基于强化学习的集群多目标分配与智能决策方法

朱建文[1]，赵长见[2]，李小平[1]，包为民[1,3]
(1. 西安电子科技大学 空间科学与技术学院，陕西 西安 710126；2. 中国运载火箭技术研究院，北京 100076；
3. 中国航天科技集团有限公司，北京 100048)

摘要： 为提升高动态协同攻击条件下的攻防效能，研究基于强化学习的集群多目标智能分配与决策方法。建立综合攻击性能评估准则，包括基于相对运动信息的攻击优势度评估以及基于目标固有信息的威胁度评估。综合攻击性能、突防概率以及攻击消耗，设计攻防效费比性能指标。构建基于强化学习的多目标决策架构，设计以分配向量为基本元素的动作空间，以及基于量化性能指标的状态空间，利用 Q-Learning 方法对协同攻击方案，包括导弹选取以及分配形式进行智能决策。仿真结果表明，强化学习能够实现攻防效能最优的多目标在线决策，其计算效率相对于粒子群优化算法具有更明显的优势。

关键词： 目标分配；协同攻击；攻防效能；智能决策；强化学习
中图分类号： TJ761.1+4 **文献标志码：** A **文章编号：** 1000-1093(2021)09-2040-09
DOI： 10.3969/j.issn.1000-1093.2021.09.025

Multi-target Assignment and Intelligent Decision Based on Reinforcement Learning

ZHU Jianwen[1], ZHAO Changjian[2], LI Xiaoping[1], BAO Weimin[1,3]
(1. School of Aerospace Science and Technology, Xidian University, Xi'an 710126, Shaanxi, China;
2. China Academy of Launch Vehicle Technology, Beijing 100076, China;
3. China Aerospace Science and Technology Corporation, Beijing 100048, China)

Abstract: A reinforcement learning-based swarm intelligent decision-making method of cooperative multi-target attack under high-dynamic situation is proposed. The composite evaluation criteria of attack performance is established, including the evaluation of attack superiority based on relative motion information and the threat evaluation based on the inherent information of target. To evaluate the attack-defence effectiveness, a cost-effectiveness ratio index is designed by combining attack performance, penetration probability and attack cost together. In addition, a multi-target decision-making architecture based on reinforcement learning is constructed, and an action space with allocation vectors as basic elements and a state space based on quantified performance indicators are designed. Q-Learning is employed to make intelligent decisions on cooperative attack plans, including missile selection and target assignment. The simulated results show that reinforcement learning can achieve multi-target online decision-making with the optimal offensive and defensive effectiveness, and its computational efficiency has more obvious advantages than that of particle swarm optimizer.

收稿日期： 2020-10-13
基金项目： 国家自然科学基金项目(61703409)；中国博士后科学基金项目(2019M66364)
作者简介： 朱建文(1987—)，男，讲师，博士。E-mail: zhujianwen1117@163.com
通信作者： 李小平(1961—)，女，教授，博士生导师。E-mail: xpli@xidian.edu.cn

Keywords: target assignment; cooperative attack; attack-defense effectiveness; intelligent decision; reinforcement learning

0　引言

随着导弹信息化与体系化能力的提升,其攻击模式由单一攻防作战拓展到多对多的群体协同对抗与博弈。多弹协同攻击能够充分利用分散的作战资源以及信息共享,是提升打击能力与突防能力的有效途径。针对多目标的分配与决策直接决定着体系的攻防性能,是协同攻击的关键技术之一[1]。

多目标决策与分配需要根据实时的攻防态势,对集群中的每个成员参与攻击与否进行决断,并分配合理的待攻击目标。攻防性能评估是目标分配的基础条件,可利用弹目相对运动信息来评估制导的难易程度以及攻击性能,而目标的威胁度可基于自身价值与运动特性来评估[2-3]。集群决策与分配是一个以攻防性能评估结果为模型、以攻防性能最大为性能指标的寻优过程[3]。倾向性和主观性是集群攻防性评估不可避免的因素,为此刘树衍等[4]综合利用专家系统与神经网络构建行为决策基础模型,进而建立智能指挥系统以优化目标分配。另一种典型方法是将分配问题转换为数学规划问题,进而利用枚举法、分支界定法或整数规划来求解[5-6]。然而,随着攻防双方规模的增加,寻优的复杂度会急剧增大,导致计算耗时呈指数型增长[7]。因此,具有灵活性、自适应能力强以及计算相对简单的智能优化方法,在求解复杂多目标决策与分配中具有较大的优势,遗传算法与粒子群优化(PSO)算法为其典型代表[8]。PSO 算法利用种群中个体运动位置和整体最优位置的记忆与学习,在解空间中朝着最优的方向运动,该算法相对于遗传算法具有更高的计算效率,但其精细程度与全局搜索能力不足[9-10]。

高动态的集群攻防为决策的最优性与实效性提出了极高的需求,其复杂多变的攻防态势需要进行多次在线决策与分配。上述优化方法在计算效率、全局最优性以及多次决策的继承性上存在不足。集群决策与目标分配中能够影响攻防性能的分配矩阵是离散的,而且多目标决策与分配满足马尔可夫决策过程[9]。本文利用强化学习对集群攻击的导弹选取以及目标分配矩阵进行决策判断,具体包含攻防性能评估、非线性攻防效费比指标构建、强化学习框架的搭建、离散化动作空间、状态空间以及奖励函数的设计。

1　综合攻击性能评估

以多发导弹对地球表面运动的目标群进行协同攻击为背景,对其攻击性能进行评估。多对多的攻防态势包括导弹自身的攻击优势度以及目标的威胁度[3]。在攻击优势度中,主要考虑弹目相对角度、距离以及速度的优势模型;目标的威胁度可基于固有特性与运动信息来评估。

1.1　基于相对运动信息的攻击优势度评估

1.1.1　攻击角度优势度评估

由于导弹在攻击目标时需要满足速度倾角约束并消除航向误差,攻击角度优势度评估需要综合考虑速度倾角与方位角。在纵向通道,当实时速度倾角与终端约束相等时,制导越容易实现,意味着攻击优势度随角度差的减小而增大。在侧向通道,导弹制导的主要目标为消除航向误差 $\Delta\sigma$,因此该误差的绝对值越大,制导任务越艰巨。相反地,若 $\Delta\sigma=0$,则导弹对该目标的优势最大。因此,可构造角度优势模型为

$$\begin{cases} S_{M\theta}=e^{-|\theta-\theta_f|}, \\ S_{M\sigma}=e^{-|\sigma-\sigma_{LOS}|}, \end{cases} \tag{1}$$

式中:θ 为速度倾角;θ_f 为终端速度倾角约束;σ 为速度方位角;σ_{LOS} 为视线方位角;$S_{M\theta}$ 与 $S_{M\sigma}$ 分别为基于速度倾角 θ 与方位角 σ 的攻击优势度。

1.1.2　相对距离优势度评估

导弹与目标之间的距离必然影响制导指令的生成与打击目标的实现,当距离过近时导弹的反应时间太短,为制导指令的执行带来了巨大压力。相反,当距离太远时导弹的探测精度受到不良影响,并且过大的能量损耗也将影响打击任务的完成。因此,相对距离的优势模型可构造为

$$S_{Mr}=e^{-\frac{|r-R_0|}{R_0}}, \tag{2}$$

式中:S_{Mr} 为基于弹目距离的攻击优势度;r 为弹目距离;R_0 为综合考虑探测能力与机动能力而确定的距离。(2)式中基于距离优势度评估的物理意义为:当导弹与目标的距离为 R_0 时优势最强;弹目距离与 R_0 相差越大,则优势越弱。

1.1.3　攻击过载优势度评估

由于导弹的机动与控制能力直接体现在可用过载上,并且过载能够同时包含弹目相对角度、距离以

及速度大小。因此,本文进一步引入过载为变量,以表征导弹对不同目标的优势度。具体方法如下:基于导弹当前的飞行状态与目标信息,采用最优制导方法计算导弹在侧向的需要过载指令。过载指令越大,意味着待飞时间越短、打击任务更加艰巨,过大的过载指令将超过导弹的控制能力,导致打击任务失败。越小的过载指令意味着越小的控制能力需求以及更加平直的弹道,但是平直的弹道将降低突防性能。因此,基于过载的优势模型为

$$S_{\mathrm{M}n}=\mathrm{e}^{-\left(\frac{|n|-n_0|}{n_0}\right)^2}, \tag{3}$$

式中:$S_{\mathrm{M}n}$为基于过载的攻击优势度;n 为过载;n_0为基于控制能力确定的过载基准量,$n_0>0$ g.

1.2 基于目标固有信息的威胁度评估

目标群中不同目标具有不同的战略价值与威胁程度,对于重要目标应当分配更多的导弹进行打击,以增强打击效果。本文考虑了易于获取的目标体积信息与速度信息作为威胁度评估的标准,体积代表弹载量与威胁度,速度表示目标的动力与机动性能,进一步将二者加权平均以综合评估目标威胁度,用于后续的目标分配。

1.2.1 目标体积威胁度评估

不同体积的目标具有不同的作战性能以及威胁程度,目标体积越大,则受威胁程度越大。因此,基于体积信息的目标威胁模型可构建为

$$S_{\Gamma_{\mathrm{t}}}=\frac{\Gamma_{\mathrm{t}j}}{\sum_{j=1}^{N_{\mathrm{t}}}\Gamma_{\mathrm{t}j}}, \tag{4}$$

式中:$S_{\Gamma_{\mathrm{t}}}$为基于体积的目标威胁度;$\Gamma_{\mathrm{t}j}$为第 j 个目标的体积大小;N_{t}为目标的数量。目标体积威胁模型(4)式的物理意义为:获取所有目标的体积,则第 j 个目标的威胁度可用其在整个目标群中的体积占比来表述。

1.2.2 目标速度威胁度评估

目标的航行速度对其威胁程度存在较大影响。目标的机动性能随速度的增大而增大,但由于目标动力性能的限制,过大的速度意味着目标在体积与质量上存在不足。因此基于速度信息的威胁模型为

$$S_{v_{\mathrm{t}}}=\mathrm{e}^{-\left|\frac{v_{\mathrm{t}}-v_{\mathrm{t}0}}{v_{\mathrm{t}0}}\right|}, \tag{5}$$

式中:$S_{v_{\mathrm{t}}}$为基于速度的目标威胁度;v_{t} 为目标的实际航行速度;$v_{\mathrm{t}0}$为预先设定的速度。目标速度威胁模型(5)式的物理意义为:当目标速度为 $v_{\mathrm{t}0}$时,越具有威胁性,过大或过小的速度都将降低威胁度。

1.3 综合攻击优势度评估

基于攻击优势模型与目标威胁模型,可建立用于目标分配的综合攻击优势度模型如下:

$$S=S_{\mathrm{a}}+S_{\mathrm{t}}, \tag{6}$$

式中:S_{a}为攻击优势度模型,

$$\begin{cases}S_{\mathrm{a}}=k_\theta S_{\mathrm{M}\theta}+k_\sigma S_{\mathrm{M}\sigma}+k_r S_{\mathrm{M}r}+k_n S_{\mathrm{M}n},\\ k_\theta+k_\sigma+k_r+k_n=1,\end{cases} \tag{7}$$

k_θ、k_σ、k_r、k_n 为加权系数,不同参数设置对应不同的重要程度;S_{t}为基于目标体积与速度的威胁度模型,

$$\begin{cases}S_{\mathrm{t}}=k_\Gamma S_{\Gamma_{\mathrm{t}}}+k_v S_{v_{\mathrm{t}}},\\ k_\Gamma+k_v=1,\end{cases} \tag{8}$$

k_Γ、k_v分别为体积与速度的加权系数。针对上述模型,需要给出以下 3 点说明:

1) 不同加权系数意味着不同的关注度,可根据具体攻击任务进行设计;

2) 针对不同目标需要考虑的因素存在差异,该模型主要针对地球表面航行的大型目标群;

3) 除上述威胁模型外,还可根据需要考虑目标电磁辐射情况、预设目标的重要程度以及其他能够反映目标特性的重要因素。

2 攻防一体性能指标构建

多目标分配与决策需要以综合攻击优势度 S 为基础,通过优化方法获得分配矩阵 X,实现攻击性能的最大化。首先,只考虑导弹运动信息与目标固有信息建立如下线性攻击性能指标:

$$\max J_{1,\mathrm{a}}=\sum_{i=1}^{N_{\mathrm{M}}}\sum_{j=1}^{N_{\mathrm{T}}}S_{ij}X_{ij}, \tag{9}$$

式中:$J_{1,\mathrm{a}}$为攻击性能指标;N_{M}与 N_{T}为导弹与目标的数量;S_{ij}为导弹 i 对目标 j 的量化综合攻击优势度;X_{ij}为导弹群对目标群分配矩阵中的元素。评估模型(6)式与性能指标(9)式构成了典型的整数规划问题,可利用内点法等方法进行寻优求解[9]。

进一步考虑导弹的突防概率,建立目标的毁伤性能指标:

$$\max J_{\mathrm{o,d}}=\sum_{j=1}^{N_{\mathrm{T}}}S_{\mathrm{t}j}\left(1-\prod_{i=1}^{N_{\mathrm{M}}}X_{ij}(1-P_{ij})\right)_{P_{ij}\neq 0}, \tag{10}$$

式中:$J_{\mathrm{o,d}}$为毁伤性能指标;$S_{\mathrm{t}j}$为第 j 个目标的价值;P_{ij}为导弹 i 对目标 j 的突防概率(0 ~ 1 之间取值)。另外,导弹攻击必然造成导弹的消耗,因此导弹协同攻击的成本指标为

$$\min J_{\mathrm{c}}=\sum_{i=1}^{N_{\mathrm{M}}}\left(c_i\sum_{j=1}^{N_{\mathrm{T}}}X_{ij}\right), \tag{11}$$

式中:J_{c} 为导弹消耗指标;c_i为导弹 i 的成本。综合

考虑 $J_{1,a}$、$J_{o,d}$以及 J_c，则可得协同攻击的综合效费性能指标为

$$\max J_t = [J_{1,a}, J_{o,d}, J_c]. \tag{12}$$

指标(12)式的目的是获得最大的效费比，但其中包含两个相互矛盾的性能指标：$J_{1,a}$与 $J_{o,d}$的目标是获得最大的攻击与毁伤性能，J_c 的目标是获得最小的攻击成本。因此，进一步引入效费比来描述单一导弹的效能，将(12)式中的两个性能指标进行整合，进而利用整合之后的单一性能指标进行优化设计。其中：

攻击效费比指标 J_a为

$$\max J_a = \frac{J_{1,a}}{J_c} = \frac{\sum_{i=1}^{N_M}\sum_{j=1}^{N_T} S_{ij}X_{ij}}{\sum_{i=1}^{N_M}\left(c_i\sum_{j=1}^{N_T}X_{ij}\right)}, \tag{13}$$

毁伤效费比指标 J_d为

$$\max J_d = \frac{J_{o,d}}{J_c} = \frac{\sum_{j=1}^{N_T} S_{tj}\left(1-\prod_{i=1}^{N_M}P_{ij}\right)_{P_{ij}\neq 0}}{\sum_{i=1}^{N_M}\left(c_i\sum_{j=1}^{N_T}X_{ij}\right)}, \tag{14}$$

攻防效费比指标 J_t 为

$$\max J_t = \frac{J_{1,a}J_{o,d}}{J_c} = \frac{\sum_{i=1}^{N_M}\sum_{j=1}^{N_T} S_{ij}X_{ij} + \sum_{j=1}^{N_T} S_{tj}\left(1-\prod_{i=1}^{N_M}X_{ij}(1-P_{ij})\right)_{P_{ij}\neq 0}}{\sum_{i=1}^{N_M}\left(c_i\sum_{j=1}^{N_T}X_{ij}\right)}. \tag{15}$$

性能指标(15)式的物理意义为：基于矩阵形式的综合攻击优势度 S、突防概率 P_{ij}以及导弹的成本 c_i，确定相同维度的分配矩阵 X，使得性能指标(15)式即攻防效费比最大。在协同攻击的多目标分配与决策过程中，必须满足的约束模型为

$$\text{s.t.}\begin{cases}\sum_{j=1}^{N_T}X_{ij}=1,\ i=1,\cdots,N_M,\\ 1\leqslant\sum_{i=1}^{N_M}X_{ij}\leqslant T_j,\ j=1,\cdots,N_T,\\ X_{ij}=\{0,1\}.\end{cases} \tag{16}$$

约束模型(16)式的物理意义为：目标分配结果以分配矩阵的形式表征，被攻击的目标标记为 1，否则标记为 0，即目标分配矩阵 X 的元素只能够是{0，1}中的某一值。由于每一发导弹最多只能攻击一个目标，矩阵中的每一行元素数值之和必为 1. 另外，需要保证每一个目标至少分配 1 发导弹进行攻击，并且目标分配矩阵中每一列元素之和不小于 1，且分配至某一目标的导弹数量最多为 T_j.

3　基于强化学习的多目标分配

性能指标(15)式是严格的非线性方程，本文利用强化学习方法实现多目标的智能分配。强化学习又称再励学习、评价学习或增强学习，该方法需要智能体与环境进行反复信息交互，通过学习策略或规则实现回报或指标的最优化[11]。

3.1　强化学习与 Q-Learning 逻辑

强化学习是一种试探、评价与更新的过程，智能体选择一个动作作用于环境，环境在执行完动作之后产生回报(奖励)信号发送至智能体，该信号包含对动作的定量评价；不同的动作对应不同的奖励值，智能体在接收回报信号之后，选择下一动作以获得更大的奖励[12]。

强化学习是迭代优化的过程，包含值迭代与策略迭代。Q-Learning 是强化学习最常用的值函数迭代更新方法，设 $Q(s,a)$ 为状态行为值函数，其物理意义为在当前策略 π 下，当前状态 s 与动作 a 对应值函数的具体取值[13]。若状态集合为 p 维、动作集合为 q 维，则 $Q(s,a)$ 为 $p\times q$ 维表格，因此可称之为 Q 表。Q-Learning 中值函数的更新方法[14]为

$$Q(s,a)\leftarrow Q(s,a)+\alpha\left[R+\gamma\max_a Q(s',a)-Q(s,a)\right], \tag{17}$$

式中：α 为值函数迭代的校正系数；γ 为折扣系数；R 与 s'分别为执行当前动作获得的回报值与下一时刻的状态。

具体的 Q-Learning 方法步骤[15]如下：

步骤 1　人为初始化 $Q(s,a)$ 表格。

步骤 2　对于每次学习训练，给定一个初始状态 s.

步骤 3　执行以下操作：

①利用当前的 Q 值，依据策略 π，确定当前的行为 a；

②执行当前的行为 a，获得量化的回报 R 与下一状态 s'；

③根据(17)式更新 Q 表；

④更新当前的状态 $s\leftarrow s'$；

⑤当状态满足终止状态时，结束当前回合的学习。

步骤 4　基于已更新的 Q 表，重复执行步骤 3，直至满足学习次数。

3.2 基于 Q-Learning 的多目标分配

在多目标分配与决策中，不同形式的 0-1 分配矩阵对应不同的攻防效费比。由于攻防性能只与当前和未来分配矩阵相关，而与过去的信息无关，因此集群决策与分配矩阵的确定符合马尔可夫决策过程。根据强化学习与 Q-Learning 方法的需求，需要根据实际优化任务对搭建智能分配模型，设计状态与动作空间以及回报函数，并利用典型的 ε-greedy 学习策略以探索更多的动作[16]。基于 Q-Learning 算法的多目标智能分配流程如图 1 所示。

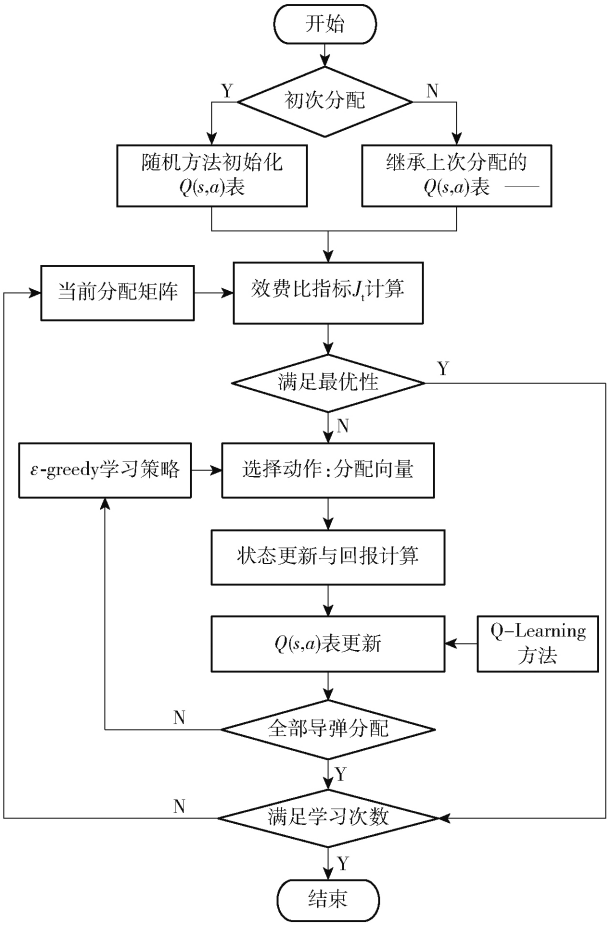

图 1 Q-Learning 智能决策迭代计算流程

Fig. 1 Iterative calculation of intelligent decision by Q-Learning method

图 1 给出了多目标智能分配的流程，其核心步骤为行为策略、动作空间、状态空间以及奖励函数的设计。

3.2.1 行为策略设计

采用 ε-greedy 策略实现多目标分配。为了充分发挥强化学习的探索和寻优能力，利用随机方法对 Q 表进行初始化，在学习前期 ε 可选择较大，以探索更多的状态与动作；在学习后期 ε 逐渐减小，以使得目标分配在已有经验基础上做出正确的动作。

3.2.2 动作空间设计

根据强化学习中对动作空间的定义，动作需要对上述状态产生影响。过于复杂的动作空间将增大动作的搜索空间，进而影响学习效率。针对该问题，设计动作为能够直接影响飞攻防性能的目标分配情况，本文称为分配向量。分配向量中，某一个具体动作 a_i 表示导弹选择目标 i，即行向量表示的动作 a_i 中，第 i 个元素为 1，其余都为 0. 若存在 N_T 个目标，则存在 N_T 个具体动作，意味着动作空间为 N_T 维。

(18) 式给出了 N_T 维的动作空间，选择第 1 个目标的动作 1 为 $a_1 = [1 \quad 0 \quad \cdots \quad 0]$，相应地选择第 2 个目标的动作 2 为 $a_2 = [0 \quad 1 \quad \cdots \quad 0]$，以此类推。

$$\boldsymbol{A} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{N_T} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}}_{N_T}. \tag{18}$$

3.2.3 状态空间设计

状态空间是强化学习中必不可少的部分，是反映当前状态或者终端状态的数据集合，并且必须包含所有可能的状态参数取值。本文设计状态空间为量化攻防效费比评估值组成的数据集合，基于性能指标(15)式构建攻防效费比函数为

$$f(X_{ij}(k)) = \frac{\sum_{i=1}^{N_M}\sum_{j=1}^{N_T} S_{ij}X_{ij}(k) + \sum_{j=1}^{N_T} S_{ij}\left(1 - \prod_{i=1}^{N_M} X_{ij}(1 - P_{ij})\right)_{P_{ij}\neq 0}}{\sum_{i=1}^{N_M}\left(c_i \sum_{j=1}^{N_T} X_{ij}(k)\right)}, \tag{19}$$

式中：$X_{ij}(k)$ 表示第 k 个分配矩阵的元素。针对 N_M 发导弹攻击 N_T 个目标的工况，忽略(16)式中的约束条件，则存在 $N_T^{N_M}$ 个分配矩阵。设置状态空间的下边界为 $N_T^{N_M}$ 个分配矩阵中攻防效费比函数的最小值 $f_{\min}$，上边界为性能函数的最大值 $f_{\max}$. 则状态空间的范围为

$$S = [f(X_{ij}(k))_{\min}, f(X_{ij}(k))_{\max}], \quad k = 1, 2, \cdots, N_T^{N_M}, \tag{20}$$

进一步将状态范围(20)式离散为等间隔的状态空间，进而获得目标分配的状态空间。

3.2.4 回报函数设计

量化的回报函数用来判断动作的性能，是强化学习的核心。在目标分配中，利用强化学习方法确定分配矩阵以获得最优的攻防性能。因此根据分配

需求，设计回报函数如下：

$$R=\begin{cases}-5, \sum_{j=1}^{N_T} X_{ij}(k) \neq 1, i=1,\cdots,N_M;\\ -5, \sum_{i=1}^{N_M} X_{ij}(k) < 1, j=1,\cdots,N_T;\\ -\|f(X_{ij}(k)) - 1.2f(X_{ij}(k))_{\max}\|, \text{其他}。\end{cases} \tag{21}$$

(21) 式中回报函数的物理意义是：当某一动作即目标分配矩阵满足所有攻击约束时，回报函数值为实际攻防量化值与最大值 1.2 倍的差。当不满足攻击约束即某一导弹分配了多个目标，或者某一目标未分配到导弹时，给予 −5 的回报值。

4　多目标决策仿真验证

采用数值仿真的方法对多目标智能分配与决策进行验证。在攻击优势度评估中，设置距离优势模型中的 $R_0=100$ km，过载优势模型中的 $n_0=1$ g，各项的加权系数分别为：$k_\theta=0.2$，$k_\sigma=0.2$，$k_r=0.2$，$k_n=0.4$。在目标威胁建模中，设置(5)式中的 $v_{t0}=20$ m/s，3 个目标的速度分别为 $v_{tA}=25$ m/s、$v_{tB}=22$ m/s 和 $v_{tC}=20$ m/s，归一化后的体积分别为 $\Gamma_{tA}=1$、$\Gamma_{tA}=1.2$ 和 $\Gamma_{tA}=1.5$，加权系数为 $k_\Gamma=0.6$、$k_v=0.4$。各发导弹属于同一类型，即 $c=1$。

在强化学习中 $\gamma=0.2$，采用 ε-greedy 策略实现决策目标，学习次数 $N_{\text{Q-Learning}}$ 的范围为 1 ~ 1 000，时变参数 $\varepsilon=\exp(-N_{\text{Q-Learning}}/100)$。参数 ε 设置的目的是：在学习前期更大地探索新的动作，在后期则保证学习的最优性。

4.1　导弹数量固定的智能分配

设置 6 发导弹攻击 3 个目标，各导弹对目标的量化综合攻击优势度以及突防概率如表 1 所示。从表 1 中可见，第 1 发导弹 M1 对目标 B 最具有优势，对目标 C 最无优势。

表 1　各导弹对目标的量化综合攻击优势度与突防概率

Tab. 1　Attack dominance and penetration probability

导弹	目标 A		目标 B		目标 C	
	优势度	突防概率	优势度	突防概率	优势度	突防概率
M1	1.60	0.79	1.74	0.78	1.24	0.76
M2	1.34	0.71	1.62	0.73	1.80	0.79
M3	1.40	0.68	1.54	0.75	1.76	0.70
M4	1.38	0.76	1.70	0.73	1.40	0.71
M5	1.60	0.62	1.24	0.63	1.54	0.66
M6	1.64	0.73	1.32	0.69	1.76	0.72

选择表 1 中前 4 发导弹 M1、M2、M3、M4 攻击 3 个目标，利用本文研究的强化学习方法实现目标分配，目标分配矩阵为

$$\boldsymbol{X}=\begin{bmatrix}0&1&0\\0&0&1\\0&0&1\\1&0&0\end{bmatrix}. \tag{22}$$

由(22)式可知，慢速航行的大目标 C 具有较大的威胁度，因此分配矩阵中 $X_{12}=1$，$X_{23}=1$，$X_{33}=1$，$X_{41}=1$，即导弹 M2 与 M3 都用于攻击目标 C，以增强整体攻防性能。随着导弹的飞行，每间隔 1 s，共进行 10 次目标分配，以充分验证智能方法的有效性，其中第 1 次与第 2 次分配的 Q-Learning 主要结果如图 2、图 3 所示。由仿真结果可知，由于第 1 次学习采用随机方法对动作以及 Q 表进行初始化，因此迭代次数较多，在大约 600 次学习之后才得以收敛，综合效费比指标 J_t 为 1.735 6。第 2 次学习继承了上一次学习获得的 Q 表，该表已经包含了优良动的动作信息与回报值，因此迭代次数与收敛速率都有大幅度改进。在经过上百次学习迭代后，Q-Learning 能够精确收敛。

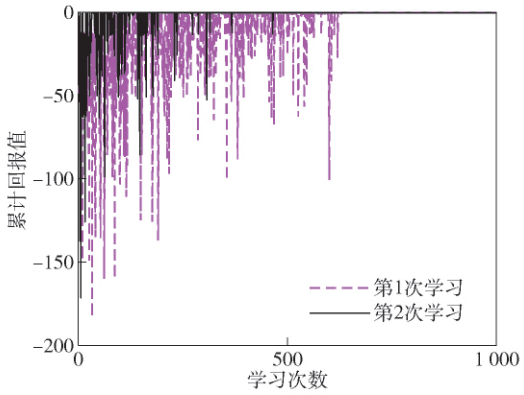

图 2　前两次分配的累计回报值

Fig. 2　Cumulative reward values of the first two assignments

在导弹飞行过程中，每间隔 1 s，分别采用强化学习与 PSO 算法实现多目标分配，两种方法的耗时与指标结果如表 2 所示(i7 8550 处理器，1.99 GHz，MATLAB 2016b 仿真环境)。由表 2 可知，强化学习与 PSO 算法都可实现多目标的自主分配，最终的综合效费比指标完全相同。然而，两种方法在计算耗时上存在一定差异，初次分配时强化学习方法耗时较长，而后续分配 PSO 算法耗时较长。对于初次分

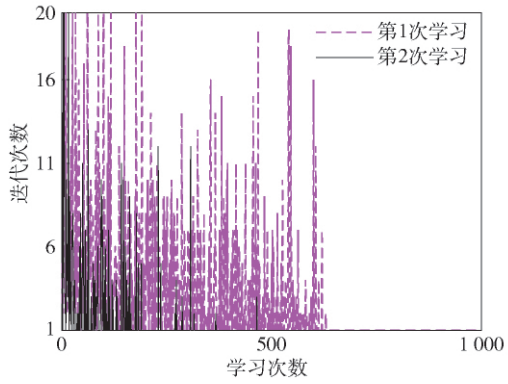

图 3 前两次分配的迭代次数

Fig. 3 Iteration steps of the first two assignments

配,强化学习方法采用随机方法进行初始化并探索更多的动作,因此耗时较长。在后续分配过程中,强化学习能够继承初次分配的结果,而 PSO 算法都需要由相同的初始状态出发进行寻优,因此强化学习耗时更短,效率更高。

表 2 强化学习方法与 PSO 算法性能对比

Tab. 2 Performance comparison between RL and PSO algorithms

分配次数	强化学习方法耗时/s	PSO 算法耗时/s	J_t
1	0.142 125	0.110 962	1.735 6
2	0.064 036	0.084 390	1.744 2
3	0.060 135	0.080 966	1.761 0
4	0.073 547	0.082 405	1.783 2
5	0.068 480	0.088 974	1.798 7
6	0.056 989	0.092 248	1.801 2
7	0.073 773	0.085 358	1.810 3
8	0.072 969	0.083 222	1.811 4
9	0.073 822	0.091 316	1.820 3
10	0.076 102	0.088 383	1.822 4

4.2 导弹数量可变的智能决策

协同攻击策略包括己方导弹的数量、成员的选择以及目标的分配情况。因此,对于给定的目标群以及价值情况,选择不同导弹数量以及编队成员,计算不同工况下的量化效费比,对攻击策略进行决策优化。利用表 1 中的 6 发导弹攻击 3 个目标,在满足(16)式所示约束条件下,根据攻击导弹的数量,可有 4 种攻击方案,每种攻击方案包含不同的导弹分组情况。例如,在 6 发导弹中选择 3 发导弹攻击 3 个目标,则存在 $C_6^3=20$ 种组合情况。4 种攻击方案与导弹分组数如表 3 所示,一共包含 42 种分组情况,即 42 种具体的攻击方案。

表 3 协同攻击方案与分组

Tab. 3 Cooperative attack program and grouping

攻击方案	导弹数量/发	导弹分组数	分组编号
1	3	$C_6^3=20$	1 ~ 20
2	4	$C_6^4=15$	21 ~ 35
3	5	$C_6^5=6$	36 ~ 41
4	6	$C_6^6=1$	42

表 3 中 42 种攻击分组情况下的攻防性能指标与效费比指标如图 4 ~ 图 9 所示。由图 4 可知,当不考虑攻击成本时,攻击导弹越多,则攻击与毁伤性能越强。当考虑攻击成本时效费比性能存在较大差异:图 5 中攻击效费比 J_a 在第 38 号编组时达到最大,此时分配 5 发导弹 M1、M2、M3、M4、M6 攻击 3 个目标;图 7 中毁伤效费比 J_d 总体上随着数量的增多而减小;图 9 中,综合考虑攻击与毁伤性能的攻防效费比 J_t 在第 23 号编组时达到最大,此时需要分配导弹 M1、M2、M3、M6 攻击目标,相应的目标分配矩阵为

$$\boldsymbol{X}=\begin{bmatrix}0&1&0\\0&0&1\\0&0&1\\0&0&0\\0&0&0\\1&0&0\end{bmatrix}. \tag{23}$$

(23)式中 $X_{12}=1$,$X_{23}=1$,$X_{33}=1$,$X_{61}=1$,其余元素均为 0,对应的物理意义是:导弹 M1 攻击目标 B,M2 与 M3 都用于攻击目标 C,M6 攻击目标 A,量化攻防效费比指标为 1.756.

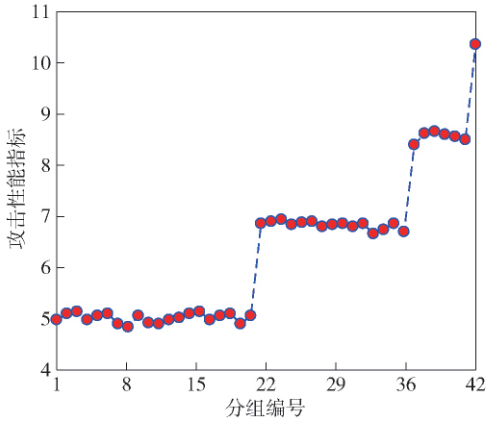

图 4 攻击性能指标 $J_{1,a}$

Fig. 4 Attack performance index $J_{1,a}$

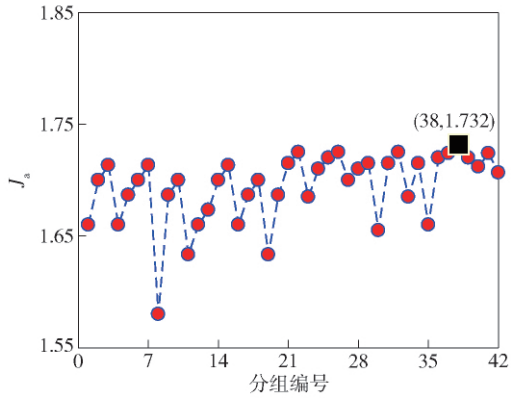

图 5　攻击效费比指标 J_a

Fig. 5　Attack cost-effectiveness ratio index J_a

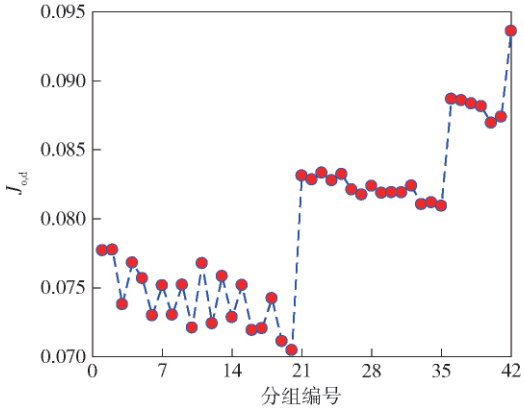

图 6　毁伤性能指标 $J_{o,d}$

Fig. 6　Damage performance index $J_{o,d}$

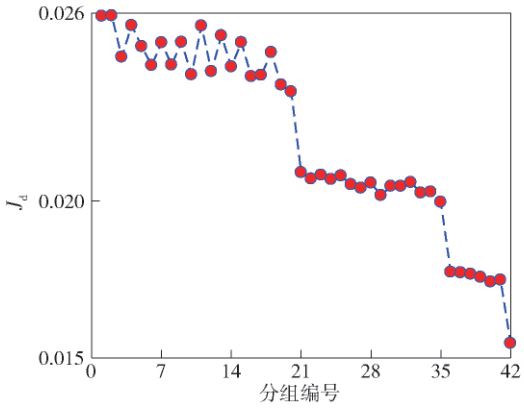

图 7　毁伤效费比指标 J_d

Fig. 7　Damage cost-effectiveness ratio index J_d

5　结论

本文采用强化学习方法研究了复杂多变且高动态环境下多目标协同攻击智能决策方法，建立了攻防性能评估准则，包括基于相对运动信息的攻击优势度评估以及基于目标固有信息的威胁度评估。综合攻击性能、毁伤性能以及攻击消耗，设计了攻防效费比性能指标。构建了基于强化学习的多目标决策架构，设计了目标分配的动作空间与状态空间，利用 Q-Learning 方法对协同攻击方案，包括导弹的数量、分组选取以及目标分配进行了智能决策。得出以下主要结论:

1) 基于相对运动信息与目标固有信息，可实现对攻击优势度与目标威胁度的评估，结合突防概率模型，可构建攻防效费比指标模型。

2) 多目标协同攻击的目标是使得攻防性能最优化，攻击导弹的选取以及目标分配的决策结果与性能指标以及决策模型密切相关。

3) 强化学习能够用于协同攻击中多目标的在线决策与分配，与 PSO 算法相比，其计算效率在非初次决策中具有更明显的优势。

本文研究的是一种基于强化学习的基础性、通用性的目标分配与智能决策方法。只需要建立矩阵形式的分配模型，便可利用该方法进行分配与决策。

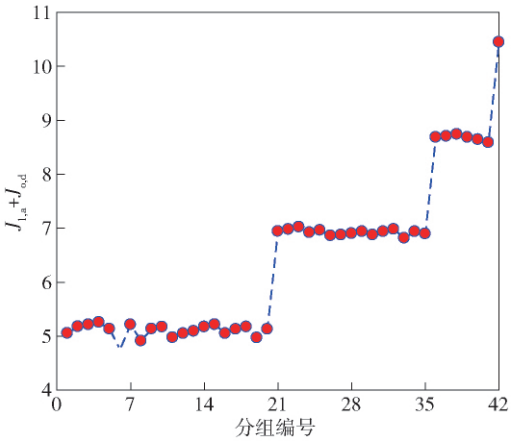

图 8　攻防性能指标 $J_{l,a}+J_{o,d}$

Fig. 8　Attack-defense performance index $J_{l,a}+J_{o,d}$

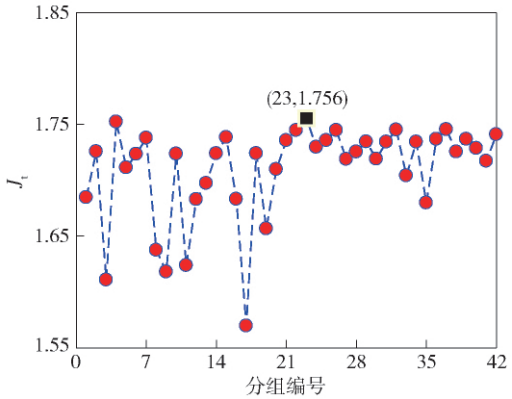

图 9　攻防效费比指标 J_t

Fig. 9　Attack-defense cost-effectiveness ratio index J_t

参考文献(References)

[1] 任章,郭栋,董希旺. 飞行器集群协同制导控制方法及应用研究[J]. 导航定位与授时,2019,6(5):1-9.
REN Z, GUO D, DONG X W. Research on the cooperative guidance and control method and application for aerial vehicle swarm systems[J]. Navigation Position & Timing, 2019, 6(5): 1-9. (in Chinese)

[2] BOGDANOWICZ Z R, TOLANO A, PATEL K, et al. Optimization of weapon-target pairings based on kill probabilities[J]. IEEE Transactions on Cybernetics, 2013, 43(6): 1835-1844.

[3] 卢森堂. 导弹自主编队协同制导控制技术[M]. 北京: 国防工业出版社,2015: 88-96.
LU S T. Cooperative guidance & control of missiles autonomous formation[M]. Beijing: National Defense Industry Press, 2015: 88-96. (in Chinese)

[4] 刘树衍,王航宇,卢发兴. 多枚反舰导弹协同攻击在线目标分配[J]. 指挥控制与仿真,2016,38(1):38-40,52.
LIU S K, WANG H Y, LU F X. Online target assignment for cooperative attack of anti-ship of multiple missiles[J]. Command Control & Simulation, 2016, 38(1): 38-40,52. (in Chinese)

[5] ZHAO M, ZHAO L L, SU X H, et al. Improved discrete mapping differential evolution for multi-unmanned aerial vehicles cooperative multi-targets assignment under unified model[J]. International Journal of Machine Learning & Cybernetics, 2017, 8(3): 765-780.

[6] DING Y F, YANG L Q, HOU J Y, et al. Multi-target collaborative combat decision-making by improved particle swarm optimizer[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2018, 35(1): 181-187.

[7] SUN J J, LIU C S. Finite-horizon differential games for missile-target interception system using adaptive dynamic programming with input constraints[J]. International Journal of System Science, 2018, 49(2): 264-283.

[8] 吴蔚楠. 多无人飞行器分布式任务规划技术研究[D]. 哈尔滨: 哈尔滨工业大学,2018: 20-32.
WU W N. Research on distributed mission planning for multiple unmanned aerial vehicles[D]. Harbin: Harbin Institute of Technology,2018: 20-32. (in Chinese)

[9] CHEN W N, ZHANG J, CHUNG H S H, et al. A novel-based particle swarm optimization model for discrete optimization problems[J]. IEEE Transactions on Evolutionary Computation, 2010, 14(2): 278-300.

[10] 费爱国,张陆游,刘刚,等. 基于粒子群拍卖混合算法的空空导弹制导权移交技术[J]. 宇航学报,2013,34(3):340-346.
FEI A G, ZHANG L Y, LIU G, et al. The technique for air-to-air missile guidance superiority handover based on particle swarm auction hybrid algorithm[J]. Journal of Astronautics, 2013, 34(3): 340-346. (in Chinese)

[11] PRASHANT B, FARUK K, NAVDEEP S. Reinforcement learning based obstacle avoidance for autonomous underwater vehicle[J]. Journal of Marine Science and Application, 2019, 18(2): 228-238.

[12] JUNELL J J, VAN KAMPENY E J, VISSER C D, et al. Reinforcement learning applied to a quadrotor guidance law in autonomous flight[C]// Proceedings of AIAA Guidance, Navigation, and Control Conference. Kissimmee, FL, US: AIAA, 2015.

[13] GAUDET B, FURFARO R. Missile homing-phase guidance law design using reinforcement learning[C]// Proceedings of AIAA Guidance, Navigation, and Control Conference. Minneapolis, MN, US: AIAA, 2012.

[14] GAUDET B, FURFARO R, LINARES R. Reinforcement learning for angle-only intercept guidance of maneuvering targets[C]// Proceedings of AIAA SciTech Forum. Orlando, FL, US: AIAA, 2020.

[15] 张秦浩,敖百强,张秦雪. Q-learning强化学习制导律[J]. 系统工程与电子技术,2020,42(2):414-419.
ZHANG Q H, AO B Q, ZHANG Q X. Reinforcement learning guidance law of Q-learning[J]. Systems Engineering and Electronics, 2020, 42(2): 414-419. (in Chinese)

[16] 刘冰雁,叶雄兵,岳智宏,等. 基于多组并行深度Q网络的连续空间追逃博弈算法[J]. 兵工学报,2021,42(3):663-672.
LIU B Y, YE X B, YUE Z H, et al. Continuous space pursuit-evasion game algorithm based on multi-group deep Q-network[J]. Acta Armamentarii, 2021, 42(3): 663-672. (in Chinese)