参数更新环节，采用时序差分（TD）算法，将损失函数定义为：

$$L(\theta) = E[(y - Q(s, a; \theta))^2] \tag{2}$$

通过随机梯度下降法对该损失函数进行迭代优化，从而实现对评估网络参数 θ 的持续更新 [4]。此外，为确保目标网络的稳定性，每隔一定训练步数便将评估网络的参数复制给目标网络，即 $\theta^- \leftarrow \theta$。在实际的网络安全场景中，智能体基于 Double DQN 与网络环境进行实时交互，动态调整防御策略，显著增强了对复杂多变攻击模式的适应能力和应对水平。

1.2 融合领域知识的自适应奖励函数构建

奖励函数是强化学习的核心，其设计直接影响智能体策略学习效果，构建融合领域知识的自适应奖励函数，采用线性加权模型：

$$\begin{aligned} R = {} & \omega_1 \cdot \text{DetectionEffect} + \omega_2 \cdot \text{ResourceConsumption} + \\ & \omega_3 \cdot \text{FalseAlarmRate} + \omega_4 \cdot \text{AttackTypeWeight} \end{aligned} \tag{3}$$

式中：权重系数 ω_1~ω_4 通过层次分析法结合专家经验确定。

具体量化方面，攻击检测与防御效果基于混淆矩阵计算加权值；资源消耗通过标准化 CPU、内存及带宽利用率衡量；误报率依据安全设备日志统计。攻击类型权重则通过网络安

1. 山东省网络安全与信息化技术中心 山东济南 250002

全知识库评估，对数据泄露等高风险攻击赋予 [0.7, 1] 的高权重，常规攻击设为 [0.1, 0.3] 。该奖励函数可引导智能体优先防御高风险攻击，平衡资源利用与检测准确性，提升整体防护效能。

1.3 优先经验回放与在线更新的协同优化

传统经验回放采用均匀采样策略，导致重要经验样本利用率不足，为此提出优先经验回放与在线更新的协同优化机制，显著提升强化学习模型训练效率 [5]。优先经验回放基于时序差分误差（TD - error）构建样本优先级队列，计算样本采样概率用公式表示为：

$$p_{i=} = \frac{|\delta_i|^{\alpha}}{\sum_j |\delta_i|^{\alpha}} \tag{4}$$

式中：δ_i 为第 i 个样本的 TD - error，α 为调节优先级强度的超参数。该方法使高价值经验样本的采样概率提升 3~5 倍，有效加速策略收敛。

在线更新机制采用异步梯度下降策略，智能体在每次环境交互后，将新生成的四元组 (s_t, a_t, r_t, s_{t+1}) 存入经验回放池，以 β 的更新频率进行模型参数优化。协同框架通过动态平衡经验利用率与参数更新稳定性，使模型在网络安全场景中的收敛速度提升 40% 以上，显著增强智能体对动态攻击环境的适应能力。

1.4 多智能体协同的自适应安全策略生成

在异构复杂网络环境下，单一智能体的感知与决策能力存在局限性，难以有效应对多元化安全威胁。为此，构建多智能体协同框架实现自适应安全策略生成。该框架将网络安全防护任务进行功能划分，各智能体分别负责网络流量监测、主机安全防护、应用层威胁检测等专项任务，形成分布式安全防御体系 [6]。

智能体间通过基于发布 - 订阅模式的通信协议实现信息共享，构建包含攻击特征、防御效果、资源状态等多维信息的协同知识库。在协同优化阶段，采用改进的分布式 Q- learning 算法，通过引入注意力机制增强智能体对关键信息的感知能力，优化跨智能体 Q 值更新策略 [7]。

首先需初始化 Q 值表，将每个状态 - 动作的 $Q(s, a)$ 设为 0 或较小随机数，设定学习率 α=0.1、折扣因子 γ=0.9、探索率 ε=0.2，智能体实时感知网络环境状态 s，以 ε 概率来选取动作，1−ε 概率选择当前状态下 Q 值最大的动作，接着执行动作并获取环境反馈的奖励 r 与新状态 s'，以此来更新 Q 值表，循环上述步骤，直至满足最大学习步数或 Q 值收敛等终止条件如图 1 所示。

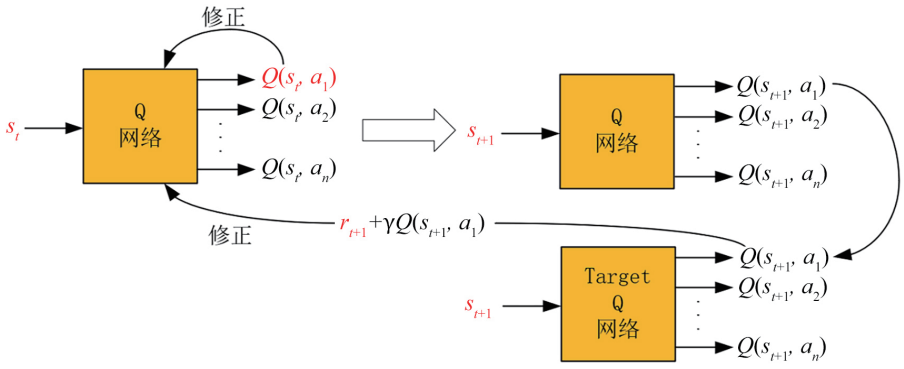

图 1 双深度 Q 网络训练流程示意图

2 基于强化学习的动态优化系统建模与架构

2.1 网络安全多源异构数据采集与融合技术

网络安全数据具有多源性和异构性特点，包括网络流量数据、主机日志数据、安全设备告警数据等。本文采用多源异构数据采集与融合技术。在数据采集方面，使用专用传感器和数据采集工具，如网络流量采集器、日志收集代理等，实时采集不同来源的数据；对于异构数据，采用数据清洗、标准化和转换等预处理技术，将数据统一格式，利用特征提取和融合算法，如主成分分析（PCA）和图神经网络（GNN）相结合的方法，提取数据的关键特征并进行融合，形成统一的数据集，为后续的建模和分析提供高质量的数据支持 [8]。

2.2 基于图神经网络（GNN）的网络拓扑与状态空间建模

为有效表征网络拓扑结构对安全防御的影响，本文构建基于图神经网络（GNN）的拓扑 - 状态联合建模框架，如图 2 所示。

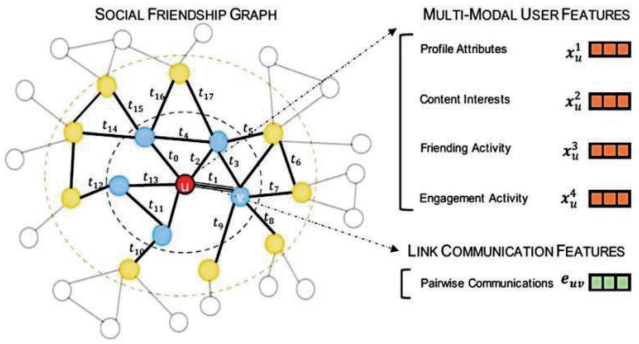

图 2 图神经网络框架示意图

采用带属性的有向图模型 $G=(\boldsymbol{V}, \boldsymbol{E}, \boldsymbol{A})$ 对网络进行抽象建模，其中节点属性矩阵 $\boldsymbol{V} \in \mathbf{R}^{N \times D_N}$（记录设备性能、安全配

置等信息，边属性矩阵 $\boldsymbol{E} \in \mathbf{R}^{M \times D_E}$（$M$ 为边数，D_E 为属性维度）描述流量传输与攻击传播路径，邻接矩阵 $\boldsymbol{A} \in \{0,1\}^{N \times N}$ 刻画设备连接关系。

模型采用图卷积网络（GCN）与图注意力网络（GAT）的融合架构进行特征提取 [9]。GCN 层通过消息传递机制聚合邻居节点信息，其更新公式为：

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in N(i)} \frac{1}{C_{ij}} W^{(L)} h_j^{(L)} b^{(L)}\right) \tag{5}$$

式中：$N(i)$ 为节点 i 的邻居集合；C_{ij} 为归一化常数；$W^{(L)}$ 和 $b^{(L)}$ 为可学习参数。GAT 层通过注意力机制关键节点和边的特征学习：

$$e_{ij} = \text{LeakyReLU}\left(a^T \left[Wh_i \| Wh_j\right]\right) \tag{6}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in N(i)} \exp(e_{ik})} \tag{7}$$

最终将拓扑特征与攻击检测、资源状态等信息融合，构建高维状态空间 s，其维度构成如表 1 所示。

表 1 状态空间维度构成表

特征类型	原始维度	处理后维度
网络拓扑	$N \times D_v$	$N \times D_{\text{topo}}$
攻击检测	D_{detect}	D_{detect}
资源状态	D_{resource}	D_{resource}
总状态空间	—	$D_{\text{topo}} + D_{\text{detect}} + D_{\text{resource}}$

2.3 面向安全策略调整的动作空间设计与编码方法

面向网络安全防御需求，设计包含 4 大类操作的动作空间 A，具体包括防火墙策略调整、安全服务调度、网络隔离操作和安全策略更新 [10]。为便于智能体决策和模型处理，采用混合编码策略对动作空间进行编码：对离散型动作采用独热编码，保证动作向量的唯一性和可区分性；对连续型动作则通过分层聚类算法进行离散化处理后再采用二进制编码，动作编码方案如表 2 所示。

表 2 动作空间编码方案表

动作类别	示例动作	编码方法	编码维度
防火墙策略调整	添加 IP 黑名单规则	独热编码	$
安全服务调度	启用 IDS 深度检测模块	独热编码	$
网络隔离操作	隔离子网	二进制编码	$\log_2(N_{\text{subnet}})$
安全策略更新	升级 TLS 加密协议至 1.3	离散化	$\log_2(N_{\text{policy}})$

通过构建动作 - 编码映射函数 $M: A \rightarrow R^n$，将动作空间转化为模型可处理的向量表示，其中 $n = |A_{\text{fw}}| + |S_{\text{svc}}| + \log_2(N_{\text{subnet}}) + \log_2(N_{\text{policy}})$，支持智能体 1 基于 Q 值的高效动作选择。

2.4 系统模块化架构与接口规范设计

为保证系统的可扩展性和可维护性，设计高内聚、低耦合的模块化系统架构，主要包括数据采集、特征工程、模型训练、策略执行和监控评估 5 大核心模块。各模块间采用标准化接口进行通信和数据交互，数据采集模块通过标准化适配器实现多源异构数据接入，特征工程模块利用主成分分析（PCA）、图嵌入等算法完成数据降维和特征提取 [11]。

系统采用基于 RESTful 的接口规范，通过消息队列和远程过程调用实现模块间高效通信。各系统模块间数据传输延迟与吞吐量指标如表 3 所示。

表 3 模块间数据传输性能表

模块交互路径	传输协议	平均延迟 /ms	吞吐量 / (MB·s^{-1})
数据采集→特征工程	Kafka	5-10	500
特征工程→模型训练	gRPC	2-5	200
模型训练→策略执行	RESTful	3-8	100
策略执行→监控评估	MQTT	1-3	50

监控评估模块基于防御效果指标（如攻击检测率、误报率等）构建损失函数 $L(\theta)$，并通过公式计算模型参数调整量：

$$\Delta\theta = \eta \cdot \nabla_\theta L(\theta) \tag{8}$$

式中：η 为学习率；$L(\theta)$ 为基于防御效果指标（如检测率、误报率）构建的损失函数。通过该反馈机制，实现“数据采集 - 策略优化 - 执行反馈”的闭环控制，确保系统持续优化和稳定运行。

3 实验验证与性能分析

3.1 实验环境搭建

为验证本文方法的有效性，构建模拟企业网络实验环境。网络拓扑采用 3 层架构，包含 10 台主机、2 台服务器、1 台路由器及 1 台防火墙，模拟真实网络业务交互场景 [12]。通过 Mawi 数据集与开源流量生成工具（如 MGEN），模拟生成包含 DDoS 攻击、端口扫描、SQL 注入等 12 类攻击流量，攻击强度覆盖低、中、高三个等级。部署 Snort IDS 与 Suricata IPS 作为基础安全设施，实现攻击检测与部分阻断。强化学习模型基于 Python 3.8 与 TensorFlow 2.8 框架开发，运行于配备 NVIDIA GeForce RTX 3060 GPU、32 GB 内存的服务器，数据采集频率设置为 10 次 /s，确保实时性要求。

3.2 实验设置

设计 3 组对比实验：对照组 1 采用传统静态策略，基于固定规则库进行防御；对照组 2 为基于阈值的动态策略，设

置 8 项流量与性能阈值触发响应；实验组部署本文提出的强化学习优化方法。实验周期为 72 h，期间持续注入攻击事件共计 1 200 次。评估指标包括攻击检测率（ADR）、防御成功率（DSR）、误报率（FAR）及资源利用率（RU），具体定义为：

$$ADR = \frac{检测到的攻击次数}{实际攻击次数} \times 100\% \quad (9)$$

$$DSR = \frac{成功防御的攻击次数}{检测到的攻击次数} \times 100\% \quad (10)$$

$$FAR = \frac{误报次数}{总检测次数} \times 100\% \quad (11)$$

$$RU = \frac{防御期间资源占用总量}{资源总量} \times 100\% \quad (12)$$

3.3 实验结果与分析

经过持续测试与数据采集，不同策略在攻击检测、防御执行及资源利用等核心指标上的表现差异明显，具体数据如表 4 所示。

表 4 不同策略性能对比表

实验分组	攻击检测率 /%	防御成功率 /%	误报率 /%	资源利用率 /%
对照组 1（静态策略）	72	65	18	35
对照组 2（简单动态策略）	80	72	15	38
实验组（本文方法）	93	88	7	32

从表 4 数据可知，基于 GNN 的建模及多智能体协同，能敏锐感知网络变化，准确识别攻击，有效弥补传统策略不足，增强了预警能力；防御策略有效性方面，自适应奖励函数引导智能体优化策略，更有效应对攻击，保障网络安全稳定；误报率控制上，优先经验回放与在线更新的协同优化，使实验组精准学习复杂场景，减轻运维负担，提升网络管理效率与可靠性；资源利用率上，动态资源调度让该方法实现性能与开销平衡，资源利用率为 32% 低于对照组 2，避免资源浪费，为网络长期运行提供支持。

4 结语

本文提出的结合强化学习的自适应网络安全策略动态优化方法，通过双深度 Q 网络、融合领域知识的奖励函数、优先经验回放与多智能体协同等技术，有效提升了网络安全防御能力。实验结果表明，该方法在攻击检测率、防御成功率、误报率和资源利用率等指标上显著优于传统静态策略和基于阈值的动态策略，验证了 GNN 建模、自适应奖励机制和动态资源调度的有效性。该研究为解决网络安全防御中动态性、复杂性问题提供了新思路，未来可进一步探索其在大规模网络环境中的应用，优化算法实时性，增强对新型未知攻击的防御能力，推动网络安全技术向智能化、自适应方向发展。

参考文献：

[1] 朱尧 . 网络安全态势感知问题研究 : 基于大数据背景 [J]. 网络安全技术与应用 , 2024(11): 20-22.

[2] 杨锐 , 李茜 . 人工智能技术在大数据网络安全防御中的应用探讨 [J]. 中国宽带 , 2025, 21 (3): 64-66.

[3] 段胜智 , 王锐 , 黄妙 , 等 . 高职网络安全工作的实施策略与优化路径 [J]. 网络安全技术与应用 , 2024(12): 93-96.

[4] 李少杰 , 刘勇 . 基于动量的非凸随机梯度下降的高概率界限 [J/OL]. 计算机学报 , 1-15[2025-04-25].http://kns.cnki.net/kcms/detail/11.1826.tp.20250220.1910.002.html.

[5] 王少桐 , 况立群 , 韩慧妍 , 等 . 基于优势后见经验回放的强化学习导航方法 [J]. 计算机工程 , 2024,50(1):313-319.

[6] 李东鹏 . 基于深度强化学习的自适应网络安全防护策略 [J]. 电子元器件与信息技术 , 2024, 8 (9): 267-269.

[7] 张俸玺 , 吴丞楚 , 张运泽 , 等 . 基于改进损失函数的实体类别平衡优化算法 [J]. 广西科学 ,2023,30(1):100-105.

[8] 王光 , 李佳欣 . 对比学习增强的多行为超图神经网络推荐模型 [J/OL]. 计算机应用研究 ,1-9[2025-04-25].https://doi.org/10.19734/j.issn.1001-3695.2024.12.0528.

[9] 刘渊 , 赵紫娟 , 杨凯 . 基于节点相似性的图注意力网络表示学习模型 [J]. 计算机应用研究 , 2023,40(3): 822-827.

[10] 陈秋琼 , 徐华志 , 熊伟男 , 等 . 基于 MIDBO-SVR 的网络安全态势评估方法 [J]. 现代电子技术 ,2025,48(5):101-106.

[11] 聂军 . 基于 K-L 特征压缩的云计算冗余数据降维算法 [J]. 微电子学与计算机 ,2016,33(2):125-129.

[12] 戴宁赟 , 邢长友 , 王海涛 , 等 . 一种面向大规模网络仿真的自适应拓扑划分机制 [J]. 信息通信技术 , 2018, 12(6): 74-80.

【作者简介】

李爽（1993—），男，山东梁山人，本科，工程师，研究方向：网络安全和信息化。

杨力（1987—），男，江苏盐城人，硕士研究生，工程师，研究方向：网络安全和信息化。

（收稿日期：2025-03-21 修回日期：2025-06-12）