基于 *Q*-network 强化学习的超视距空战机动决策

张　强，杨任农，俞利新，张　涛，左家亮
（空军工程大学空管领航学院，西安，710051）

摘要　考虑到空空导弹对空战胜负的重要影响，针对空战态势状态特征连续、多维的情况以及传统方法缺乏对空战对抗中敌方策略的考虑，将强化学习应用到 1vs1 超视距空战机动决策。首先，建立了同时为对抗双方进行机动决策的强化学习框架，提出 ε-纳什均衡策略来选取机动动作，并通过导弹攻击区优势函数来修正奖赏函数；其次，基于记忆库和目标网络训练 *Q*-network，形成超视距空战机动决策的“价值网络”；最后，设计了 *Q*-network 强化学习决策模型，并将机动决策过程分为了学习阶段与实战阶段。仿真结果表明：智能体可以感知空战的态势并作出合理的超视距空战机动决策。

关键词　超视距空战；机动决策；强化学习；纳什均衡
DOI　10.3969/j.issn.1009-3516.2018.06.002
中图分类号　V325　　**文献标志码**　A　　**文章编号**　1009-3516(2018)06-0008-07

BVR Air Combat Maneuvering Decision by Using *Q*-network Reinforcement Learning

ZHANG Qiang, YANG Rennong, YU Lixin, ZHANG Tao, ZUO Jialiang
(Air Traffic Control and Navigation College, Air Force Engineering University, Xi'an 710051, China)

Abstract: In consideration of the great impact of missiles on air combat, the continuous and multidimensional state space and the weakness of traditional approaches in ignoring opponent's strategy in the air combat, reinforcement learning is applied to 1vs1 beyond visual range (BVR) air combat maneuvering decisions. Firstly, a new reinforcement learning framework is built to decide both sides' maneuvers. In this framework, ε-Nash equilibrium strategy is proposed to choose action, and reward function is revised by missile attack zone scoring function. Then, by using a memory base and a target network, *Q*-network can be trained, forming a "value network" for BVR air combat maneuvering decisions. Finally, *Q*-network reinforcement learning model is designed, and the whole maneuvering decision is divided into learning part and strategy forming part. In the simulation, considering that the enemy in the air combat confrontation adopts a fixed maneuver and the two sides are both agents, the former agent wins, and the latter has the advantage of the situation to win, verifying that the agent can perceive the situation of air combat and make a reasonable BVR air combat maneuver.

Key words: beyond visual range air combat; maneuvering decision; reinforcement learning; Nash equilibrium

收稿日期：2018-07-04
作者简介：张　强(1994—)，男，河北邯郸人，硕士生，主要从事智能空战研究。E-mail:2593122568@qq.com

引用格式：*张强，杨任农，俞利新，等. 基于Q-network强化学习的超视距空战机动决策[J]. 空军工程大学学报(自然科学版)，2018，19(6)：8-14. ZHANG Qiang，YANG Rennong，YU Lixin，et al. BVR air combat maneuvering decision by using Q-network reinforcement learning[J]. Journal of Air Force Engineering University (Natural Science Edition)，2018，19(6)：8-14.*

超视距空战是空战的“第一回合”，战机作为搭载武器的平台，通过机动决策获得武器发射优势，并规避敌方的导弹攻击。因此，机动决策是打赢现代化空战的关键。

国内外学者对此做了大量研究，文献[1]将自主空战的机动决策分为基于对策方法以及基于人工智能 2 种，前者比较有代表性的是基于矩阵对策的方法[2]、基于影响图的方法[3]以及微分对策法[4]等。但这些方法求解困难，难以满足实时性。因此，基于人工智能的机动决策成为当前机动决策研究的热点，代表性的方法有：专家系统法[5-6]、神经网络法[7]、遗传算法[8]、人工免疫系统法[9]以及强化学习方法[10-13]等。大多数文献对于空战对抗的另一方决策的处理，一般给定其运动规律，或者已知对方的决策，均与实际空战不符。

本文通过引入导弹攻击区，建立其与强化学习中奖赏函数的相关关系；针对空战态势状态特征连续、多维的实际，基于记忆库和目标网络训练 Q-network，获得值函数的近似值；考虑到双方的对抗，同时为对抗双方进行机动决策，提出 ε-纳什均衡策略以选取机动动作。

1　超视距空战关键要素描述

1.1　超视距空战态势描述

1vs1 超视距空战态势的描述主要是对红蓝双方 2 架战机的空间位置以及相对运动状态的描述。描述空战态势的典型参数有：红蓝双方两机的空间位置、速度、方位角、进入角以及距离变化率。

如图 1 所示，建立地面坐标系，在大地上选取某一固定点作为原点 o，ox 轴取正北方向；oy 轴取铅垂方向，向上为正；oz 轴取正东方向。

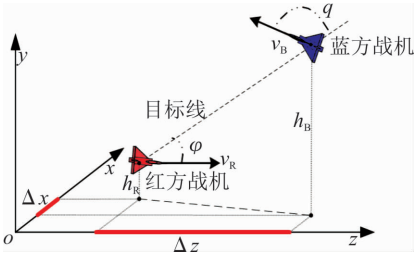

图 1　空战态势几何关系

Fig. 1　Geometrical relationship between two sides in air combat

给定红方战机的空间坐标 $\boldsymbol{c}_R$：(x_R, y_R, z_R) 以及速度矢量 $\boldsymbol{v}_R$，蓝方战机的空间坐标 $\boldsymbol{c}_B$：(x_B, y_B, z_B) 以及速度矢量 $\boldsymbol{v}_B$。目标线 RB 指红方战机 R 到蓝方战机 B 的连线。

目标方位角 φ 指红方战机速度方向与目标线的夹角，右偏为正。其中，$0 \leqslant |\varphi| \leqslant 180°$，有：

$$|\varphi| = \left| \arccos\left(\frac{\boldsymbol{v}_R \cdot (\boldsymbol{c}_B - \boldsymbol{c}_R)}{|\boldsymbol{v}_R|\,|\boldsymbol{c}_B - \boldsymbol{c}_R|} \right) \right| \frac{180}{\pi} \tag{1}$$

目标进入角 q 指蓝方战机速度方向与目标线延长线的夹角，右偏为正。其中 $0 \leqslant |q| \leqslant 180°$，有：

$$|q| = \left| \arccos\left(\frac{\boldsymbol{v}_B \cdot (\boldsymbol{c}_B - \boldsymbol{c}_R)}{|\boldsymbol{v}_B|\,|\boldsymbol{c}_B - \boldsymbol{c}_R|} \right) \right| \frac{180}{\pi} \tag{2}$$

1.2　超视距空战优势条件

超视距空战的首要原则是先敌发现、先敌成功完成导弹的拦射[15]。本文假设红蓝双机的态势透明，主要研究先敌成功完成导弹的发射。利用红蓝双方空战态势，结合导弹攻击区的计算，判定是否有一方进入对方的导弹攻击区内，一旦进入，即可发射导弹。一方成功完成导弹的发射，可以认为空战结束，率先达到导弹发射条件的一方获胜。

导弹攻击区的近似计算通常有数值拟合[16]以及神经网络等方法[13]。本文直接采用文献[16]中的方法，利用空战态势的要素，对某型中程空空导弹攻击区进行拟合，分别得到在迎头与尾后攻击态势下攻击区的远、近边界。

2　基于 Q-Network 强化学习的超视距空战机动决策

强化学习是智能体在一个未知环境中优化其行为的一种途径，其任务通常用马尔可夫决策过程来描述[17]。强化学习中，策略是指状态到动作的映射，学习的目的则是寻求最大化累计奖赏的策略。即寻求最优策略：

$$\pi^* = \arg\max_{\pi} E_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t \right] \tag{3}$$

式中：π^* 为最优的策略；γ 为折扣率；r_t 为 t 时刻的即时奖赏。

2.1　超视距空战的强化学习框架

建立超视距空战的强化学习框架，本文将红蓝双方的战机均作为智能体，即对抗的双方同时进行学习。

环境的状态为红蓝双方共同感知；智能体根据状态，为红蓝双方各决策一个动作；由于是确定性的环境，环境的下一时刻状态仅由当前状态和当前动作唯一确定；结合导弹攻击区的计算，环境为智能体反馈奖赏。智能体再次感知环境的状态，至此，完成了一次循环。

如图 2 所示，其模型表示为智能体与环境的相互作用过程。

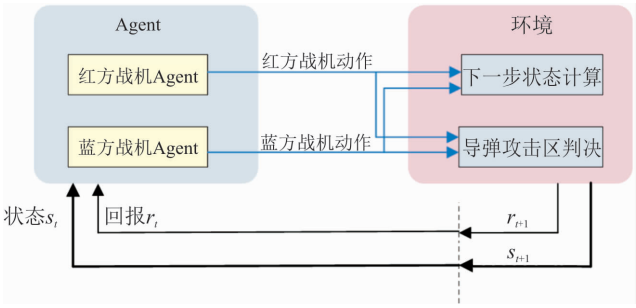

图 2 强化学习框架

Fig. 2 Framework of reinforcement learning

利用强化学习方法进行超视距空战机动决策时,首先需要确定状态空间与动作空间。

状态空间可以通过若干个状态特征来表示。利用红蓝双方战机的空间相对位置以及速度,并结合据此计算出的描述空战态势的相关角度,就可以表示整个超视距空战的态势。并考虑到计算导弹攻击区与作战飞机的高度密切相关,故而选取 8 个状态特征表征状态空间:$s=\{\Delta x, y_R, y_B, \Delta z, \gamma_R, \psi_R, \gamma_B, \psi_B\}$。式中:$\Delta x, \Delta z$ 为红蓝双方战机在正北方向、正东方向的位置差;y_R, y_B 为红蓝双方战机的高度;$\gamma_R, \psi_R, \gamma_B, \psi_B$ 分别为红蓝双方战机航迹倾角和航迹偏角。

为简化动作空间,选取空战中 5 个基本机动动作,即:向左转弯,向右转弯,跃升,俯冲以及保持当前运动状态。即动作空间定义为:{turnleft, tureright, up, down, stay}。

2.1.1 奖赏函数

奖赏函数是指导智能体完成预期目标的关键,经典强化学习任务中一般根据达成目标的情况制定有限个离散的奖惩值。在超视距空战中,以对方是否进入导弹攻击区,简单定义奖赏函数。但在模拟空战中大部分是平局,强化信号变得尤为稀疏,且奖赏函数的离散性,使得智能体的学习过程非常漫长。文献[12]提出以近距空战的相对优势定义态势得分来修正奖赏函数,解决了强化信号的稀疏性,并将奖赏函数连续化,加快了学习的速度。本文将导弹攻击区的优势态势进行量化,以此修正奖赏函数,可以在整个状态空间中给出奖赏信号,提升了强化学习的效率。

由拟合的导弹攻击区可知[16],满足导弹发射条件有 2 条:①两机的相对角度处于导弹发射允许范围;②两机的距离在导弹攻击区的远近边界范围之间。定义 $\Omega_1, \Omega_2, \Omega_3, \Omega_4$ 表示空战状态分别为:我机处于迎头攻击态势,且满足导弹发射的第 1 个条件;我机处于尾后攻击态势,且满足导弹发射的第 1 个条件;我机处于迎头攻击态势,不满足导弹发射的第 1 个条件;我机处于尾后攻击态势,不满足导弹发射的第 1 个条件。

导弹的攻击效能与距离和弹目相对空间几何关系均有关,结合两机距离 D 与导弹攻击区远近边界 D_{max}, D_{min} 的关系,并分别考虑迎头和尾后态势下的弹目空间位置关系,定义态势得分函数 $A(s)$:

$$A(s)=\begin{cases}\left[\dfrac{|q|}{180}\omega_1+\dfrac{180-|\varphi|}{180}(1-\omega_1)\right]\cdot \\ \exp\left(-\dfrac{|D-0.5(D_{max}+D_{min})|}{k}\right), s\in\Omega_1 \\ \left[\dfrac{180-|q|}{180}\omega_2+\dfrac{180-|\varphi|}{180}(1-\omega_2)\right]\cdot \\ \exp\left(-\dfrac{|D-0.5(D_{max}+D_{min}|}{k}\right), s\in\Omega_2 \\ \left[\dfrac{|q|}{180}\omega_1+\dfrac{180-|\varphi|}{180}(1-\omega_1)\right]\times 0.1, s\in\Omega_3 \\ \left[\dfrac{180|q|}{180}\omega_2+\dfrac{180-|\varphi|}{180}(1-\omega_2)\right]\times 0.1, s\in\Omega_4 \\ 0, s\notin\{\Omega_1\parallel\Omega_2\parallel\Omega_3\parallel\Omega_4\}\end{cases} \tag{4}$$

式中:ω_1, ω_2 为目标方位角与目标进入角的权重因子,k 用来权衡空战态势的角度因素与导弹攻击区远、近边界对空战优势的影响。可以得出,$A(s)\in[0,1]$。

由此,以红方战机为例,对奖赏函数修正如下:

$$r_R(s)=\omega e_R(s)+(1-\omega)A_R(s) \tag{5}$$

式中:ω 为权重,且有 $\omega\in(0,1)$;$r_R(s)$ 为修正过的红方战机的奖赏函数,包括 2 部分:$e_R(s)$ 指示红方是否满足导弹攻击条件,满足时为 1,否则为 0;$A_R(s)$ 为红方态势得分函数。同理,给出蓝方的奖赏函数 $r_B(s)$。

值得注意的是,本文的强化学习模型需要学习红蓝双方的空战策略,奖赏函数的设置需要权衡双方的收益。由于蓝方的奖赏即红方的损失,反之亦然,最后给出全局的奖赏函数:

$$r(s)=r_R(s)-r_B(s) \tag{6}$$

2.1.2 动作选取

红蓝双机对抗是一个博弈的过程,一方面争取本机的最大优势;另一方面避免遭受敌机的攻击。因此,智能体在进行动作的选取上,不能简单类比传统方法选择达到最优 Q 值的动作对。作为对抗的双方,各自争取利益的最大化,且一方的优势就是另一方的损失,红蓝双机的动作选取有限,各有 5 个动作可供选择;可将该博弈过程视为 2 人有限零和博弈,还可以视为矩阵博弈。对于一个动作对$<a_R, a_B>$,相应红方有一个赢得值,为 $Q(s, a_R, a_B)$。为简化描述,使用 Q_{ij} 表示,其中 i, j 表示红蓝双方战机选择的动作编号。故而赢得矩阵为:

$$\boldsymbol{A}=\begin{bmatrix} Q_{11} & \cdots & Q_{15} \\ \vdots & & \vdots \\ Q_{51} & \cdots & Q_{55} \end{bmatrix}$$

在博弈过程中，若存在最优纯策略，即满足：

$$\max_i \min_j Q_{ij}=\min_j \max_i Q_{ij}=Q_{i^*j^*} \tag{7}$$

这对于红蓝双方都是最稳妥的行为。任何一方背离了该动作，都会令自身利益损失，相反，另一方会受益。

然而，这种纯策略并不一直存在，2 人有限零和博弈在混合策略意义下的平衡局势则一定存在。此时，要满足纳什均衡，红蓝双方并不选取单独的一个动作，而是各自以某一概率选取动作。即红方以 $\boldsymbol{P}_r=[p_1,p_2,p_3,p_4,p_5]$ 的概率选择 5 个动作；蓝方以 $\boldsymbol{P}_b=[q_1,q_2,q_3,q_4,q_5]$ 的概率选择 5 个动作。任一矩阵对策的求解等价于一对互为对偶的线性规划问题[18]。

当存在最优纯策略时，红蓝双方选择达到该平衡的动作；当纳什均衡以混合策略给出时，则红蓝双方各以一定的概率选择动作。在混合策略下的动作选取算法，本文采用轮盘赌选法确定两方的动作。以红方为例，给出在混合策略下的动作选取算法。

Input：p_i，　$i=1,2,3,4,5$

Init：动作编号$=1$，$i_{\text{new}}=1$

rnd$=U(0,1)$；

$\boldsymbol{P}^*=[p_1,\sum_{i=1}^{2}p_i,\sum_{i=1}^{3}p_i,\sum_{i=1}^{4}p_i,\sum_{i=1}^{5}p_i]$

While($P^*_{i_{\text{new}}}<$rnd)

　　$i_{\text{new}}=i_{\text{new}}+1$；

　　$=i_{\text{new}}$；

End

Output：动作编号

为探索到更广泛的状态空间，本文采用 ε-纳什均衡策略：以 ε 的概率选择随机动作，以 1−ε 的概率选择满足纳什均衡的动作。由于学习初期对状态感知并不准确，应加大探索的力度；随着学习的深入，加大利用的力度。因此，给出 ε 的计算公式：

$$\varepsilon=\max\{0.05,1-\frac{\text{训练次数}}{1\ 000}\} \tag{8}$$

用$\underset{a_R,a_B}{\text{Nash}}Q(s,a_R,a_B)$表示采用 ε-纳什均衡策略进行动作的选取时对应的 Q 值。

2.1.3　状态转移

智能体采取机动动作后，环境的状态将发生改变。根据机动动作对应的控制量求解出作战飞机的状态变化量 Δs，即根据控制量 $\boldsymbol{u}=\{\phi,n\}$ 结合当前状态量 $s=\{\Delta x,y_R,y_B,\Delta z,\gamma_R,\psi_R,\gamma_B,\psi_B\}$，计算新的状态量 $s'\leftarrow s+\Delta s$。其中 ϕ 为滚转角，n 为法向过载。

作战飞机在航迹坐标系上的质点运动学方程为：

$$\begin{aligned}\dot{x}&=v\cos\gamma\cos\psi\\ \dot{y}&=v\sin\gamma\\ \dot{z}&=-v\cos\gamma\sin\psi\end{aligned} \tag{9}$$

式中：x,y,z 为作战飞机在地面坐标系中的坐标；$\dot{x},\dot{y},\dot{z}$ 表示速度在 3 个坐标轴上的分量；联立作战飞机动力学方程：

$$\begin{aligned}\dot{\gamma}&=\frac{g}{v}(n\cos\phi-\cos\gamma)\\ \dot{\psi}&=\frac{gn\sin\phi}{v\cos\gamma}\end{aligned} \tag{10}$$

通过求解作战飞机运动学与动力学方程，结合执行动作周期 Δt 求解出作战飞机的状态变化量：$\Delta s=\Delta t\{\dot{x}_B-\dot{x}_R,\dot{y}_R,\dot{y}_B,\dot{z}_B-\dot{z}_R,\dot{\gamma}_R,\dot{\psi}_R,\dot{\gamma}_B,\dot{\psi}_B\}$

下面分别给出红蓝双方战机的不同动作的控制量，并设置红方跃升的法向过载有优势：

①向左转弯：$\begin{cases}\phi_R=60^\circ,n_R=1\\ \phi_B=60^\circ,n_B=1\end{cases}$

②向右转弯：$\begin{cases}\phi_R=-60^\circ,n_B=1\\ \phi_B=-60^\circ,n_B=1\end{cases}$

③跃升：$\begin{cases}\phi_R=0^\circ,n_R=6\\ \phi_B=0^\circ,n_B=5\end{cases}$

④俯冲：$\begin{cases}\phi_R=0^\circ,n_R=-3\\ \phi_B=0^\circ,n_B=-3\end{cases}$

⑤保持当前运动状态：$\dot{\gamma}=0,\dot{\psi}=0$。

2.2　Q-network 模型建立

由 2.1.2 节可知，超视距空战的机动策略的给出需要已知 $Q(s,a_R,a_B)$。由于状态空间的状态特征量是连续的、多维的、离散化后的状态量随空间维数呈指数增长，故而 $Q(s,a_R,a_B)$ 无法通过经典强化学习的基于表格值的算法给出。文献[14]提出了 DQN(Deep Q-network，深度 Q 网络)，利用深度卷积神经网络对 Q 值进行估计。本文利用结构类似于 DQN 的神经网络来近似 $Q(s,a_R,a_B)$，解决状态空间连续所带来的“维数灾难”。

2.2.1　构建 Q-network 结构

Q-network 由输入层、隐含层和输出层构成，采用 3 层全连接的网络结构。状态特征作为输入，输入层有 8 个节点；“状态-动作对值”的估计 $Q(s,a_R,a_B)$ 作为输出，输出层有 25 个节点。从输入层到隐含层的激活函数为 Sigma 函数，从隐含层到输出层为线性输出。

强化学习的函数近似方法中，由于神经网络属于非线性方法，容易产生 Q 估计值不收敛[14]。

DQN 利用了 2 个处理方式解决该问题:定义记忆库,将以往的状态执行情况记录下来,并打乱顺序,实现"经验回放";定义目标网络,该网络的结构与 Q-network 相同,且目标网络参数 θ^- 由 Q-network 网络参数 θ 确定,唯一区别在于目标网络参数 θ^- 固定一定周期后,再更新为当前的 Q-network 网络参数 θ。实验表明,应用上述处理,可以有效避免 Q 估计值不收敛的问题。

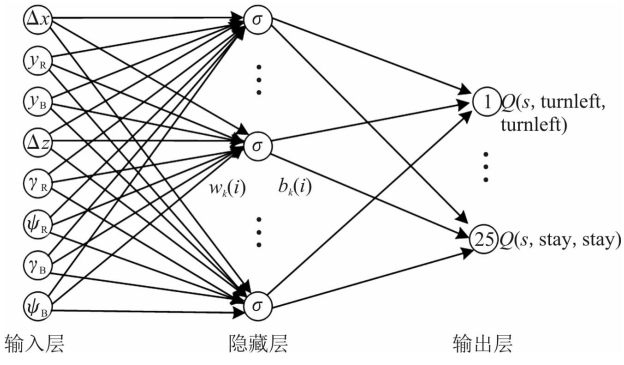

图 3 Q-network 网络结构

Fig. 3 Structure of Q-network

2.2.2 构建记忆库与目标网络

记忆库是 Q-network 的训练样本库,每一组数据记录着策略的执行情况$<s,a_R,a_B,r,s'>$:即当前的状态 s,红蓝双方采取的动作 a_R,a_B,对应的奖赏函数值 r 以及下一时刻状态 s'。在学习过程中,不断有新的策略执行数据加入到记忆库,为保证最新的数据能够被用来训练网络,需设置记忆库的更新规则。设置记忆库的最大容量 $M_{\max}$,新进入的第 n 组数据在记忆库的位置记为 M_{index},利用取模运算计算该位置,有:$M_{\text{index}}=n(\text{mod } M_{\max})$,并取代之前在该位置的数据。

目标网络实际上是另一个 Q-network,其网络参数 θ^- 不是通过训练进行更新的,而是直接复制 Q-network 网络参数 θ,但该复制过程是每隔固定时间进行一次。即网络参数 θ^- 每隔固定时间更新一次,其余时间保持不变。构建目标网络是为了生成 Q-network 的目标输出,通过强化学习中的 TD 误差来更新网络参数 θ。本文设置 Q-network 每更新 100 次,目标网络参数复制一次。

2.2.3 Q-network 的训练

Q-network 训练的目标是对 Q 值尽可能准确地作出估计。利用以往执行策略的数据,通过网络参数的更新以减小 TD 误差,使得 Q-network 尽可能逼近准确的 Q 值。

利用 TD 误差,定义损失函数 loss:

$$\text{loss}=\frac{1}{2}\left[r(s)+\gamma \underset{a'_R a'_B}{\text{Nash}}Q(s',a';\theta^-)-\underset{a_R a_B}{\text{Nash}}Q(s,a;\theta)\right]^2 \quad (11)$$

式中:$r(s)$ 为奖赏函数;γ 为折扣率;$\underset{a'_R a'_B}{\text{Nash}}Q(s',a';\theta^-)$,$\underset{a_R a_B}{\text{Nash}}Q(s,a;\theta)$ 分别表示在 ε-纳什均衡的动作选取策略下,分别利用目标网络估计的 $Q(s',a')$ 值,和利用 Q-network 估计的 $Q(s,a)$ 值。

利用随机梯度下降(SGD)算法,更新网络参数,设置小样本的数量为 100 组。整个训练过程见图 4。

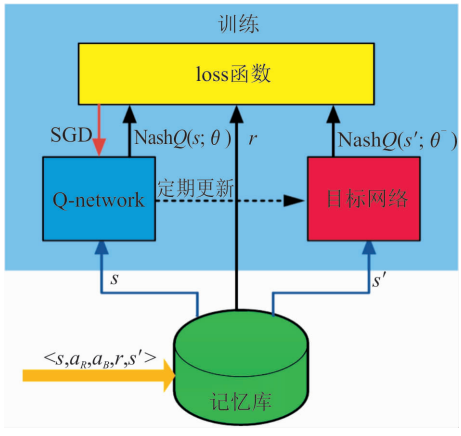

图 4 Q-network 训练过程示意

Fig. 4 Training process of Q-network

2.2.4 Q-network 超参数设置

Q-network 的隐含层节点数的设置影响到模型的精度和训练时间的长短。本文以损失函数为衡量标准,确定合适的隐含层节点数。由于损失函数的变化较为剧烈,难以直接比较,本文采用 S-G 滤波器对数据进行平滑处理,见图 5。

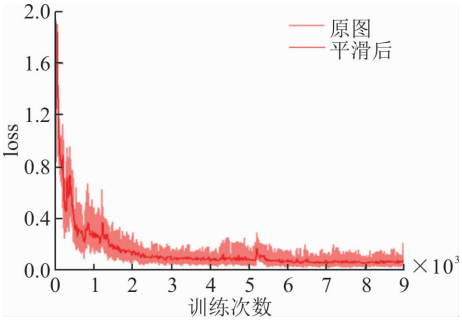

图 5 损失函数的平滑

Fig. 5 Smoothness of loss function

对比平滑后的损失函数值,见图 6,容易发现:在训练前期,节点数的变化会对 loss 函数值有比较大的影响,节点数越多,loss 函数值下降得越快。但在训练后期,loss 函数值已经收敛到基本相同的值。训练相同的片段数,训练次数越少,则时间越短,即能在较短时间内学习到一样多的知识。故而本文选取隐含层节点数为 200。

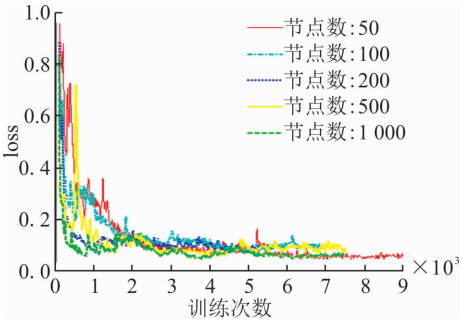

图 6　损失函数与节点数

Fig. 6　Loss function and number of nodes

2.3　基于 Q-network 的强化学习

基于 Q-network 的强化学习应用于超视距空战机动决策分成 2 个阶段：

第 1 阶段是学习阶段，首先结合初始化的 Q-network，利用强化学习框架进行 Q 值的更新，将状态特征、动作对、奖赏函数值等更新到记忆库中。给定初始位置，利用强化学习框架进行动作的选取和状态转移，空战的数据实际上在强化学习过程中产生。

由于状态空间连续，学习速度会非常缓慢，故而本文考虑将空战的初始状态固定。从空战的初始状态出发，当红蓝中的任意一方或双方同时达到导弹发射条件，或当执行动作次数达到设置的最大次数时，表示该片段结束。应当注意的是，在学习阶段动作的选择采用的是 ε-纳什均衡策略。

当进入到记忆库的数据组数达到 Q-network 训练小样本时所需的数量，即开始增量训练。新的 Q-network 重新作用于强化学习的过程，直到强化学习的片段数达到期望值，学习过程结束。

第 2 阶段为实战阶段，在该阶段，关闭“记忆库”，不再对 Q-network 的权值和偏置进行更新。由于网络已经训练完毕，Q-network 可以快速输出估计的 Q 值，进而采用纳什均衡策略选取我方的动作，能够满足实时性的要求。

3　仿真实验

设定空战决策周期为 2 s，红蓝双方的初始态势固定：红方战机的空间坐标 $\boldsymbol{c}_R$ 为(3 000，4 000，8 000)，红方战机航迹倾角和航迹偏角分别为：$\gamma_R=0°$，$\psi_R=0°$，速度大小 $v_R=350$ m/s。蓝方战机的空间坐标 $\boldsymbol{c}_R$ 为(30 000，4 000，9 000)，蓝方战机航迹倾角和航迹偏角分别为：$\gamma_B=0°$，$\psi_B=180°$，速度大小 $v_B=300$ m/s。奖赏函数的权重 ω 为 0.7，目标方位角与目标进入角的权重因子 ω_1，ω_2 分别取 0.25，0.4，用来权衡空战态势的角度因素与导弹攻击区远、近边界对空战优势的影响的 k 取 0.000 1，强化学习中折扣率 γ 取 0.9，神经网络的学习率 η 取 0.001。利用以上参数，结合 Q-network 强化学习对红蓝双方的空战机动决策进行仿真。

当红蓝双方空战对抗 1 000 个片段后，停止 Q-network 的学习，红蓝双方根据该价值网络进行机动决策。红方战机采用固定的机动动作，保持平飞，蓝方战机利用纳什均衡策略选取动作，双方空战对抗见图 7；作为对照，红蓝双方均采用纳什均衡策略选取动作，双方空战对抗见图 8。

实验采用的硬件平台为 Intel(R) Xeon(R) CPU ES-2643，主频 3.50 GHz，内存 8 GB。软件配置为 Microsoft Windows 7 旗舰版 32 位操作系统，Matlab2014a 运行环境。决策一次行动的时间平均为 0.02 s，满足实时性的要求。

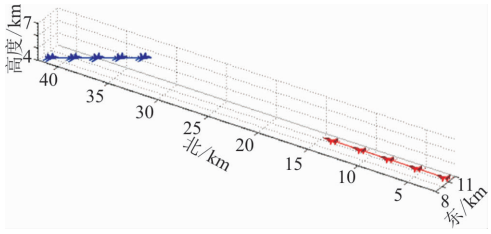

图 7　红方采取固定机动的交战轨迹

Fig. 7　Fight trajectory when red use fixed strategy

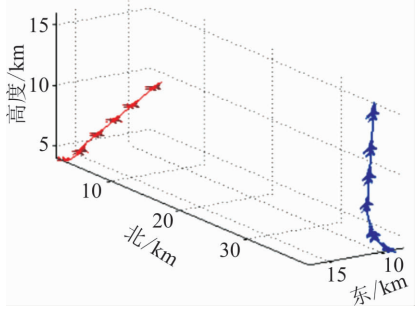

图 8　双方均采用纳什均衡策略交战轨迹

Fig. 8　Fight trajectory when both fighters use Nash equilibrium strategy

在红方采取固定动作时，蓝方保持爬升，抓住高度优势，率先获得先敌发射条件。在双方均采用纳什均衡策略时，起初两方均采用爬升动作以增大导弹攻击区的范围，由于红方具有速度优势和过载优势，蓝方感知到态势的劣势，一方面保持垂直爬升的姿态，利用高度的优势，另一方面保持与红方的距离，尽量不被对方攻击到。最终，还是红方率先获得先敌发射条件。

仿真表明，智能体能够良好地感知空战态势，作出合理的机动策略获取空战的优势，并能够权衡态势，作出相应的规避。

4　结语

本文设计了基于 Q-network 的强化学习方法，同时为对抗双方决策机动动作。实验表明，在未知

对方的机动时,智能体仍能作出合理的机动决策。且决策一次动作的时间约为 0.02 s,达到实时性的要求。该研究对于智能自主空战的机动决策具有重要的理论价值和现实意义,不足之处在于动作的选取上与实际空战仍有一定的差距,下一步的工作将针对机动控制量连续取值的问题进行研究。

机载雷达在超视距空战中的作战效能日益凸显;导弹攻击区有多种形式的表示,本文采用其中的动力攻击区远、近边界,而不同导弹攻击区的攻击效能也有区别。这是空战机动决策的未来发展方向。

参考文献(References):

[1] 周思雨,吴文海,张楠,等.自主空战机动决策方法综述[J].航空计算技术,2012,42(1):27-31.
ZHOU S Y,WU W H,ZHANG N,et al. Overview of Autonomous Air Combat Maneuver Decision[J]. Aeronautical Computing Technique,2012,42(1):27-31. (in Chinese)

[2] AUSTIN F,CARBONE G,MICHAEL F, el al. Game Theory for Automated Maneuvering during Air-to-Air Combat[J]. Journal of Guidance 1990, 13(6): 1143-1147.

[3] VIRLANEN K, RAIVIO T, HÄMÄLÄINEN R P. Decision Theoretical Approach to Pilot Simulation[J]. Journal of Aircraft, 1999,27(4):632-641.

[4] HORIE K,CONWAY B. Optimal Fighter Pursuit-Evasion Muneu Vers Found via Two-Sided Optimization [J]. Journal of Cuidance,Control and Dynamics,2006, 29(1):105-112.

[5] MCMANUS J W,CHAPPELL A R,ARBUCKLE D P. Situation Assessment in the Paladin Tactical Decision Generation System[R]. AGARD Conference Proceedings of Air Vehicle Mission Control and Management,1992:1-10.

[6] 祝世虎,董朝阳,张金鹏,等.基于神经网络与专家系统的智能决策支持系统[J].电光与控制,2006,13(1):8-11.
ZHU S H,DONG C Y,ZHANG J P,et al. An Intelligent Desicion-Making System Based on Neural Networks and Expert System [J]. Electronics Optics&Control,2006,13(1):8-11. (in Chinese)

[7] ROGER W S,ALAN E B. Neural Network Models of Air Comb-at Maneuvering[D]. New Mexico: New Mexico State University,1992.

[8] 张涛,于雷,周中良,等.基于变权重伪并行遗传算法的空战机动决策[J].飞行力学,2012,30(5):470-474.
ZHANG T,YU L,ZHOU Z L,et al. Decision-Making for Air Combat Maneuvering Based on Variable Weight Pseudo-Parallelgenetic Algorithm[J]. Flight Dynamics,2012,30(5):470-474. (in Chinese)

[9] KRISHNA K K, KANESHIGE J. Artificial Immune System Approach for Air Combat Maneuvering[C]// Proc of the 41st Aerospace Sciences Meeting&Exhibit. Reno, Nevada,2003:1-10.

[10] MCG J S,HOW J P,WILLIAMS B,et al. Aircombat Strategy Using Approximate Dynamic Programming [J]. Journal of Guidance,Control,and Dynamics,2010, 33(5):1641-1654.

[11] 徐安,于雷,寇英信,等.基于 MDP 框架的飞行器隐蔽接敌策略[J].系统工程与电子技术,2011,33(5):1063-1068.
XU A,YU L,KOU Y X,et al. Stealthy Eagement Maneuvering Strategy for Air Combat Based on MDP[J]. Systems Engineering and Electronics, 2011, 33(5): 1063-1068. (in Chinese)

[12] 徐安,寇英信,于雷,等. 基于 RBF 神经网络的 Q 学习飞行器隐蔽接敌策略[J].系统工程与电子技术,2012,34(1):97-101.
XU A,KOU Y X,YU L,et al. Stealthy Engagement Maneuvering Strategy with Q-Learning Based on RBFNN for Air Vehicles[J]. Systems Engineering and Electronics,2012,34(1):97-101.(in Chinese)

[13] 左家亮,杨任农,张滢,等.基于启发式强化学习的空战机动智能决策研究[J].航空学报,2017,38(10):321168.
ZUO J L,YANG R N,ZNANG Y,et al. Research on Air Combat Maneuvering Intelligence Decision-Making Based on Heuristic Reinforcement Learning[J]. Acta Aeronautics et Astronautics Sinica, 2017, 38(10): 321168. (in Chinese)

[14] VOLODYMYR M,KORAY K, DAVID S,et al. Humanlevel Control through Deep Reinforcement Learning[J]. NATURE,2015,518:529-533.

[15] 吴文海,周思羽,高丽,等.超视距空战过程分析[J].飞行力学,2011,29(6):45-47.
WU W H,ZHOU S Y,GAO L,et al. Analysis of BVR Air Combat Process[J]. Flight Dynamics,2011,29(6): 45-47. (in Chinese)

[16] 吴文海,周思羽,高丽,等.基于导弹攻击区的超视距空战态势评估改进[J].系统工程与电子技术,2011,33(12):2679-2685.
WU W H,ZHOU S Y,GAO L,et al. Improvements of Situation Assessment for Beyond-Visual-Range Air Combat Based on Missile Launching Envelope Analysis [J]. Systems Engineering and Electronics, 2011, 33(12):2679-2685. (in Chinese)

[17] 陈兴国,俞扬.强化学习及其在电脑围棋中的应用[J]自动化学报,2016,42(5): 685-695.
CHEN X G,YU Y. Reinforcement Learning and Its Application to the Game of Go[J]. Acta Automatica Sinica,2016,42(5):685-695. (in Chinese)

[18] 钱颂迪. 运筹学[M]. 3 版. 北京:清华大学出版社,2005:381-393.
QIAN S D. Operational Research[M]. 3rd Ed. Beijing: Tsinghua University Press,2005:381-393. (in Chinese)

(编辑:徐敏)