doi:10.11835/j.issn.1000-582X.2025.06.008
引用格式:杨云,梁花,魏兴慎,等.结合博弈论与强化学习的态势感知与路径预测[J].重庆大学学报,2025,48(6):84-97.

结合博弈论与强化学习的态势感知与路径预测

杨 云[1],梁 花[2],魏兴慎[3,4],李 洋[2],刘 俊[5]
(1.国网重庆市电力公司 重庆 400014;2.国网重庆电力公司电力科学研究院 重庆 401123;3.国网电力科学研究院有限公司 南京 211106;4.南瑞集团有限公司 南京南瑞信息通信科技有限公司 南京 211106;5.重庆邮电大学 软件工程学院 重庆 400000)

摘要:网络安全态势感知技术对评估网络安全状况及预测攻击行为路径,辅助管理员做出有效防御有重要意义。传统的网络态势评估方法大多偏重在理论层面进行静态分析,难以实际运用,传感器收集到的数据庞大繁杂,易造成存储空间负载过大。针对上述问题,结合博弈论算法与强化学习算法,提出一种结合博弈论与强化学习的网络攻防动态感知模型以分析网络态势安全及预测攻击路径。首先,设计带有优先级关系矩阵的层次分析法计算系统损失及安全态势;其次,引入Boltzmann概率分布法计算混合策略纳什均衡;最后,改进Q-Learning与博弈论算法对网络状态转移进行动态分析,达到准确预测攻击路径、选择最优防御策略的目的。通过网络仿真实验,验证模型的有效性和可行性。

关键词:强化学习;Q-learning;博弈论;态势感知;层次分析法;纳什均衡
中图分类号:TP393 **文献标志码:**A **文章编号:**1000-582X(2025)06-084-14

Situational awareness and path prediction combining game theory and reinforcement learning

YANG Yun[1], *LIANG Hua*[2], *WEI Xingshen*[3,4], *LI Yang*[2], *LIU Jun*[5]
(1. State Grid Chongqing Electric Power Company, Chongqing 400014, P. R. China; 2. Electric Power Research Institute of State Grid Chongqing Electric Power Company, Chongqing 401123, P. R. China; 3. State Grid Electric Power Research Institute Co., Ltd., Nanjing 211106, P. R. China; 4. Nanjing NARI Information Communication Technology Co., Ltd., NARI Group Co., Ltd., Nanjing 211106, P. R. China; 5. School of Software Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400000, P. R. China)

Abstract: Cybersecurity situational awareness technology plays a critical role in assessing network security status, predicting potential attack paths, and assisting administrators in implementing effective defenses. Traditional methods for network situation assessment mostly rely on theoretical analysis, limiting their practicality in real-world networks. Additionally, the complexity of sensor-collected data often results in excessive storage demands.

收稿日期:2020-10-12
基金项目:重庆市电力公司科技项目(520626190067)。
Supported by Science and Technology Projects of State Grid Chongqing Electric Power Company (520626190067).
作者简介:杨云(1964—),男,高级工程师,主要从事电力通信技术、智能电网和信息安全技术方向研究,(E-mail) yy@cq.sgcc.com.cn。
通信作者:刘俊(1978—),男,硕士生导师,(E-mail) junliu@cqupt.edu.cn。

To address these challenges, this paper proposes a dynamic network attack-defense perception model that integrates reinforcement learning and game theory to enhance situational awareness and predict potential attack paths. The approach begins with the design of a hierarchical analytic process using a priority relation matrix to calculate system losses and assess security posture. Next, the Boltzmann probability distribution is employed to calculate the mixed-strategy Nash equilibrium, identifying optimal strategic responses. Finally, an improved Q-learning algorithm, in combination with game-theoretic principles, is used to dynamically model network state transitions, enabling accurate prediction of attack paths and supporting defenders in selecting optimal defense strategies. Simulation results validate the model's effectiveness and practicality in complex network environments.
Keywords: reinforcement learning; Q-learning; game model; situational awareness; analytic hierarchy process; Nash equilibrium

面对网络威胁的进化，安全研究人员对现有攻击威胁、网络脆弱性[1]等进行了深入研究，研究成果例如防火墙、入侵检测技术等，但大多技术有局限性，影响网络安全防御效率。面对网络的多样性、异构性，近年来网络安全研究重点转移到数据融合[2-3]。现阶段面对传统防御体系抵抗攻击能力不够的风险，网络态势安全感知技术能够全面感知网络安全威胁，擅长主动防御、洞悉及评判网络健康状态，通过全流量分析技术实现完整的网络攻击溯源取证，帮助安全人员更好地采取针对性响应处置措施[4-5]。

国内外针对网络态势感知研究开展了大量研究[6]，包括可视化技术的网络安全态势感知、基于层次化分析的网络安全态势感知、基于数据融合技术的网络安全态势感知等，但目前还没有成熟的模型、评估方法和统一评判标准。按照评估依据的理论技术，可以将网络安全评估方法分为：知识理论方法、人工智能方法和基于数学模型的方法。

真实网络攻防中，攻防参与者如果足够理性，博弈论可以为攻防参与者选取收益最大的策略，达到纳什均衡[7-8]。强化学习中各代理人可以自行对环境进行探索，得到满足纳什均衡的最稳定策略集组成的最大概率攻击路径，帮助安全员预测攻击方的攻击路径[9]。理性的攻防参与者会尝试推理出对方战略，攻防回合操作较复杂，对于运用博弈论与强化学习算法解决态势感知问题来说是极大挑战。近年来一些研究小组采用博弈论的方法去解决此问题，但现有博弈模型大多是静态的，难以反映攻击防御动态变化和实时评估现实环境中的实现。考虑上述问题，结合 Q-learning 算法实现对网络态势及预测攻击路径的动态评估。

1　相关研究

网络安全评估方法主要分为：知识理论方法、人工智能方法和基于数学模型的方法。知识推理模型可以大致分为基于证据理论和图模型[10]推理，基于证据理论推理包括：D-S[11]、ER 等证据理论，但 D-S 证据理论的基本概率分配函数较难确定，并要求证据之间的相互独立。图模型推理方法包括贝叶斯网络[12]、模糊认知图[13]等，图模型推理方法便于理解，大多情况下计算条件概率与权重矩阵比较困难。人工智能的评估方法包含人工神经网络[14-15]、支持向量机等方法。但神经网络建模没有统一网络结构标准，结果随机性大。基于数学模型[16]的方法统计各类网络安全态势要素数据。

基于层次化的算法（analytic hierarchy process，AHP）[17]较为常用，该算法为计算某时刻攻击对网络系统的威胁，将服务器分为服务层、主机层、网络层。以一种先局部后整体，从下向上的策略对安全威胁量化评估，但 AHP 在检验判断矩阵是否一致较困难。张勇等[18]提出了一种基于马尔可夫博弈模型的网络态势感知方法，该模型对网络系统的脆弱性及威胁传播性进行分析，为管理员提供最优防御方案。实验表明，该模型评估准确。王华等[19]为了满足对抗性网络环境中的网络状态评估和预警，提出一种可调整生成序列的自适应灰色 Verhulst 模型。实验证明，该模型能对中长期网络安全态势进行合理预测。而博弈进化论[20]综合考虑各方面的影响因素，准确性较高，但目前大多博弈进化论的研究停留在静态博弈。因此，研究提出一种结合博弈论与强化学习的攻防动态感知模型，评估网络状态。该模型可以定义态势指标，描述攻击者和防御者双

方的关系,为网络行为提供有效的表达方式。

2 模型指标度量

笔者提出结合博弈论与强化学习的网络攻防动态感知模型(Q-learning and game theory network situation awareness model,QGNSAM),模型分为2部分,第1部分如图1所示为层次化分析,第2部分为$\boldsymbol{Q}$矩阵学习状态转移。层次化分析根据传感器及扫描工具得到基本配置信息,将配置信息传入模型,存储于数据库,依次求出服务层、主机层及网络层的网络系统损失值,得到当前状态的网络态势,采集攻防双方策略,生成攻防博弈图,计算混合策略纳什均衡,得到当前最稳定策略以及最新$\boldsymbol{Q}$学习矩阵。

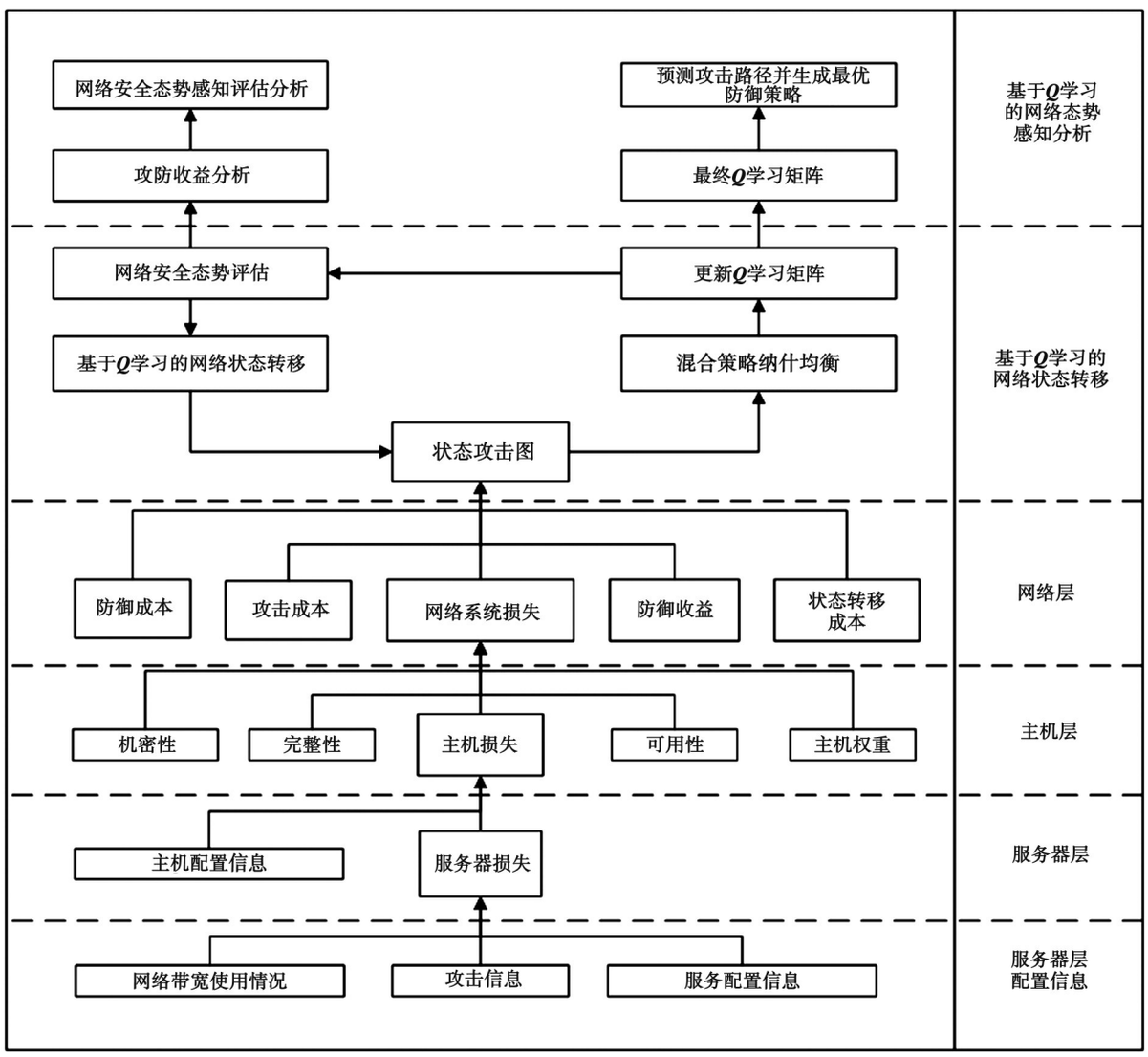

图1 QGNSAM模型框架

Fig.1 QGNSAM model framework

2.1 网络系统指标度量

攻防双方网络系统安全收益指标的量化在网络安全状态转移和网络态势分析计算中十分重要,根据林肯实验室总结出的各种攻防策略,研究提出了结合层次化分析的攻防策略指标量化方法。

2.1.1 系统损失

将攻击对目标网络系统造成的威胁程度作为系统损失值,引入层次分析法对系统损失(system cost, SyCost)进行分析。将网络系统分为服务层、主机层与网络系统层。攻击首先对服务造成威

胁,根据攻击的严重程度与网络带宽占用率计算服务威胁指数Rservice,根据服务威胁指数以及服务权重指数计算主机威胁指数RHost,最后,攻击根据主机威胁指数以及主机的重要性权重计算系统的威胁指数RNet。

1)服务层

$$\mathrm{Rservice}_k(t)=\sum_{j=1}^{p}\mathrm{Rservice}_{k_j}(t)=f(D_{k_j}(t),\mathrm{Net}(t))=\left(\sum_{j=1}^{p}D_{k_j}(t)\right)\times 100\mathrm{Net}(t),\tag{1}$$

式中:$\mathrm{Rservice}_k(t)$表示主机k中服务层损失值;$D_{k_j}(t)$为网络转移状态t,攻击j对主机k服务层的攻击威胁度。将攻击威胁度分为低、中、高,分别用0.1、1、10表示,以体现不同威胁的攻击对服务的伤害差距,说明100次低威胁攻击与1次高威胁攻击对系统的实际伤害相等,突出高威胁攻击的$\mathrm{Net}(t)=(\mathrm{Net}_1,\cdots,\mathrm{Net}_v)$风险。$\mathrm{Net}(t)$表示网络转移状态$t$中网络带宽占用率;$v$为该网络转移状态时间段内的分析时间窗口数。$100\times\mathrm{Net}(t)$的系数100是为了将网络带宽占用率转化为整数方便评估。服务威胁指数$\mathrm{Rservice}_{k_j}(t)$越大,攻击$j$对当前系统的服务层的威胁程度越高。

2)主机层

$$\mathrm{RHost}_k(t)=(\mathrm{Rservice}_k(t),AV_k(t))=\sum_{j=1}^{SV_k}\frac{(WR_{k_j}\times\mathrm{Rservice}_{k_j}(t))}{SV_k}+AV_k(t),\tag{2}$$

式中:WR_{k_j}表示主机k中服务j在目标网络中的权重,WR_{k_j}量化为1、2、3分别代表低权重、中等权重、高权重,其取值由网络安全员确定。计算主机k中每个服务j的威胁性,归一化得到主机k的总服务威胁性,最后与主机k可访问性求和,得到主机k威胁程度。

系统主机层损失指标选取对最终计算结果影响很大,研究采用模糊权重计算各个衡量主机层资源的重要性。$AV_k(t)$表示网络转移状态t阶段,主机k的可访问性。SV_k表示主机k中的服务总数。主机k由n种计算机资源组成,计算机资源的重要程度分为高、低,用1、0.5衡量,以指标i相关重要程度表示为$c(i)$。建立指标优先级关系矩阵$\boldsymbol{F}=(f_{ij})_{n\times n}$,$f_{ij}$定义如公式(3)所示。

$$f_{ij}=\begin{cases}0, & c(i)<c(j),\\ 0.5, & c(i)=c(j),\\ 1.0, & c(i)>c(j)。\end{cases}\tag{3}$$

利用公式(3)求得指标i的重要程度,并利用公式(4)(5)求得指标权重向量$\boldsymbol{w}^0=(w_1,\cdots,w_i,\cdots,w_n)^{\mathrm{T}}$,$w_i$表示第$i$个计算机资源指标权重。

$$h_i=\sum_{j=1}^{n}f_{ij}-0.5,\ (i=1,2,\cdots,n)。\tag{4}$$

$$w_i=\frac{h_i}{\sum_{j=1}^{n}h_i}。\tag{5}$$

利用公式(6)计算主机i的可访问性AV_i,$\mathrm{Ack}_i(t)$表示网络转移t状态中主机k中指标i的可访问性,分为可访问和不可访问,表示为1、0。

$$AV_k=\sum_{j=1}^{n}\mathrm{AcK}_i(t)\cdot w_i。\tag{6}$$

3)网络系统层

$$\mathrm{SyCost}(t)=\mathrm{RNet}(t)=\sum_{k=1}^{m}\mathrm{RHost}_k(t)(\mathrm{criticality}\times WH_k\times P_{AV_k}+\mathrm{InCost}_k(t)\times P_{in_k}+\mathrm{ConCost}_k(t)\times P_{\mathrm{co}_k}),\tag{7}$$

式中:m为该集群中主机总个数;criticality表示被攻击的目标的资源损失;InCost表示网络系统完整性,ConCost表示网络系统机密性[21];P_{AV_k}表示k的主机可用性权重;P_{in_k}表示主机k的完整性权重;P_{co_k}表示主机k的机密性权重。主机权重WH_k、可用性权重、完整性权重、机密性权重由安全员根据经验给出,RNet表示网络系统的损失度。将每一个状态求出的RNet值作为此时网络态势值。

2.1.2 攻击成本

攻击成本量化了每次攻击时攻击者的付出,包括攻击消耗时间、攻击操作成本、计算机资源损耗及攻击者自身学习成本,法律风险等[22]。由于目前网络攻击法律风险较小,可忽略不计,暂不考虑攻击造成的法律风险,只考虑攻击操作成本AOP($\phi\boldsymbol{a}_{c1},\phi\boldsymbol{a}_{c2},\cdots,\phi\boldsymbol{a}_{cc}$),由于多个攻击者在某一状态进行不同攻击,同一状态不同攻击时间间隔较近,因此,看作同一状态下多人同时攻击对网络造成了影响,影响结果为当前CPU时间CPUt(t)(min)、网络带宽占用率Net(t)、内存的变化ΔMem,如公式(8)所示

$$\mathrm{AtCost}(t)=\mathrm{AOP}(\phi\boldsymbol{a}_{i1},\phi\boldsymbol{a}_{i2},\cdots,\phi\boldsymbol{a}_{ic})=\frac{\mathrm{CPUt}(t)}{100}+100\mathrm{Net}(t)+\Delta\mathrm{Mem}。\tag{8}$$

2.1.3 防御成本

防御成本指防御方采取防御操作时产生的自身成本,防御成本规定为人力成本、残余成本和负面成本之和,人力成本(human cost, HCost)表示防御者在采取防御措施时消耗的时间和计算资源以及人力权重,人力权重由人为给予,取值为[0,1],如公式(10);负面成本(negative cost, NCost)表示采取关闭服务或系统导致的系统异常运行,包括:系统异常、系统无法维持服务等消极影响,如公式(11)。其中,$\boldsymbol{r}(\phi\boldsymbol{d}_{j1},\cdots,\phi\boldsymbol{d}_{jv},\phi\boldsymbol{a}_{i1},\cdots,\phi\boldsymbol{a}_{ic})$表示防御策略组($\phi\boldsymbol{d}_{j1},\phi\boldsymbol{d}_{j2},\cdots,\phi\boldsymbol{d}_{jv}$)用来抵御攻击($\phi\boldsymbol{a}_{i1},\phi\boldsymbol{a}_{i2},\cdots,\phi\boldsymbol{a}_{ic}$)对系统可用性造成的负面影响程度,取值为[0,1];残余成本(residual cost,RsCost)用公式(12)表示,其中$\boldsymbol{e}(\phi\boldsymbol{d}_{j1},\cdots,\phi\boldsymbol{d}_{jv},\phi\boldsymbol{a}_{i1},\cdots,\phi\boldsymbol{a}_{ic})$表示防御策略组合($\phi\boldsymbol{d}_{j1},\phi\boldsymbol{d}_{j2},\cdots,\phi\boldsymbol{d}_{jv}$)对攻击策略组合($\phi\boldsymbol{a}_{i1},\phi\boldsymbol{a}_{i2},\cdots,\phi\boldsymbol{a}_{ic}$)造成的残余系统损失程度,取值为[0,1][23]。

$$\mathrm{HCost}(t)=\mathrm{Host}_1(t)+\mathrm{Host}_2(t)+\cdots+\mathrm{Host}_v(t)\left(\mathrm{Host}_i(t)\in[0,1]\right),\tag{9}$$

$$\mathrm{NCost}(t)=\mathrm{AtCost}(t)\times\boldsymbol{r}\left(\phi\boldsymbol{d}_{j1},\cdots,\phi\boldsymbol{d}_{jv},\phi\boldsymbol{a}_{i1},\cdots,\phi\boldsymbol{a}_{ic}\right),\tag{10}$$

$$\mathrm{RsCost}(t)=\mathrm{SysCost}(t)\times\boldsymbol{e}\left(\phi\boldsymbol{d}_{j1},\cdots,\phi\boldsymbol{d}_{jv},\phi\boldsymbol{a}_{i1},\cdots,\phi\boldsymbol{a}_{ic}\right),\tag{11}$$

$$\mathrm{DfCost}(t)=\mathrm{HCost}(t)+\mathrm{RsCost}(t)+\mathrm{NCost}(t)。\tag{12}$$

2.1.4 防御回报

在真实网络中,防御者对攻击采取防御策略,即使该防御手段防御效果不佳,也能收集到一些可用信息,对防御者分析之后的攻击路径、攻击意图、攻击规模、攻击规律、攻击实力等有一定意义,在某种程度上对之后的攻击行为形成一定威慑,侧面提升了防御者收益。研究将某状态的多人防御者采取的防御手段看作是同时进行,整体计算一个状态结束时的防御回报。防御回报(defense return,DfRe)分成5大类,具体如表1所列。

表1 防御回报描述

Table 1 Defensive return

防御回报点描述	防御回报点 DfRe
防御措施能够完全抵挡住攻击	5
防御措施能够抵挡大部分攻击	4
防御措施能够保护部分资产,抵御一定程度的攻击	3
防御小部分攻击,攻击者已侵犯大部分资产,系统损失较大	2
防御措施对攻击行为防御效果可以忽略,系统受到严重损害	1

2.1.5 状态转移成本

由于网络态势感知模型结合了Q-Learning算法,每受到一次攻击行为就会产生一次状态转移,在攻防过程中,网络状态是不断转移的。状态转移成本(reverse cost, RvCost)指的是每一次攻击者采取攻击措施时系统受到的损害差值,一个状态结束时的状态转移成本为在此状态受到的2×10^4多人攻击所造成的状态转移成本之和。

2.2 基于博弈论的定量分析

根据现实网络攻击经验,非理性攻击者占比少于理性攻击者,非理性攻击者经过100次攻击迭代最终一

定会遵照理性攻击者的策略,因此忽略非理性攻击者。研究将博弈论和强化学习中多状态的Q-learning算法引入网络攻防动态感知中,由收集的资源信息及攻防动作生成攻防博弈图,根据攻防策略、基于博弈论思想,建立结合强化学习的网络态势感知5元组模型QGNSAM={$\boldsymbol{N},\phi,\boldsymbol{P},\boldsymbol{U},S$},各元素的具体描述如下:

1)攻防代理人$\boldsymbol{N}=(\boldsymbol{N}^1,\boldsymbol{N}^2)$,$\boldsymbol{N}^1$表示攻击者的代理空间,$\boldsymbol{N}^2$表示防御者的代理空间,$\boldsymbol{N}$为双方的博弈空间。其中:$\boldsymbol{N}^1=(\boldsymbol{N}_1^1,\boldsymbol{N}_2^1,\cdots,\boldsymbol{N}_c^1)$,$\boldsymbol{N}^2=(\boldsymbol{N}_1^2,\boldsymbol{N}_2^2,\cdots,\boldsymbol{N}_v^2)$,分别代表攻击方参与者和防御方参与者;$c$为攻击者人数;$v$为防御者人数。

2)策略集合。$\phi=(\phi\boldsymbol{A},\phi\boldsymbol{B})$,攻击者$\boldsymbol{N}^1$策略集合$A$中攻击分为6个大类,$\phi\boldsymbol{A}=(\phi\boldsymbol{A}_1,\phi\boldsymbol{A}_2,\phi\boldsymbol{A}_3,\phi\boldsymbol{A}_4,\phi\boldsymbol{A}_5,\phi\boldsymbol{A}_6)$,攻击者策略总数记为$m_1$,包含了U2R、DATA、USER、DOS、PROBING和其他类别的攻击,包括不响应,攻击描述及威胁度如表2所列。

表2　攻击及攻击威胁度描述

Table 2　Attack and threat level

攻击	类型	描述	攻击威胁等级D
$\phi\boldsymbol{A}_1$	U2R	未授权的本地超级用户特权访问	10.0
$\phi\boldsymbol{A}_2$	DATA	数据非法被篡改或拷贝	5.0
$\phi\boldsymbol{A}_3$	USER	非法获得普通用户权限	1.0
$\phi\boldsymbol{A}_4$	DOS	拒绝服务攻击	5.0
$\phi\boldsymbol{A}_5$	PROBING	端口监视或扫描	0.5
$\phi\boldsymbol{A}_6$	Others	其他攻击,包括不响应	—

防御方$\boldsymbol{N}^2$策略集合为$\phi\boldsymbol{B}$,包括升级补丁、关闭服务、切断网络、扫描病毒漏洞、不响应等,$\boldsymbol{N}^2$策略总数设为m_2。

3)概率分布。$\boldsymbol{P}=\{\boldsymbol{p}(\boldsymbol{A}),\boldsymbol{p}(\boldsymbol{D})\}$表示攻防多代理人的状态转移策略概率分布,其中,$\boldsymbol{p}(\boldsymbol{A})$表示攻击策略概率集合,$\boldsymbol{p}(\boldsymbol{D})$表示防御策略概率集合,$\boldsymbol{P}=\{\boldsymbol{p}(\boldsymbol{A}),\boldsymbol{p}(\boldsymbol{D})\}$,$\boldsymbol{p}(\boldsymbol{A})=[\boldsymbol{p}^1(\boldsymbol{A}),\boldsymbol{p}^2(\boldsymbol{A}),\cdots,\boldsymbol{p}^c(\boldsymbol{A})]$表示$c$个攻击参与者的各自的策略集合,防御策略概率集合同理。

$$\boldsymbol{p}^{j1}(\boldsymbol{A})=[\boldsymbol{p}^{j1}(\phi\boldsymbol{a}_1),\cdots,\boldsymbol{p}^{j1}(\phi\boldsymbol{a}_i),\cdots,\boldsymbol{p}^{j1}(\phi\boldsymbol{a}_{m_1})],\tag{13}$$

$$\boldsymbol{p}^{j2}(\boldsymbol{D})=[\boldsymbol{p}^{j2}(\phi\boldsymbol{d}_1),\cdots,\boldsymbol{p}^{j2}(\phi\boldsymbol{d}_i),\cdots,\boldsymbol{p}^{j2}(\phi\boldsymbol{d}_{m_2})],\tag{14}$$

$$\sum_{i=1}^{m_1}\boldsymbol{p}^{j1}(\phi\boldsymbol{a}_i)=1,\ \sum_{j=1}^{m_2}\boldsymbol{p}^{j2}(\phi\boldsymbol{d}_j)=1,$$

式中:$\boldsymbol{p}^{j1}(\phi\boldsymbol{a}_{i1})$表示攻击者$j1$选取第$i1$攻击策略的概率;$\boldsymbol{p}^{j2}(\phi\boldsymbol{d}_{i2})$表示防御者$j2$选取第$i2$个防御策略的概率;$m_1$为攻击策略总数;$m_2$为防御策略总数。在采用Q-learning结合博弈算法时,需多次解决纳什均衡问题,普遍采用ε-贪婪算法或Boltzmann概率分布法[24]。由于贪婪算法是指代理人在每个时间节点,选择受益最高的策略,该方法易陷入局部极值。而Boltzmann概率分布法通过概率来选择行动,和Q-learning结合起来具有一定的自适应学习能力。Boltzmann概率分布规定在t状态下,攻击方$c1$选择行动$\phi\boldsymbol{a}_i$的概率如公式(15),防御者同理。

$$\boldsymbol{p}^{c1}(\phi\boldsymbol{a}_i)=\mathrm{e}^{\boldsymbol{Q}^{c1}(S,\phi\boldsymbol{a}_i)/\lambda}/\sum_{\phi\boldsymbol{a}\in\boldsymbol{A}}\mathrm{e}^{\boldsymbol{Q}^{c1}(S,\phi\boldsymbol{a}_i)/\lambda}。\tag{15}$$

公式(15)为博弈阶段t的概率分布函数,λ代表了博弈中代理人选择策略的随机性,随着λ增大而变大,$\lambda=5\times0.999^t$。每个参与者都需要计算自己的纳什均衡,并拥有不一定相同的策略概率分布。

4)支付函数。当攻击者们选择策略$\phi\boldsymbol{a}_{c1},\phi\boldsymbol{a}_{c2},\cdots,\phi\boldsymbol{a}_{cc}$,防御者选择策略$\phi\boldsymbol{d}_{v1},\phi\boldsymbol{d}_{v2},\cdots,\phi\boldsymbol{d}_{vv}$时,博弈空间存在一个奖赏对偶$(\boldsymbol{U}_{A_1}^a,\ \boldsymbol{U}_{D_1}^d)$,$\boldsymbol{U}_{A_1}^a$表示攻击者在$(\phi\boldsymbol{a}_{i1},\phi\boldsymbol{a}_{i2},\cdots,\phi\boldsymbol{a}_{ic})$策略组合时的期望收益,$\boldsymbol{U}_{D_1}^d$表示在$(\phi\boldsymbol{d}_{j1},\phi\boldsymbol{d}_{j2},\cdots,\phi\boldsymbol{d}_{jv})$策略组合防御者期望收益。如公式(18),矩阵$\boldsymbol{U}$中$\boldsymbol{a}_i$表示某一攻击者采取攻击$\phi\boldsymbol{a}_i$的博弈奖赏,$\boldsymbol{d}_i$表示某一防御者选择防御策略$\phi\boldsymbol{d}_i$的博弈奖赏,攻击收益计算如公式(16),防御收益如公式(17)

$$\boldsymbol{a}_i=\mathrm{SyCost}(t)-\mathrm{DfCost}(t)-\mathrm{AtCost}(t)-\mathrm{RvCost}(t),\tag{16}$$

$$\boldsymbol{d}_i = \mathrm{DfRt}(t) - \mathrm{SyCost}(t) - \mathrm{DfCost}(t) - \mathrm{RvCost}(t), \tag{17}$$

收益矩阵表示为

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{a}_1, \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_1, \boldsymbol{a}_{m_1} & \boldsymbol{a}_1, \boldsymbol{d}_1 & \cdots & \boldsymbol{a}_1, \boldsymbol{d}_{m_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{a}_{m_1}, \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_{m_1}, \boldsymbol{a}_{m_1} & \boldsymbol{a}_{m_1}, \boldsymbol{d}_1 & \cdots & \boldsymbol{a}_{m_1}, \boldsymbol{d}_{m_2} \end{bmatrix}。 \tag{18}$$

该收益矩阵适合当前环境下所有参与者的行动，当攻击者在状态t选取策略$(\phi\boldsymbol{a}_{c1}, \phi\boldsymbol{a}_{c2}, \cdots, \phi\boldsymbol{a}_{cc})$，防御者选取策略$(\phi\boldsymbol{d}_{v1}, \phi\boldsymbol{d}_{v2}, \cdots, \phi\boldsymbol{d}_{vv})$时，攻击者总收益为$\boldsymbol{U}_a(t)$，防御者总收益为$\boldsymbol{U}_d(t)$。若$\boldsymbol{U}_a(t) + \boldsymbol{U}_d(t) = 0$，该博弈为零和博弈，否则为非零和博弈。由于在真实网络中，攻防双方博弈除彼此的影响外，还有环境的影响，包括机器老化、网络延迟等因素，导致非零和博弈概率较大，即$\boldsymbol{U}_a(t)$与$\boldsymbol{U}_d(t)$之和不一定等于0，更为贴近现实。式(19)中$\boldsymbol{U}_{1i}^{a}(t)$表示在状态$t$时，攻击者1选定为$\phi\boldsymbol{a}_i$攻击策略攻击期望收益公式，防御期望收益$\boldsymbol{U}_{vj}^{d}(t)$，不再赘述。某一状态的攻击总收益为每个攻击者此状态下的攻击收益之和，即$\boldsymbol{U}_a(t) = \sum_{j=1}^{c} \boldsymbol{U}_j^a$。

$$\boldsymbol{U}_{1i}^{a}(t) = \boldsymbol{p}(\phi\boldsymbol{d}_1) \times \boldsymbol{a}_i + \boldsymbol{p}(\phi\boldsymbol{d}_2) \times \boldsymbol{a}_i + \cdots + \boldsymbol{p}(\phi\boldsymbol{d}_{m_1}) \times \boldsymbol{a}_i。 \tag{19}$$

5)状态空间。$S = (S_1, \cdots, S_t, \cdots, S_T)$，$S$表示的是目前博弈的状态空间，$S_t$表示此博弈空间的第$t$个博弈状态。

2.3 博弈中混合策略纳什均衡

混合策略纳什均衡是非零和博弈中的最稳定策略组合的一个概念解，也称非合作博弈均衡。混合策略是指任一代理人不清楚其他代理人的行动及偏好，不能采用固定的、明确的策略概率加以衡量，而是以一种随机的、最大化自己利益的概率分布衡量当前策略空间。混合策略纳什均衡对于攻击方或是防御方都是最优解。局中代理人总数为$(c+v)$人，即$\boldsymbol{N} = \{\boldsymbol{N}^1, \boldsymbol{N}^2\}$，策略分别为攻击策略防$\phi\boldsymbol{A} = (\phi\boldsymbol{a}_1, \phi\boldsymbol{a}_2, \cdots, \phi\boldsymbol{a}_{m_1})$御策略$\phi\boldsymbol{B} = (\phi\boldsymbol{d}_1, \phi\boldsymbol{d}_2, \cdots, \phi\boldsymbol{d}_{m_2})$，若任一方采取中$[0, c]$种策略或$(\phi\boldsymbol{d}_1, \cdots, \phi\boldsymbol{d}_{j-1}, \cdots, \phi\boldsymbol{d}_{j+1}, \cdots, \phi\boldsymbol{d}_{m_2})$中的$[0, v]$种策略，整体收益都较$(\phi\boldsymbol{d}_{j1}, \cdots, \phi\boldsymbol{d}_{jc}, \phi\boldsymbol{a}_{i1}, \cdots, \phi\boldsymbol{a}_{iv})$更低，则$(\phi\boldsymbol{d}_{j1}, \cdots, \phi\boldsymbol{d}_{jc}, \phi\boldsymbol{a}_{i1}, \cdots, \phi\boldsymbol{a}_{iv})$策略组合成为当前状态下的最稳定策略，达到纳什均衡，用公式表示为

$$\boldsymbol{U}_a \geqslant \boldsymbol{U}_{Ai}^{a}, (\boldsymbol{U}_a^i \text{表示其他策略组合}), \tag{20}$$

$$\boldsymbol{U}_d \geqslant \boldsymbol{U}_{Di}^{d}, (\boldsymbol{U}_d^i \text{表示其他策略组合}), \tag{21}$$

式中：$\boldsymbol{U}_{Ai}^{a}$表示攻击方的其他策略组合。最后利用线性规划方法可以求出混合策略纳什均衡的概念解。纳什均衡计算复杂度为$|m_1| \times |m_2|$。

3 结合博弈论与强化学习的网络攻防动态感知模型

由于多智能体与单智能体的差别，单纯的Q-learning算法难以适用于多智能体的情况，会造成难以收敛的问题，甚至与马尔科夫的静态不变性相违背，因此，独立的Q-learning算法在多智能体系统中失效。研究将Q-learning算法与博弈算法相结合，提出QGNSAM算法，突破单智能体的局限。QGNSAM分为2个部分，首先根据扫描的漏洞信息、攻防策略、网络系统配置、网络拓扑结构等生成网络攻击图，建立环境和博弈模型并应用Q-learning算法，随着状态改变，更新网络状态转移$\boldsymbol{Q}$学习矩阵，分析纳什均衡策略，直到达到目标状态，最终产生完整$\boldsymbol{Q}$学习矩阵，得到预测攻击路径与防御者的最优防御手段。

基于上述定量分析与计算，接下来将Q-learning算法引入，计算多参与者多状态的网络态势评估。研究对传统Q-learning进行改进，结合博弈算法，提出QGNSAM模型，该问题针对每个状态具有一组可操作策略。参与者在状态之间的移动称为动作，参与者在特定状态下采取行动会产生收益。针对每个参与者，强化学习目标是最大化自己的总回报。研究将在同一环境中的多参与者组成的强化学习博弈模型视为混合关系模型，即参与者与参与者之间非纯竞争关系或纯协作关系。参与者的每个动作的质量取决于对应收益，该收益是来自环境反馈。攻防行动是多人行动，笔者将参与者看作Q-learning算法学习的对象，将多人攻防行动看作多智能体学习，对此采用公式(22)的Q-learning博弈算法计算$\boldsymbol{Q}$值矩阵。初始$\boldsymbol{Q}$值矩阵为$(m_1 + m_2)^2$规模的零矩阵，每一状态的每个参与者都将更新一次自己的$\boldsymbol{Q}$值矩阵，研究设计所有参与者使用在完全相同的环

境中进行对抗和防御，因此，采用统一的 $\boldsymbol{Q}$ 值矩阵，且每个参与者在每一状态可以采取一个动作策略。

$$\boldsymbol{Q}^{z}(S_t,\phi\boldsymbol{a}_i)\leftarrow\times\boldsymbol{Q}^{z}(S_t,\hat{\phi}\boldsymbol{a}_{i1},\cdots,\phi\boldsymbol{a}_i,\cdots,\hat{\phi}\boldsymbol{a}_{ic},\hat{\phi}\boldsymbol{d}_{j1},\cdots,\hat{\phi}\boldsymbol{d}_{jv})\leftarrow(1-\alpha_t)\left(\sum_{i=1}^{c}\boldsymbol{Q}^{z}(\phi\boldsymbol{a}_i,\hat{\phi}\boldsymbol{a}_{ci})\right)+\alpha_t[\boldsymbol{r}_t(S_t,\phi\boldsymbol{a}_i)+\gamma\max_{a_j}\boldsymbol{Q}^{z}(S_t,\phi\boldsymbol{a}_j)](i=1,2,\cdots,c,\phi\boldsymbol{a}_j\in\phi\boldsymbol{A}),\tag{22}$$

其中：$\boldsymbol{Q}^{z}(S_t,\phi\boldsymbol{a}_i)$ 表示攻击方 z 在网络转移 S_t 状态采取 $\boldsymbol{a}_i$ 策略得到的 $\boldsymbol{Q}$ 学习矩阵更新，α_t 是介于 0，1 之间的学习速率参数，学习速率随着时间减小，越接近目标状态，收敛速度越小。若 α_t=0，则表示不更新 $\boldsymbol{Q}$ 值；$\boldsymbol{r}_t(S_t,\phi\boldsymbol{a}_i)$ 表示在 S_t 状态获得的瞬时奖赏，即本次博弈收益，$\max_{a_j}\boldsymbol{Q}(S_t,\phi\boldsymbol{a}_j)$ 表示攻击者 z 选择使下一个状态 $\boldsymbol{Q}$ 值最大的策略 $\phi\boldsymbol{a}_j$；γ 表示折现因子，越接近 0，表示攻击者倾向于考虑立即收益，越接近 1，代表攻击者更愿意延迟奖励；$\hat{\phi}\boldsymbol{a}_{i1}$ 表示在此次行动中，参与者 z 对攻击者 $c1$ 观测或评估的策略，$\hat{\phi}\boldsymbol{d}_{j1}$ 表示参与者 z 对防御者 $v1$ 观测或评估的策略总策略组 $(\hat{\phi}\boldsymbol{a}_{i1},\cdots,\phi\boldsymbol{a}_i,\cdots,\hat{\phi}\boldsymbol{a}_{ic},\hat{\phi}\boldsymbol{d}_{j1},\cdots,\hat{\phi}\boldsymbol{d}_{jv})$ 表示参与者 z 对其他参与者预测或评估的策略。

参与者一直探索并对目标系统采取动作，到达每一状态时检查是否到达目标主机或目标操作，若未达到，则开始下一轮攻击，若达到，算法收敛。每一组博弈必须满足混合策略纳什均衡。若未满足则重新选择。并重新计算并更新 $\boldsymbol{Q}$ 值矩阵。该过程复杂度为 $\left|(c+v)\times(m_1+m_2)\right|$，若为单智能体，只计算攻击方的 $\boldsymbol{Q}$ 值矩阵，复杂度为 $\left|m_1\right|$，可以看出，若采用多智能体博弈学习，复杂度增加明显。具体算法流程如算法 1 所示：

算法 1 结合博弈论与强化学习的网络攻防动态感知算法

输入：网络系统基础配置信息、攻防双方策略

输出：$\boldsymbol{Q}$ 学习矩阵（攻击者最优行动策略）

1. 初始化：扫描漏洞及网络系统基本配置信息、攻防策略、初始化 $\boldsymbol{Q}$ 学习矩阵 $\boldsymbol{Q}=0$、阶段 $t=0$、折现因子 γ、学习速率 a_0、每个防御者第 1 次采取各策略的概率、攻防收益、状态转移成本；
2. 利用工具得到网络攻击图；
3. while 当前状态！=目标状态 do；
4. 更新 S_t，网络状态 $S_t\in S$；
5. 利用改进的层次分析法计算网络系统态势值；
6. 利用 Boltzmann 概率分布法和此时的各参与者的 $\boldsymbol{Q}$ 学习矩阵计算采取每个策略的概率，计算当前网络状况下的混合策略纳什均衡，从攻防策略集合中各取一种策略作为一组攻防策略组合，再根据收益矩阵中对应收益值结合 Q-learning 算法调整 $\boldsymbol{Q}$ 矩阵。最后利用线性规划计算出纳什均衡；
7. 各参与者采取行动并观察结果；
8. 更新每个参与者的 $\boldsymbol{Q}$ 值矩阵（该公式为更新参与者 z 的 $\boldsymbol{Q}$ 值矩阵，应对不同的参与者执行，共执行 $c+v$ 次）

$$\boldsymbol{Q}^{z}(S_t,\phi\boldsymbol{a}_i)\leftarrow\times\boldsymbol{Q}^{z}(S_t,\hat{\phi}\boldsymbol{a}_{i1},\cdots,\phi\boldsymbol{a}_i,\cdots,\hat{\phi}\boldsymbol{a}_{ic},\hat{\phi}\boldsymbol{d}_{j1},\cdots,\hat{\phi}\boldsymbol{d}_{jv})\leftarrow(1-\alpha_t)(\sum_{i=1}^{c}\boldsymbol{Q}^{z}(\phi\boldsymbol{a}_i,\hat{\phi}\boldsymbol{a}_{ci}))+\alpha_t[\boldsymbol{r}_t(S_t,\phi\boldsymbol{a}_i)+\gamma\max_{j}\boldsymbol{Q}^{z}(S_t,\phi\boldsymbol{a}_j)](i=1,2,\cdots,c,\phi\boldsymbol{a}_j\in\phi\boldsymbol{A})$$

9. t=t+1，调整学习速率 α_t；
10. End While；
11. 结束并返回 $\boldsymbol{Q}$ 值矩阵。

4　实验分析

4.1　实验环境

为了验证提出的 QGNSAM 模型的可行性和有效性，搭建了如图 2 的小型实验网络。实验网络系统 NW，主要由 2 台防火墙，4 台主机 1 台工作站，2 个攻击者组成，2 个防火墙将实验环境分为 3 部分，攻击代理人所处网络为外部网络，主机 H1~H4 以及工作站 H5 处于内部网络，其中主机 H1 处于 DMZ 隔离区，其余 4 台主机及工作站处于可信网络区域。外部网络攻击代理人只能通过 80 端口访问 DMZ 区域中的 Web 服务器，并与 Web 管理员进行通信尝试与内部网络取得联系，再访问可信网络区域中的图形工作站。在可信网络区域中，服务器可以互相联络。从针对 5 台主机的攻击策略来看，攻击对主机 H1 中的 Web 服务，对主机 H2 中的数据

库服务，主机 H3 中的 FTP 服务记忆图形工作站的可用性造成了消极影响，而对系统的完整性与机密性影响可忽略不计。因此，这 4 台主机的 P_{AV_k}=1，而完整性和机密性 P_{in_k}=P_{co_k}=0，对于负责认证服务的主机 H4，攻击主要对 H4 的机密性造成了影响，对主机可用性和完整性影响较小，由于研究采用了层次分析法，可用性不能为 0，否则会造成服务层损失计算无效，因此，将可用性 P_{AV_k} 设为 0.2，机密性 P_{co_k}=1，完整性 P_{in_k}=0。并将主机 H1 的目标资源损失值 criticality 设为 4，主机 H2 和主机 H4 的目标资源损失值设为 4。

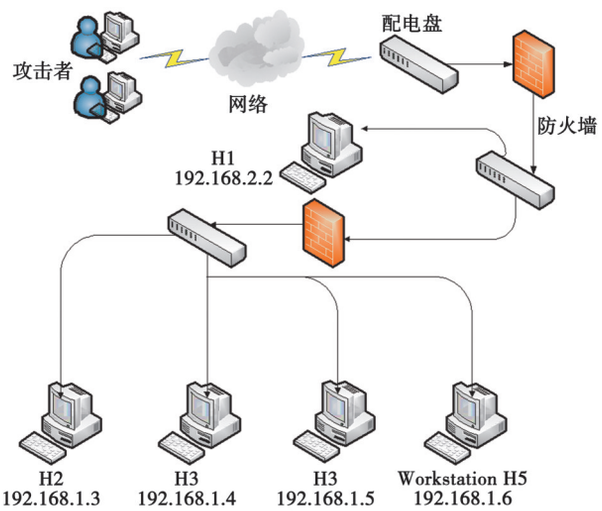

图 2 实验网络拓扑结构

Fig. 2 Experimental network topology

4.2 实验数据

利用 Nessue 扫描 NW 网络系统，得出该系统中主机权重与服务分布如表 3 所列，表 4 展示了各服务的权重，并假定攻击者最终攻击目标为主机 H4 的认证服务器。

表 3 主机及服务描述

Table 3 Hosts description

主机号	IP	服务名称	主机权值
H1	192.168.2.2	Web	2
H2	192.168.1.3	Database	3
H3	192.168.1.4	FTP	3
H4	192.168.1.5	Authentication	5
H5	192.168.1.6	graphic software	3

表 4 服务描述

Table 4 Services description

服务名称	服务描述	服务权重
Web	Apache-1.3.20	1
Database	Mysql -5.5.53	3
FTP	Wu-ftpd-2.6.1-18	2
Authentication	Cacert.pem	3
graphic server	—	1

实验网络中主机损失衡量指标分别为CPU占用率、内存利用率、进程状态和磁盘利用率4类。根据经验对指标重要程度评定,建立优先级矩阵$\boldsymbol{F}$,并利用公式(5)计算指标权重。根据表5计算得出上述4个指标在目标网络系统中的权重$\boldsymbol{w}$ = (0.333 3,0.125 0,0.208 3,0.333 3)。接下来,使用MulVal开源工具生成对应网络攻击图。结合历史资料及经验,可以得到本文攻击代理人可选攻击策略及其策略威胁程度,如表6所示。根据实验过程中每一步的网络状态的转移,分析防御代理人共有5种防御手段,如表7所示。状态转移成本考虑能够选取的攻防策略分别导致的系统损失,为了简化计算,事先计算状态转移成本如表8所示。

表5　主机指标权重

Table 5　Host indicators weight

F	CPU	Mem	Pro	Dis
CPU	0.5	1.0	1.0	0.5
Mem	0.0	0.5	0.0	0.0
Pro	0.0	1.0	0.5	0.0
Dis	0.5	1.0	1.0	0.5

表6　攻击策略描述

Table 6　Attack strategy description

攻击策略	攻击策略描述	类别	攻击威胁程度
$\phi\boldsymbol{a}_1$	Apache Graphical Interface Cross-Site-Script	U2R	10.0
$\phi\boldsymbol{a}_2$	Stack based buffer overruns	User	1.0
$\phi\boldsymbol{a}_3$	Oracle TNS Listener	Probing	0.5
$\phi\boldsymbol{a}_4$	Wu-FtpdSockPrintf()	U2R	10.0
$\phi\boldsymbol{a}_5$	IPsweep	Probing	0.1
$\phi\boldsymbol{a}_6$	Remote Denial of Service	DOS	1.0
$\phi\boldsymbol{a}_7$	Stealing Data	Data	l.0
$\phi\boldsymbol{a}_8$	Bypass Authentication	User	10.0
ϕ	Unresponsive	—	0

表7　防御策略描述

Table 7　Defense strategy description

防御策略	防御策略描述
$\phi\boldsymbol{d}_1$	Upgrade Patch
$\phi\boldsymbol{d}_2$	Close Services
$\phi\boldsymbol{d}_3$	Scan Virus
$\phi\boldsymbol{d}_4$	Drop Suspicious Packets
ϕ	Unresponsive

根据实验网络拓扑图及扫描漏洞信息等,可以得出实验网络被攻击后经历的7种状态S_1->S_7,状态集合表示为S={S_1(正常状态),S_2(主机H1root权限被攻击),S_3(主机H2user权限被攻击),S_4(主机H2被监听),S_5(主机H3root权限被攻击),S_6(主机H5user权限被攻击),S_7(主机H4root权限被攻击)}。主机攻击路径为H1->H2->H3->H5->H4,最终获取主机H4认证服务器的root权限。

随着攻防博弈过程的进行,$\boldsymbol{Q}$学习矩阵不断更新,取S_1阶段的$\boldsymbol{Q}$学习矩阵进行分析,此时可采用的攻防

策略集合如公式(23)中的$\phi \boldsymbol{a}(1)$和$\phi \boldsymbol{d}(1)$,矩阵(24)为攻击者1的$\boldsymbol{Q}$值矩阵,本次实验检测到攻击者第1步采取了Oracle TNS Listener行动,即$\boldsymbol{a}_3^1$,可以看到根据$\boldsymbol{Q}$值矩阵,攻击者1预测选取的下一步行动为$\boldsymbol{a}_2^1$,即Stack based buffer overruns。同理,也创建了攻击者2和2个防御者的$\boldsymbol{Q}$值矩阵为接下来的行动作出学习与决定。

表8 状态转移成本描述

Table 8 State transition cost

策略	$\phi \boldsymbol{d}_1$	$\phi \boldsymbol{d}_2$	$\phi \boldsymbol{d}_3$	$\phi \boldsymbol{d}_4$	ϕ
$\phi \boldsymbol{a}_1$	0.60	0.60	0.90	0.80	1.00
$\phi \boldsymbol{a}_2$	0.50	0.50	0.80	0.80	1.00
$\phi \boldsymbol{a}_3$	0.20	0.80	0.50	0.70	1.00
$\phi \boldsymbol{a}_4$	0.30	0.50	0.90	0.80	1.00
$\phi \boldsymbol{a}_5$	0.20	0.70	0.50	0.70	1.00
$\phi \boldsymbol{a}_6$	0.90	0.60	0.50	0.90	1.00
$\phi \boldsymbol{a}_7$	0.50	0.70	0.90	0.50	1.00
$\phi \boldsymbol{a}_8$	0.50	0.50	0.80	0.8	1.00
ϕ	0.05	0.05	0.05	0.05	0.05

S_1: $\phi \boldsymbol{a}(1) = \{ \boldsymbol{a}_1^1, \boldsymbol{a}_2^1, \boldsymbol{a}_3^1\boldsymbol{a}_4^1\}$, $= \{ \phi \boldsymbol{a}_1, \phi \boldsymbol{a}_2, \phi \boldsymbol{a}_3, \phi \boldsymbol{a}_4\}$= {Apache Graphical Interface Cross-Site-Script, IPsweep, Unresponsive, Oracle TNS Listener}

$\phi \boldsymbol{d}(1)$={ $\boldsymbol{d}_1^1, \boldsymbol{d}_2^1, \boldsymbol{d}_3^1, \boldsymbol{d}_4^1$ }={ $\phi \boldsymbol{d}_1, \phi \boldsymbol{d}_2, \phi \boldsymbol{d}_3, \phi \boldsymbol{d}_4$ }=

{Upgrade Patch, Close Services, Unresponsive, Scan Virus},

(23)

$$
\begin{array}{c}
 \\
\boldsymbol{a}_1^1 \\
\boldsymbol{a}_2^1 \\
\boldsymbol{a}_3^1 \\
\boldsymbol{a}_4^1 \\
\boldsymbol{a}_5^1 \\
\boldsymbol{a}_6^1 \\
\boldsymbol{a}_7^1 \\
\boldsymbol{a}_8^1 \\
\text{null}
\end{array}
\begin{array}{c}
\begin{array}{ccccccccc}
\boldsymbol{a}_1^1 & \boldsymbol{a}_2^1 & \boldsymbol{a}_3^1 & \boldsymbol{a}_4^1 & \boldsymbol{a}_5^1 & \boldsymbol{a}_6^1 & \boldsymbol{a}_7^1 & \boldsymbol{a}_8^1 & \text{null}
\end{array} \\
\begin{bmatrix}
1.36 & 4.72 & 3.24 & 3.60 & 2.12 & 3.11 & 1.74 & 1.33 & 0.54 \\
3.10 & 2.14 & 2.26 & 4.52 & 4.10 & 3.27 & 1.26 & 1.24 & 1.35 \\
2.13 & 3.23 & 2.32 & 1.87 & 1.09 & 2.06 & 2.13 & 1.87 & 0.56 \\
3.73 & 3.22 & 2.34 & 2.70 & 1.92 & 1.75 & 2.89 & 1.00 & 0.02 \\
2.87 & 3.78 & 1.20 & 1.41 & 2.48 & 4.49 & 2.56 & 1.22 & 0.89 \\
2.66 & 3.01 & 3.08 & 3.11 & 2.09 & 1.91 & 2.99 & 2.54 & 0.04 \\
1.32 & 1.56 & 1.76 & 1.23 & 2.32 & 2.44 & 1.29 & 1.57 & 1.01 \\
1.74 & 1.23 & 1.32 & 4.44 & 4.47 & 1.32 & 2.99 & 0.23 & 0.17 \\
4.31 & 2.26 & 4.42 & 3.27 & 1.17 & 1.09 & 1.32 & 4.01 & 0.01
\end{bmatrix}
\end{array}
\text{。} \qquad (24)
$$

对S_1状态得到的$\boldsymbol{Q}$学习矩阵执行混合策略纳什均衡,利用线性规划,得到混合策略纳什均衡:$\boldsymbol{p}^1(\boldsymbol{A}) = [0.31, 0.22, 0.19, 0.28]$, $\boldsymbol{p}^2(\boldsymbol{A}) = [0.11, 0.40, 0.23, 0.26]$, $\boldsymbol{p}^1(\boldsymbol{D}) = [0.24, 0.17, 0.17, 0.40]$, $\boldsymbol{p}^2(\boldsymbol{D}) = [0.32, 0.19, 0.34, 0.15]$。最后得出攻防博弈过程中达到稳定后每个状态的攻击收益与防御收益,如图3所示。

最终得出每个状态下的多人攻击行为对网络系统造成的态势影响以及多人防御者受益变化如图4所示。

从图3可以看出攻击收益与防御收益之和不一定为0,是由于状态转移需要成本,且系统中其他环境也会减少攻防总收益。图3为状态S_1,系统受到攻击,防御方采取策略但影响不大,导致攻击方收益下降不明显,防御方收益损失依旧严重。如图4中可见,状态S_1,网络态势值为0.645 2,实际网络态势值为0.655 6,可以看出AHP预测网络态势值较实际值更低,且相较其余2个算法,AHP在起始状态预测偏差较大;状态S_1到状态S_4时,防御方计算混

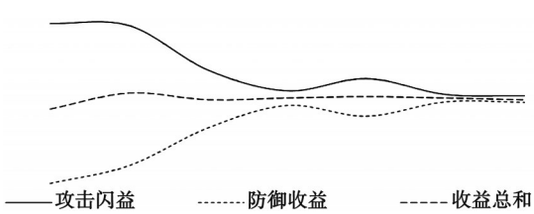

图3 攻防收益对比图

Fig.3 Benefits of attackers and defenders

合策略纳什均衡选取策略有效,在S_4状态时,防御方收益损失减小,上升接近0,同时攻击方收益明显减小,且接近0值(S_4状态时攻击收益为200,防御收益为-211),在S_4状态之后,攻击方再次发起进攻,攻击收益缓慢增加,由于受到有效攻击,此时网络态势值增加,在S_5状态达到态势局部极值4.556 2,在此过程中,研究方法较真实态势值略高,由于采用Q-learning方法进行预测,考虑下一步行动选择,有助于管理员做好充分防御准备。但防御方在S_5状态时紧急采取防御行动,防御效果明显并持久,因此,在S_6状态时将网络态势值剧烈下降至2.452。状态S_6,防御者收益逐渐接近0,攻击者收益逐渐下降为96,图3中状态S_7双方收益趋于稳定,说明防御策略已达到效果。图4中,状态S_7的网络态势值缓慢下降至1.102,说明防御方成功阻止了攻击方对系统的损害,系统逐渐趋于安全且稳定,证明该模型有效。实验中4种方法精度如图5所示。

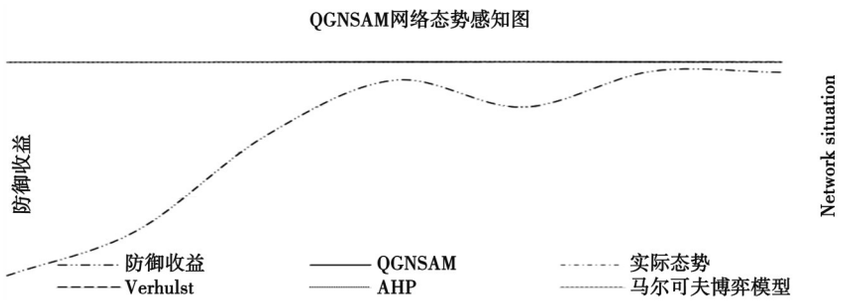

图4　各模型网络态势感知图

Fig.4　Network situation awareness of each algorithm

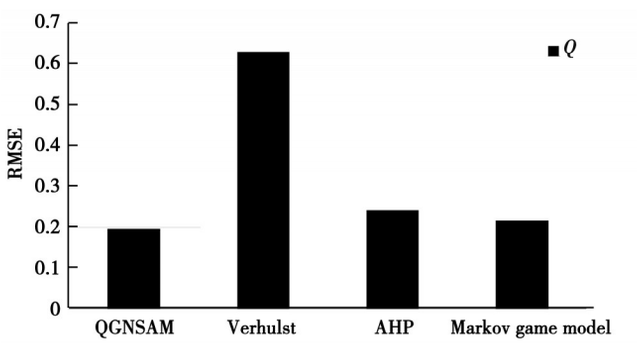

图5　各算法的方根误差

Fig.5　RMSE of each algorithms

从图5可以看出,研究方法较灰色Verhulst和AHP层次分析法预测结果方根误差RMSE明显更低,且略低于Markov Game Model方法,说明提出的方法精度更大,与实际结果更为贴近。用所提方法可实时、动态预测下一个状态的攻防策略,随时可以根据此时系统状况更新预测态势,更好地为网络安全人员提供有效的防护措施。

5　结束语

笔者提出结合博弈论与强化学习的网络动态感知模型QGNSAM,建立了主机层、服务层和网络层的网络安全状况指标体系,采用了改进的模糊层次分析法对网络安全状况指标进行量化。并采用Q-learning算法对网络系统状态转移过程进行监视与分析,用玻尔兹曼概率分布算法计算每个状态下的纳什均衡稳定策略,使双方收益最大化,且策略预测稳定,实现对网络态势的分析。在实验分析部分,以收益可视化界面展示多人攻防行动中攻防双方收益以及收益之和,利用Q-learning学习转移算法得出攻击预测路径,使安全人员对当前网络状况更加清楚。目前研究只考虑了攻击者和防御者2类博弈代理人,没有加上用户第3方博弈,对攻击代理人做$\boldsymbol{Q}$学习矩阵状态转移分析,未来将致力于考虑包括用户方的3方多人博弈与Q-learning算法解决网络态势问题,并将该模型应用到更加复杂的网络环境中。

参考文献

[1] Coffey K, Smith R, Maglaras L, et al. Vulnerability analysis of network scanning on SCADA systems[J]. Security and Communication Networks, 2018, 2018(1): 3794603.

[2] Wang J, Gao Y, Liu W, et al. An intelligent data gathering schema with data fusion supported for mobile sink in wireless sensor networks[J]. International Journal of Distributed Sensor Networks, 2019, 15(3): 155014771983958.

[3] Gao Y, Zhang S Y. A network security situation awareness method based on multi-source information fusion[C]//2nd International Forum on Management, Education and Information Technology Application. Shenzhen, China: Atlantis Press, 2018(2): 273-276.

[4] 龚正虎, 卓莹. 网络态势感知研究[J]. 软件学报, 2010, 21(7): 1605-1619.
Gong Z H, Zhuo Y. Research on cyberspace situational awareness[J]. Journal of Software, 2010, 21(7): 1605-1619. (in Chinese)

[5] Xin Z, Qiang L. Research on situation forecast of cyberspace situation awareness[C]//2016 Sixth International Conference on Instrumentation & Measurement, Computer, Communication and Control. Harbin, China: IEEE, 2016: 140-144.

[6] Han W H, Tian Z H, Huang Z Z, et al. System architecture and key technologies of network security situation awareness system YHSAS[J]. Computers, Materials & Continua, 2019, 59(1): 167-180.

[7] Greiner D, Periaux J, Emperador J M, et al. Game theory based evolutionary algorithms: a review with Nash applications in structural engineering optimization problems[J]. Archives of Computational Methods in Engineering, 2017, 24(4): 703-750.

[8] Papadimitriou C, Piliouras G. From Nash equilibria to chain recurrent sets: an algorithmic solution concept for game theory[J]. Entropy, 2018, 20(10): 782.

[9] Yousefi M, Mtetwa N, Zhang Y, et al. A reinforcement learning approach for attack graph analysis[C]//2018 17th IEEE International Conference on Trust, Security and Privacy In Computing and Communications/12th IEEE International Conference on Big Data Science and Engineering. New York, USA: IEEE, 2018: 212-217.

[10] Barik M S, Mazumdar C, Gupta A. Network vulnerability analysis using a constrained graph data model[M]//Information Systems Security. Cham: Springer International Publishing, 2016: 263-282.

[11] Tao X L, Liu L Y, Zhao F, et al. Ontology and weighted D-S evidence theory-based vulnerability data fusion method[J]. Journal of Universal Computer Science, 2013, 25: 203-221.

[12] 丁华东, 许华虎, 段然, 等. 基于贝叶斯方法的网络安全态势感知模型[J]. 计算机工程, 2020, 46(6): 130-135.
Ding H D, Xu H H, Duan R, et al. Network security situation awareness model based on Bayesian method[J]. Computer Engineering, 2020, 46(6): 130-135. (in Chinese)

[13] Zhu L. A novel social network measurement and perception pattern based on a multi-agent and convolutional neural network [J]. Computers & Electrical Engineering, 2018, 66: 229-245.

[14] Zhao D M, Liu J X. Study on network security situation awareness based on particle swarm optimization algorithm[J]. Computers&Industrial Engineering, 2018, 125: 764-775.

[15] 胡柳, 周立前, 邓杰, 等. 基于支持向量机和自适应权重的网络安全态势评估模型[J]. 计算机系统应用, 2018, 27(7): 188-192.
Hu L, Zhou L Q, Deng J, et al. Evaluation model of network security situation based on support vector machine and self-adaptive weight[J]. Computer Systems&Applications, 2018, 27(7): 188-192. (in Chinese)

[16] 郭佳. 基于数学模型的网络安全态势感知综述[J]. 中国新技术新产品, 2016(22): 187.
Guo J. Summary of network security situation awareness based on mathematical model[J]. New Technology & New Products of China, 2016(22): 187. (in Chinese)

[17] 陈秀真, 郑庆华, 管晓宏, 等. 层次化网络安全威胁态势量化评估方法[J]. 软件学报, 2006, 17(4): 885-897.
Chen X Z, Zheng Q H, Guan X H, et al. Quantitative hierarchical threat evaluation model for network security[J]. Journal of Software, 2006, 17(4): 885-897. (in Chinese)

[18] 张勇, 谭小彬. 一种基于隐Markov模型的网络安全态势感知方法研究[J]. 信息网络安全, 2011, 11(10): 47-51.
Zhang Y, Tan X B. An approach to network security situation awareness based on hidden Markov model[J]. Netinfo Security, 2011, 11(10): 47-51. (in Chinese)

[19] 王华. 基于自适应灰色维尔斯模型的C4ISR安全态势预测方法[J]. 内蒙古师范大学学报(自然科学汉文版), 2018, 47(3): 232-236.
Wang H. Security situation forecasting method of C4ISR based on adaptive gray verhulst model[J]. Journal of Inner Mongolia

Normal University(Natural Science Edition), 2018, 47(3): 232-236. (in Chinese)

［20］刘景玮, 刘京菊, 陆余良, 等. 基于网络攻防博弈模型的最优防御策略选取方法[J]. 计算机科学, 2018, 45(6): 117-123.
Liu J W, Liu J J, Lu Y L, et al. Optimal defense strategy selection method based on network attack-defense game model[J]. Computer Science, 2018, 45(6): 117-123. (in Chinese)

［21］刘小虎, 张明清, 张玉臣, 等. DDoS 主动防御系统防御能力量化仿真研究[J]. 信息工程大学学报, 2015, 16(6): 760-764.
Liu X H, Zhang M Q, Zhang Y C, et al. Study on quantitative simulation of DDoS active defense system’s defense capability [J]. Journal of Information Engineering University, 2015, 16(6): 760-764. (in Chinese)

［22］翁芳雨. 基于随机博弈模型的网络安全态势评估与预测方法的研究与设计[D]. 北京: 北京邮电大学, 2018.
Weng F Y. Research and design of network security situation assessment and prediction method based on stochastic game model[D]. Beijing: Beijing University of Posts and Telecommunications, 2018. (in Chinese)

［23］曾丽娇. 基于攻防演化博弈的网络安全态势研究[D]. 西安: 西安电子科技大学, 2018.
Zeng L J. Research on network security situation based on attack and defense evolutionary game[D]. Xi’an: Xidian University, 2018. (in Chinese)

［24］Chung K, Kamhoua C A, Kwiat K A, et al. Game theory with learning for cyber security monitoring[C]//2016 IEEE 17th International Symposium on High Assurance Systems Engineering. Orlando, USA: IEEE, 2016: 1-8.

（编辑　侯　湘）