基于强化学习的一对多雷达干扰资源分配策略研究

尚　熙[1]，杨革文[2]，戴少怀[2]，蒋伊琳[1]

（1. 哈尔滨工程大学 信息与通信工程学院，黑龙江 哈尔滨　150001；2. 上海机电工程研究所，上海　201109）

摘　要：针对干扰机一对多情形下的干扰突防问题，提出了一种基于强化学习的一对多干扰情形下的干扰资源分配方法，引入干扰辐射能量比和突防距离比作为评价指标，并对DQN（deep Q network）和Dueling-DQN算法引入动态调整的奖励值以增强算法的收敛能力。结合一对多干扰突防场景，对两种算法进行了验证，实验结果验证了两种算法的可行性及差异性，实现了对于干扰资源在干扰功率、时长、干扰样式及干扰雷达选取的资源分配能力，满足了一对多情形下的干扰资源实时、动态的分配需求。

关键词：干扰资源分配；强化学习；干扰辐射能量；最大突防距离；动作分配

中图分类号：TN974；TP181　　**文献标志码：**A　　**文章编号：**2096-4641（2022）01-0094-08

Research on Resource Allocation Strategy of One-to-Many Radar Jamming Based on Reinforcement Learning

SHANG Xi[1], YANG Gewen[2], DAI Shaohuai[2], JIANG Yilin[1]

（1. School of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, Heilongjiang, China; 2. Shanghai Electro-Mechanical Engineering Institute, Shanghai 201109, China）

Abstract: Aiming at the interference penetration of the jammer in the case of one-jammer to multi-radar, a reinforcement learning-based interference resource allocation method in the case of one-jammer to multi-radar interference is proposed. The interference radiation energy ratio and penetration distance ratio are introduced as evaluation indicators, and the dynamically adjusted reward values are used for DQN (deep Q network) and Dueling-DQN algorithms to enhance the convergence ability of the algorithm. By building a one-jammer to multi-radar interference penetration scenario, DQN and Dueling-DQN algorithms were verified, the experimental results verify the feasibility and difference of the two algorithms, and realize the resource allocation ability for interference resources in interference power, duration, interference pattern and interference radar selection, and meet the real-time and dynamic interference resource allocation requirement in the case of one-jammer to multi-radar.

Keywords: jamming resource allocation; reinforcement learning; jamming radiation energy; maximum penetration distance; action allocation

0　引　言

目前，自适应雷达对抗技术[1]已经成为现代电子对抗研究的重点问题。在飞机进行突防的过程中，所面对的“地/海面雷达信号环境”大多是复杂、多变的；并且，随着多功能雷达的发展，当前雷达的工作模式[2]可以发生较大的变化，单一干扰策略的压制效果并不理想。当干扰方对雷达方进行干扰时，通常是处于信

收稿日期：2021-07-27；**修订日期：**2022-01-18

基金项目：上海航天科技创新基金（SAST2019-001）

作者简介：尚熙（1996—），男，硕士研究生，主要研究方向为雷达干扰资源调度。Email: shangxi@hrbeu.edu.cn

通信作者：蒋伊琳（1980—），男，博士，副教授，主要研究方向为图像处理与电子对抗。Email: jiangyilin@hrbeu.edu.cn

息非对称的情形，这样，就需要干扰方的干扰设备具有一对多的干扰能力[3]。现有的能够动态分配干扰资源策略且与环境进行互动变化的算法主要是强化学习，如：黄星源等[4]对信息对称情况下的多对多干扰资源分配问题进行了研究，但未涉及一对多干扰资源分配问题；周彬等[5]使用Q-learning算法对无人机路径规划问题进行了研究，但不适用于多状态及动作场景；刘松涛等[6]在对自适应干扰机的研究中，没有引入干扰辐射能量低、突防距离远的干扰资源分配理念；现有的DQN(deep Q network)算法[7-8]可应用于多状态及动作场景，已经被用于解决路径规划问题，但是对于雷达干扰资源的分配问题研究较少。

本文主要以突防距离比和干扰辐射能量比作为整个干扰对抗过程的评价标准，以要求突防距离最大和使用干扰辐射能量最少之间的反比关系互相牵制，采用引入动态调整奖励值的DQN和Dueling-DQN算法，分别记作DQN(R)和Dueling-DQN(R)，利用其合理地分配自适应干扰机的干扰资源，使得实施一对多干扰时的整体效益最大化，即：在研究复杂电磁环境下实施一对多干扰的压制效果基础上，对能量损耗和突防距离的最大化效益进行研究，并比较两种强化学习算法的优劣。

1　复杂电子对抗环境模型建立及资源分配评价标准

1.1　自适应干扰机

与传统干扰机相比，自适应干扰机[9]能够随着所接收到的参数信息实时地更改自身的干扰策略，合理、高效地对雷达实现压制干扰，根据干扰方的需求，可以满足干扰辐射总能量小、突防距离大的要求。传统干扰机往往采用单一的干扰样式和功率对雷达进行压制，只考虑压制成功与否，这种固定的、单一的策略很容易因多功能雷达工作参数的改变而达不到理想的压制效果，因此，自适应干扰机更加符合当前对抗技术的需求。自适应干扰机可以被理解为一个智能体，其接收的数据和采用的策略则可以被认为是与环境的交互，通过模拟人类大脑学习过程，对不同的事物做出不同的策略和自己的评价，最终使智能体可以面对任何环境做出合理的动作和评价。自适应干扰机的结构框图如图1所示，本文中干扰策略库主要的干扰样式有噪声调幅、噪声调频、灵巧噪声和密集假目标压制。

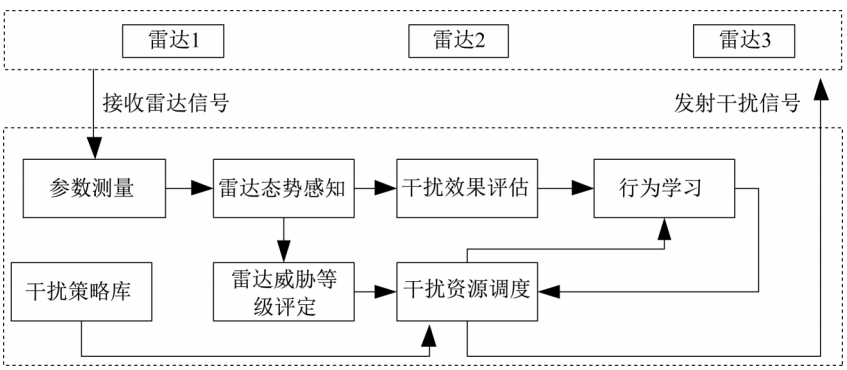

图1　自适应干扰机的结构框图

Fig. 1　Block diagram of the adaptive jammer

1.2　电子对抗场景

建立一对多的复杂电子对抗环境时，本文中进行对抗的雷达模型有搜索、跟踪、制导3种工作状态(具体状态转换关系见2.1节)，飞机突防开始状态默认雷达处于搜索状态，飞机突防失败状态默认雷达进入制导状态。因海杂波对雷达的探测性能影响较大，在整个突防的过程中引入海杂波的影响[10]，可参考文献[11]和文献[12]。考虑到随着自适应干扰机离雷达越来越近，需要考虑干扰信号从旁瓣进入的影响，可参考文献[13]。

雷达在杂波与干扰环境下的最大探测距离，如式(1)所示。

$$R_{\max}=\left[\frac{P_{\mathrm{t}}G_{\mathrm{t}}G_{\mathrm{r}}\lambda^{2}\sigma D}{(4\pi)^{3}\left(kT_{0}B_{\mathrm{n}}F_{\mathrm{n}}+P_{\mathrm{c}}+\dfrac{P_{\mathrm{j}}G_{\mathrm{j}}G_{\mathrm{t}}(\theta)\lambda^{2}\gamma_{\mathrm{j}}}{(4\pi)^{2}R_{\mathrm{j}}^{2}}\right)D_{\mathrm{s}}(n)L_{\mathrm{t}}}\right]^{\frac{1}{4}} \tag{1}$$

式中：k 为玻尔兹曼常数；T_0 为标准室温，通常取 290K；B_n 为接收机内部噪声带宽；F_n 为噪声系数；P_c 为接收到的海杂波功率；$D_s(n)$ 为接收机输出端测量的信噪比，即 n 个脉冲信噪比；G_j 为干扰机天线增益；$G_t(\theta)$ 为在偏离雷达 θ 角度时的接收增益；R_j 为干扰机到雷达的径向距离；γ_j 为极化因子；λ 为发射信号波长；σ 为目标散射截面积；D 为脉冲压缩比；L_t 为雷达接收机损耗；P_j 为干扰功率；P_t 为雷达发射功率；G_t 为雷达天线增益；G_r 为雷达接收天线增益。此处认为雷达接收信号方向为雷达天线的主瓣方向，故而有 $G_t=G_r=G$。

建立如图 2 所示的一对多情形下的电子对抗场景，其中，两雷达之间相距为 d，3 部雷达都建立在高为 H 的小岛顶部。飞机携带一部自适应干扰机从远方突防而来，自适应干扰机携带的干扰样式有噪声调幅干扰、噪声调频干扰、灵巧噪声干扰及密集假目标干扰。突防开始时，飞机与雷达 2 相距为 L，飞机飞行速度为 v，飞行高度为 H，飞行航迹指向雷达 2。以上述建立的电子对抗环境为背景，研究一部自适应干扰机在一对多情况下的干扰资源分配策略。

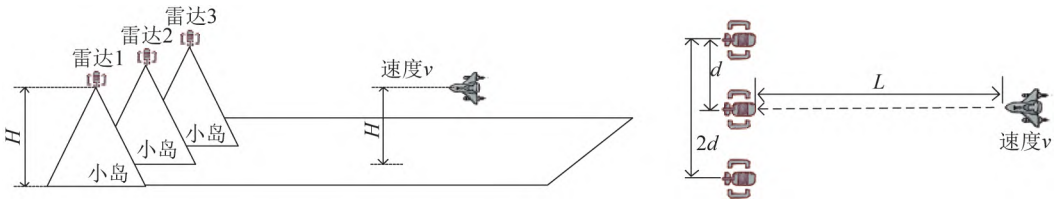

图 2 整体电子对抗场景示意图

Fig. 2 Schematic diagram of the overall electronic countermeasure scenario

1.3 干扰资源分配评价标准

1.3.1 整体干扰辐射能量占比

对于飞机携带自适应干扰机对雷达进行突防的过程，需要考虑的是利用现有资源使干扰辐射能量最小、突防距离最大，合理、动态完成干扰资源的分配。若整个需要突防的距离为 L，飞机速度为 v，则可以将整个过程分为 $q=L/v$ 步，每次重新分配干扰资源的时间步长为 1 s，在该时间步长内采用相同的干扰样式进行多次干扰。记 $E_{s1}\sim E_{s4}$ 分别为整个过程中各种压制的干扰辐射总能量，如表 1 所示。其中，约束条件有：$q=t_1+t_2+t_3+t_4$，$t_1\sim t_4$ 分别为每种干扰样式各自干扰的总时长；$t_{j1}\sim t_{j4}$ 分别为某一干扰样式且不同功率下的干扰时长；$P_{j1}\sim P_{j4}$ 分别为不同时刻采取不同干扰样式时选择的干扰功率大小。

基于此，提出整体干扰辐射能量占比 $\eta_p=E_u/E_m$，其中：$E_u=E_{s1}+E_{s2}+E_{s3}+E_{s4}$，称为实际干扰总能量，即突防过程中，干扰机总共辐射的干扰能量大小；$E_m=\max(P_j)\cdot(L/v)$，称为整体干扰总能量，即突防过程中总是施放干扰功率最大的干扰样式产生的总干扰辐射能量大小。根据定义，$0<\eta_p<1$，η_p 越大表明整个过程中干扰需要的总能量就越大，反之，需要的干扰总能量就越小。作为一对多干扰策略的目标值，η_p 越小越好。

总共4种干扰

表 1 整个过程中各种压制的干扰辐射能量计算方法

Tab. 1 The calculation method of various suppressed interference radiation energy in the whole process

压制干扰样式	干扰辐射能量
噪声调幅	$E_{s1}=\sum_{i_1=1}^{t_1}P_{j1}(i_1)\cdot t_{j1}(i_1)$
噪声调频	$E_{s2}=\sum_{i_2=1}^{t_2}P_{j2}(i_2)\cdot t_{j2}(i_2)$
灵巧噪声	$E_{s3}=\sum_{i_3=1}^{t_3}P_{j3}(i_3)\cdot t_{j3}(i_3)$
密集假目标	$E_{s4}=\sum_{i_4=1}^{t_4}P_{j4}(i_4)\cdot t_{j4}(i_4)$

记号不清晰，j表示干扰样式？i表示离散时刻

1.3.2 最大突防距离比

不能只从一个方面对整个雷达的压制干扰效果进行评价，因此，引入最大突防距离比 $\eta_d=L_u/L$，其中：L_u 表示战斗机携带干扰吊舱突防的最大距离；L 表示整个突防的距离。η_d 作为干扰效果的目标值，越大越好，与对 η_p 的要求相反。可以利用指标 η_d 和 η_p 对干扰资源分配的结果进行整体评价。

2 基于强化学习的干扰资源分配

2.1 马尔可夫建模

2.1.1 状态

在对雷达阵地的突防过程中，多功能雷达有多种

工作模式[14]。对于干扰方来说,多功能雷达主要有搜索模式、跟踪模式和制导模式。搜索模式是初始状态,制导模式是终止状态,进入制导模式后结束本次迭代。多功能雷达的工作状态变化可以用图3描述,记S_{ri}(i为雷达编号)为雷达状态值,用来描述雷达所处工作模式,搜索、跟踪、制导模式的S_{ri}值分别取0、1、2。本文假设3部雷达不进行组网,各自对目标进行独立探测。对于每部雷达,状态转换依据为:

1) 搜索状态下,如果4次探测中雷达有3次探测到目标,雷达状态进入跟踪状态,否则,保持搜索状态。

2) 跟踪状态下,如果3次探测中雷达有2次探测到目标,雷达状态从跟踪状态进入制导状态;若3次探测均未探测到目标,雷达状态返回搜索状态;否则,保持在跟踪状态。

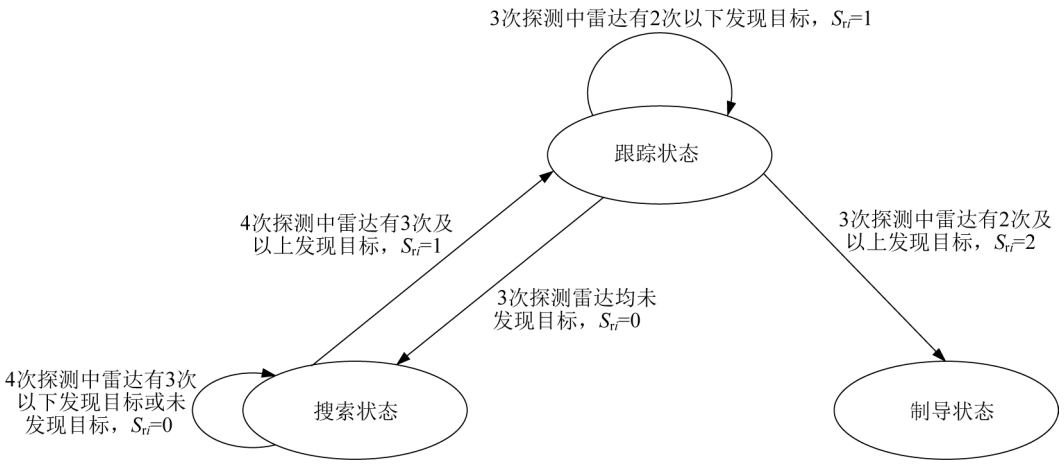

图3　雷达的工作模式转换及相应S_{ri}值变化示意图

Fig. 3　Schematic diagram of radar working mode conversion and corresponding S_{ri} changes

根据马尔可夫模型[15]建立状态函数$\boldsymbol{S}_t$

$$\boldsymbol{S}_t=\mathrm{CON}\left(R_{\mathrm{d}},\boldsymbol{S}_{\mathrm{r}}\right) \tag{2}$$

式中:R_{d}表示飞机离雷达2的径向距离;$\boldsymbol{S}_{\mathrm{r}}=[S_{\mathrm{r1}},S_{\mathrm{r2}},S_{\mathrm{r3}}]$为1×3的矩阵,包含3部雷达当前时刻各自的工作模式S_{ri},S_{ri}对应的值越大,则代表威胁等级越高,反之,则越小;CON(·)为连接函数。

2.1.2　动作

离散化

干扰动作主要从功率和干扰样式进行划分,自适应干扰机能够有效干扰的功率范围为$P_{\min}$~$P_{\max}$,若把有效干扰功率合理地划分成n个,那么,对于4种压制干扰样式,自适应干扰机可以采取的干扰动作就有$4n$种,每一种干扰动作对应不同功率下的某一种压制干扰方式。根据马尔可夫模型建立动作函数,可表达为

$$\boldsymbol{A}_t=\mathrm{CON}\left(i,a_t\right) \tag{3}$$

式中:i表示对第i部雷达进行干扰,本文i的取值为1到3;a_t表示该时刻采取$4n$中动作的一种。因此,$\boldsymbol{A}_t$表示对第i部雷达所采取$4n$中某一种干扰动作,为1×2的矩阵。

2.1.3　奖励

当奖励设置不合理时,会使得干扰策略分配难以快速收敛,导致智能体学习的速度大大降低,而合理的奖励值设置会使智能体可以快速地在与环境的交互中学习和收敛。因此,本文从3个方面设置奖励:每步干扰成功奖励、干扰功率奖励、干扰样式能量最小化奖励。

每步干扰成功奖励R_1设置为

$$R_1=\begin{cases}+1, & \boldsymbol{S}_{\mathrm{r}}\text{不变}\\ -5, & \boldsymbol{S}_{\mathrm{r}}\text{增大}\\ +5, & \boldsymbol{S}_{\mathrm{r}}\text{减小}\end{cases} \tag{4}$$

干扰功率奖励R_2设置为

$$R_2=\begin{cases}r_1, & P_{\mathrm{j}}=P_{\max}\\ \vdots & \vdots\\ r_{10}, & P_{\mathrm{j}}=P_{\min}\end{cases} \tag{5}$$

式中:r_1~r_{10}取值范围为$[-0.5,+0.5]$,r_1~r_5为正值,分别为−0.5、−0.4、−0.3、−0.2、−0.1,r_6~r_{10}为负值,分别为0.1、0.2、0.3、0.4、0.5。

干扰样式能量最小化R_3奖励设置为

$$R_3=\begin{cases}r_{\mathrm{j1}}, & R_{\mathrm{d}}\in(R_{\mathrm{d1}},L]\\ r_{\mathrm{j2}}, & R_{\mathrm{d}}\in(R_{\mathrm{d2}},R_{\mathrm{d1}}]\\ r_{\mathrm{j3}}, & R_{\mathrm{d}}\in(R_{\mathrm{d3}},R_{\mathrm{d2}}]\\ r_{\mathrm{j4}}, & R_{\mathrm{d}}\in(R_{\mathrm{d4}},R_{\mathrm{d3}}]\\ r_{\mathrm{j5}}, & R_{\mathrm{d}}\in[0,R_{\mathrm{d4}}]\end{cases} \tag{6}$$

$$r_{j1}=\begin{cases}5-\dfrac{5v}{L-R_{d1}}n_t, & \text{噪声调幅干扰}\\ -0.5, & \text{噪声调频干扰}\\ -0.5, & \text{灵巧噪声干扰}\\ -0.5, & \text{密集假目标干扰}\end{cases}\tag{7}$$

$$r_{j2}=\begin{cases}-0.5, & \text{噪声调幅干扰}\\ 5-\dfrac{5v}{R_{d1}-R_{d2}}n_t, & \text{噪声调频干扰}\\ -0.5, & \text{灵巧噪声干扰}\\ -0.5, & \text{密集假目标干扰}\end{cases}\tag{8}$$

$$r_{j3}=\begin{cases}-0.5, & \text{噪声调幅干扰}\\ -0.5, & \text{噪声调频干扰}\\ 5-\dfrac{5v}{R_{d3}-R_{d2}}n_t, & \text{灵巧噪声干扰}\\ -0.5, & \text{密集假目标干扰}\end{cases}\tag{9}$$

$$r_{j4}=\begin{cases}-0.5, & \text{噪声调幅干扰}\\ -0.5, & \text{噪声调频干扰}\\ -0.5, & \text{灵巧噪声干扰}\\ 5-\dfrac{5v}{R_{d4}-R_{d3}}n_t, & \text{密集假目标干扰}\end{cases}\tag{10}$$

$$r_{j5}=\begin{cases}-0.5, & \text{噪声调幅干扰}\\ -0.5, & \text{噪声调频干扰}\\ -0.5, & \text{灵巧噪声干扰}\\ +0.5, & \text{密集假目标干扰}\end{cases}\tag{11}$$

式中：$n_t=(L-R_d)/v$，表示飞机飞行到 n_t 步；$R_{d1}\sim R_{d4}$ 表示各个干扰样式最大功率下对雷达旁瓣干扰时雷达最大探测距离（当虚警概率为 10^{-6}、发现概率为 0.5 时）。参考第三章仿真参数，带入公式(1)可求得各种干扰下的雷达最大探测距离 $R_{d1}\sim R_{d4}$，$R_{d1}\sim R_{d4}$ 分别为 321 km、139 km、77 km、16 km。

综上，我们可得每步的总奖励值为 $R_{all}=R_1+R_2+R_3$。

2.2 基于DQN与Dueling-DQN算法的干扰资源分配策略

DQN算法是Q-Learning算法的改进，摒弃了Q-Learning算法中的Q表，利用深度神经网络（deep neural network，DNN）代替了Q表，可以适用于多状态-多动作的强化学习模型。DQN网络主要由当前网络、目标网络、环境、经验回放池、DQN误差函数构成，如图4所示。

如图4所示，其中 s、a、r、θ_r 分别代表状态、动作、奖励及网络参数，采用随机梯度下降法（stochastic gradient descent，SGD）更新网络参数 θ_r。DQN内部包含两个网络，分别是当前网络和目标网络，这两个网络的结构一模一样，均为DNN网络，把上文的状态 $\boldsymbol{S}_t$ 和动作 $\boldsymbol{A}_t$ 作为DNN网络的输入得到该状态和该动作下的值函数。控制目标网络的参数在一定的步数间隔内保持不变，把当前网络的参数直接复制给目标网络，而不是每帧都更新，目的是去除目标 Q 值和当前 Q 值的相关性，解决训练不稳定的问题，提高收敛成功率。其网络中的相关参数有：D_m（经验回放集合尺寸）、γ（奖励折扣因子）、r_L（学习率）、ε（ε-*greedy*贪婪因子，以 ε 的概率选择最优动作，$(1-\varepsilon)$ 的概率选择随机动作）、C（重置网络权重步数）、N（每次训练批量）。

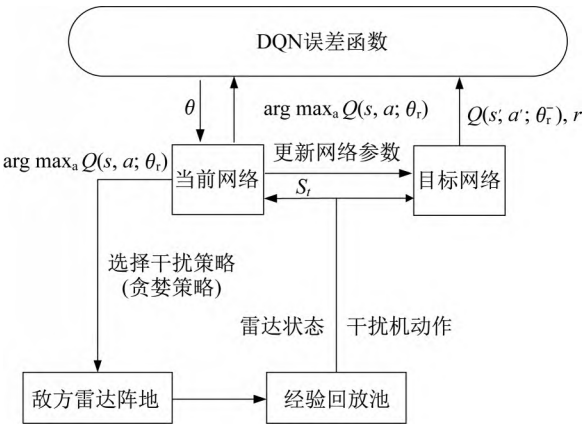

图4 本模型中DQN网络结构图

Fig. 4 DQN network structure diagram in this model

结合建立模型的DQN算法的具体实施步骤如图5所示，主要流程为：先对网络进行初始化；然后侦查获得雷达当前时刻状态，使用贪婪策略以 ε 的概率选择最优动作，以 $(1-\varepsilon)$ 的概率选择随机动作，与设定对抗环境进行交互获得下一时刻状态，并对该动作进行打分操作，将上文中的状态和动作作为DNN网络的输入，进行值函数的计算，将当前状态下采取动作的优劣度以值函数的方式进行表述；当达到训练总步数后，将状态和动作信息存储下来，用于自适应干扰机的在线干扰资源学习和分配。

Dueling-DQN算法与DQN算法的不同点在于：DQN神经网络输出的是每种动作的 Q 值，而Dueling-DQN每个动作的 Q 值是由式(12)确定的，其余部分两者完全相同。

$$Q(s,a;\theta_r)=V_s(s;\theta_r)+A(s,a;\theta_r)\tag{12}$$

式中：$V_s(s;\theta_r)$ 表示这个状态下的值；$A(s,a;\theta_r)$ 表示每个动作在这个状态上的优势。因为有时在某种状

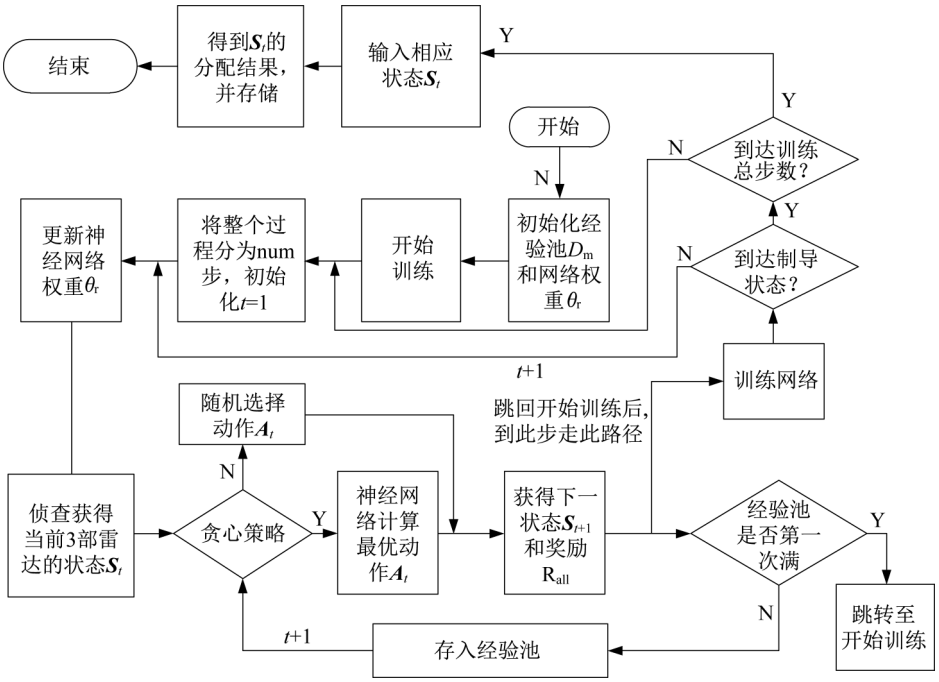

图5　结合建立模型的DQN算法流程图

Fig. 5　DQN algorithm flow chart combined with the establishment of the model

态下,无论做什么动作,对下一个状态的影响均很小。通过这种方法就能大幅提升学习效果,加速收敛。

Dueling-DQN(R)算法与DQN(R)算法分别是在Dueling-DQN算法与DQN算法的基础上,将动态调整的奖励值引入其中。

3　仿真验证

3.1　干扰对抗场景及DQN网络相关参数设置

雷达阵地与干扰机位置关系设置如图6所示,飞机携带自适应干扰吊舱对准雷达2进行飞行,速度v为300 m/s,3部雷达分别位于高度H为1 000 m的小岛上,雷达阵地间隔d为5 000 m,目标散射截面积σ为6 m^2,雷达平均功率为77 kW,雷达天线增益G_t为42 dB,脉冲宽度为6.4 us,接收机带宽为40 MHz,载频为8 GHz,雷达接收机损耗L_t为6 dB,脉冲重复频率为5 000 Hz;干扰机有效功率100～1 000 W,干扰机天线增益G_j为15 dB,干扰机带宽为400 MHz,γ_j为0.5 dB。海杂波环境中当海面风速为10～20 Kt(Kt用来描述海况信息的风速单位),即风速为19～38 km/h,X波段不同来源的海杂波σ^0的数据合成为－36 dB。3部雷达为同一体制雷达,工作参数相近,均有3种工作模式。

经验回放池大小为2 000,奖励折扣因子为0.9,学习率为0.001,ε-greedy为0.9,重置网络权重步数为1 200,每批次训练量为320。

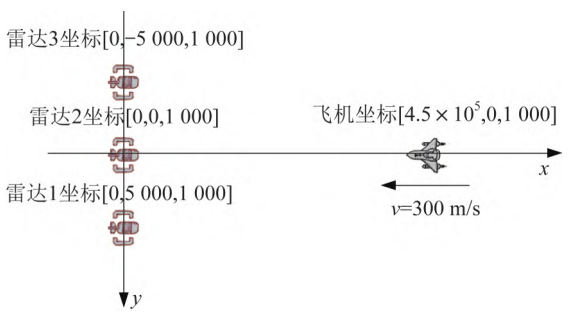

图6　雷达阵地与干扰机位置关系

Fig. 6　The relationship between the radar position and the jammer

DQN(R)算法和Dueling-DQN(R)算法训练结果对比如图7～8所示。

由图7可知:大约680次训练后Dueling-DQN(R)算法分配下的最大突防距离已经基本稳定,而DQN(R)算法则要经过约1 000次训练才能达到Dueling-DQN(R)算法的效果。由图8可知:引入干扰功率奖励值和干扰样式能量最小化奖励值后,1 200次的训练后,每次突防的整体辐射能量占比在20%～50%之间,且Dueling-DQN(R)算法下的整体辐射能量占比基本上比DQN(R)算法的整体辐射能量小,即:就干扰辐射能量的损耗情况而言,Dueling-DQN(R)算法分配的干扰策略要优于DQN(R)算法分配的干扰策略。

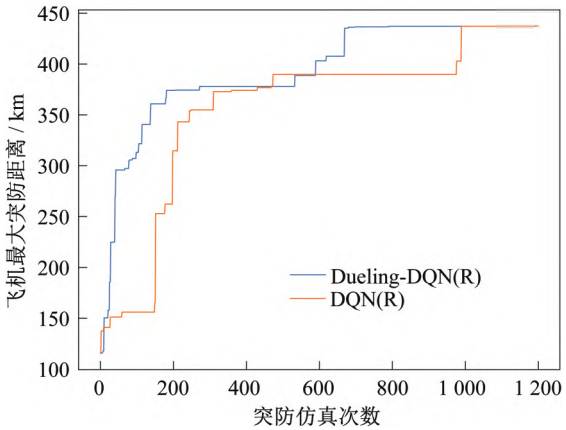

图 7 两种算法下最大突防距离

Fig. 7 Maximum penetration distance of the two algorithms

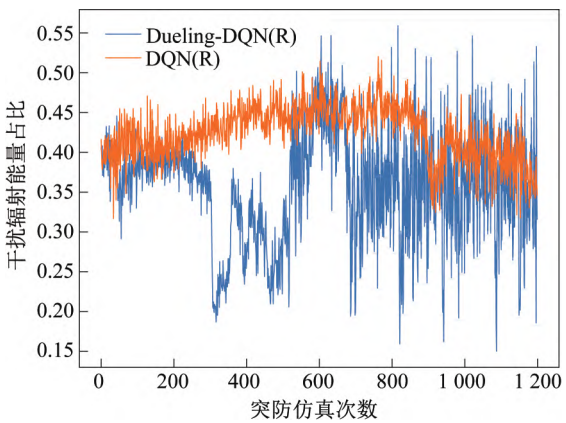

图 8 两种算法辐射能量占比

Fig. 8 The proportion of radiated energy of the two algorithms

DQN(R)算法与 Dueling-DQN(R)算法训练至 1 200 次时干扰动作分配如图 9～10 所示。

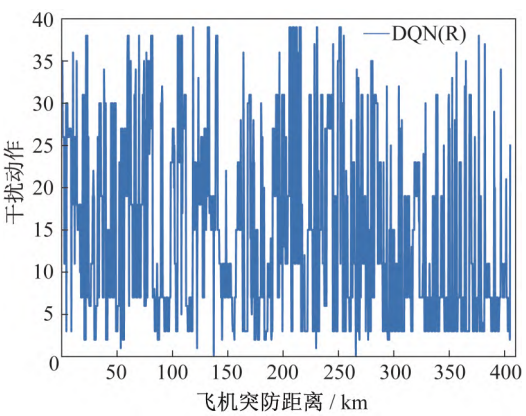

图 9 DQN(R)算法训练 1 200 次时的动作分配

Fig. 9 Action distribution when DQN(R) algorithm training reaches 1200 times

由图 9～10 中可以看出，Dueling-DQN(R)算法的干扰动作分配比较稳定，而 DQN(R)算法的分配结果

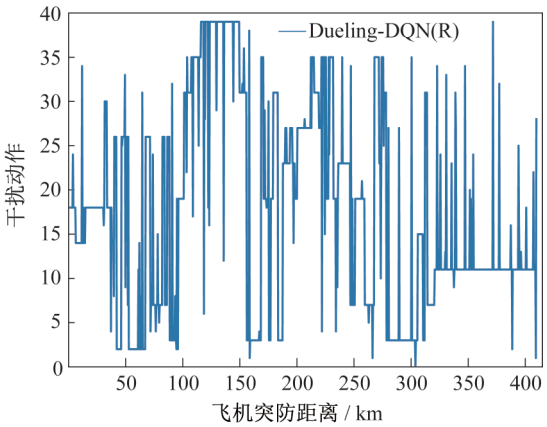

图 10 Dueling-DQN(R)算法训练 1 200 次时的动作分配

Fig. 10 Action distribution when Dueling-DQN(R) algorithm training reaches 1200 times

较为多变，结合两种算法的原理可知，这是因为 Dueling-DQN(R)算法去掉某一状态下的不敏感动作，使得其收敛能力较 DQN(R)有明显提升。并且，图 9～10 中，动作数值越大表示其干扰功率越高，大部分动作的选取所需的干扰功率未达到峰值，而较为独立的峰值大都为雷达状态改变的结果。

图 11 为 3 种算法的损失函数对比，Dueling-DQN(R)算法大致在 1 800 次网络迭代后收敛，DQN(R)算法大致在 4 000 次网络迭代后收敛，而寻常 DQN 算法大致在 9 000 次网络迭代后收敛。从其变化规律来看，3 种算法的收敛速度和稳定性从优到劣为：Dueling-DQN(R)算法、DQN(R)算法、DQN 算法，这是因为引入了随环境动态调整的奖励值 R_3 的结果，提升了前两种算法收敛的速度和稳定性。而 Dueling-DQN(R)算法、DQN(R)算法的收敛速度也从一定程度上佐证了图 7 中曲线变化梯度。

图 12 是网络迭代 14 000 次中 Dueling-DQN(R)算

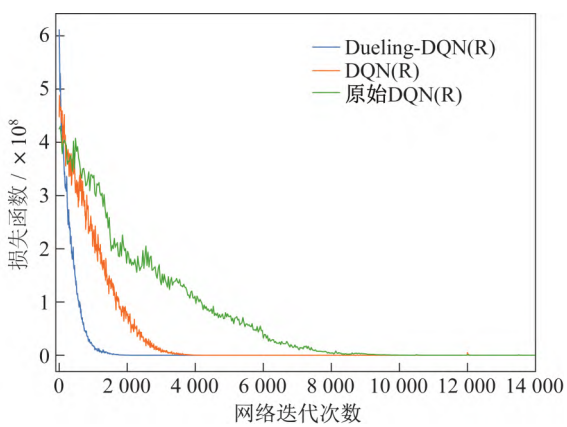

图 11 3 种算法的收敛曲线

Fig. 11 Convergence curves of the three algorithms

法下不同时刻自适应干扰机所选择干扰的目标雷达，可见在整个突防的过程，Dueling-DQN(R)算法实现了随突防距离变化更改其一对多干扰策略的能力。

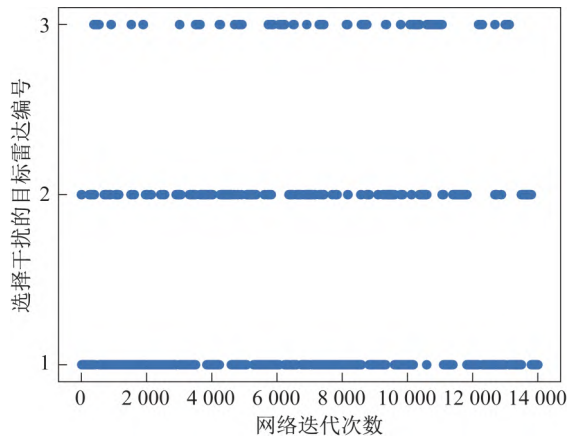

图 12　自适应干扰机选择干扰的雷达

Fig. 12　The adaptive jammer selects the radar to jam

4　结束语

为了研究一对多情形下的干扰资源的分配方法，本文提出了一种奖励值随突防距离动态调整的DQN干扰资源分配方法，以干扰、杂波下的雷达最大探测距离作为奖励值的调整标准，可以加快DQN算法和Dueling-DQN算法的收敛速度和稳定性，使得干扰资源的分配更快地收敛。本文算法较为理想地完成了对不组网雷达阵地的一对多的干扰资源分配，对干扰辐射总能量小、突防距离大的认知电子对抗以及非对称下多机协同干扰的干扰资源分配具有一定的参考价值。

参　考　文　献

[1] 赫彬，苏洪涛. 认知雷达抗干扰中的博弈论分析综述[J]. 电子与信息学报，2021，43(5)：1199-1211.

[2] 贠洁，孙闽红，官友廉. 基于雷达字建模的多功能雷达工作模式识别[J]. 杭州电子科技大学学报(自然科学版)，2020，40(6)：19-25.

[3] 郭小一，袁卫卫，黄金才. 雷达干扰资源一对多分配方法[J]. 火力与指挥控制，2008，33(12)：22-25.

[4] 黄星源，李岩屹. 基于双Q学习算法的干扰资源分配策略[J]. 系统仿真学报，2021，33(8)：1801-1808.

[5] 周彬，郭艳，李宁，等. 基于导向强化Q学习的无人机路径规划[J]. 航空学报，2021，42(9)：506-513.

[6] 刘松涛，雷震烁，温镇铭，等. 认知电子战研究进展[J]. 探测与控制学报，2020，42(5)：1-15.

[7] 黄亭飞，程光权，黄魁华，等. 基于DQN的多类型拦截装备复合式反无人机任务分配方法[J]. 控制与决策，2022，37(1)：142-150.

[8] 罗鹏，黄珍，秦易晋，等. 基于DQN的车辆驾驶行为决策方法[J]. 交通信息与安全，2020，38(5)：67-77.

[9] 陈涛，张颖，黄湘松. 基于强化学习的自适应干扰波形设计[J]. 空天防御，2021，4(2)：59-66.

[10] 陈岩，李艳艳，杨立波，等. 地海杂波统计特性研究概述[J]. 空天防御，2020，3(4)：44-51.

[11] 郭予并，魏永峰. 干扰条件下机载雷达对海探测距离数值计算分析[J]. 雷达与对抗，2014，34(4)：5-8.

[12] 尹良，刘红杰，赵晓峰，等. 基于雷达方程的AN/TPY-2雷达作用距离与干扰研究[J]. 系统工程与电子技术，2018，40(1)：50-57.

[13] 陈淦涛，许稼，高效，等. 有源压制干扰下雷达探测距离分析与计算[J]. 雷达科学与技术，2011，9(1)：13-17.

[14] 贠洁，孙闽红，官友廉. 基于雷达字建模的多功能雷达工作模式识别[J]. 杭州电子科技大学学报(自然科学版)，2020，40(6)：19-25.

[15] 刘蓉，张衡，肖颖峰. 基于改进马尔可夫决策过程模型的多机协同航路规划研究[J]. 南京理工大学学报，2021，45(1)：84-91.