文章编号:1002-0640(2019)04-0142-08

基于强化学习的无人机自主机动决策方法*

孙 楚[1],赵 辉[1],王 渊[1],周 欢[1],韩 瑾[2]
(1.空军工程大学航空航天工程学院,西安 710038;2.汾西重工有限责任公司,太原 030027)

摘 要: 提出了一种基于连续动作集强化学习的无人机机动决策方法。通过Actor-Critic强化学习构架下的NRBF神经网络输出状态真实效用值与连续动作控制变量,效用值与动作控制变量的输出层共用隐层以简化网络结构。结合相对熵理论设计隐层节点的自适应调整方法,有效减少了隐层节点数目。对输出动作控制变量,采用基于高斯分布的连续动作选择策略,并依据训练次数调整随机动作控制变量的概率分布,提升了网络对未知策略的探索能力。在3种不同空战态势下的仿真验证了该方法的有效性,结果表明该方法生成的策略鲁棒性较强,动作控制量更加精确。

关键词: 无人作战飞机,机动决策,强化学习,连续动作,神经网络
中图分类号: TJ85 **文献标识码:** A **DOI:** 10.3969/j.issn.1002-0640.2019.04.029
引用格式: 孙楚,赵辉,王渊,等.基于强化学习的无人机自主机动决策方法[J].火力与指挥控制,2019,44(4):142-149.

UCAV Autonomic Maneuver Decision- making Method Based on Reinforcement Learning

SUN Chu[1], ZHAO Hui[1], WANG Yuan[1], ZHOU Huan[1], HAN Jin[2]
(1. *School of Aeronautics and Astronautics Engineering, Air Force Engineering University, Xi'an* 710038, *China*;
2. *Shanxi Fenxi Heavy Industry Corporation, Taiyuan* 030027, *China*)

Abstract: The UCAV autonomic maneuver decision-making method based on reinforcement learning with continuous action space is put forward. A normatized RBF neural network based on Actor-Critic structure is used to approximate the true state value and continuous action control variables, the outputs of value and action share the same hidden layer to simplify the structure of the network. A autonomic adjustment basded on relative entropy is designed to reduce the nodes of the hidden layer. Continuous action space is achieved by a continuous action selector based on Gauss distribution, and the distribution is adjusted by the training episodes to improve the ability for the discovery of unknown strategies. Simulation is conducted in three different air combat circumstances, the results show the effectiveness of the proposed method, the strategies are of strong robustness and the action control variables are more accurate.

Key words: unmanned combat aerial vehicle, maneuver decision-making, reinforcement learning, continuous action, neural network

Citation format: SUN C, ZHAO H, WANG Y, et al. UCAV autonomic maneuver decision-making method based on reinforcement learning[J]. Fire Control & Command Control, 2019, 44(4): 142-149.

收稿日期: 2018-01-13 **修回日期:** 2018-03-25
* **基金项目:** 国家自然科学基金(61601505);航空科学基金资助项目(20155196022)
作者简介: 孙 楚(1992-),男,湖北武汉人,硕士研究生。研究方向:无人飞行器系统及作战技术。

0　引言

机动决策是无人战斗机（Unmanned Combat Aerial Vehicle,UCAV）自主空战决策任务系统的重要组成环节[1-2]。将强化学习技术应用于 UCAV 的机动决策研究中,能够充分发挥强化学习在无样本学习中的优势,其强大的学习能力能够极大地加强 UCAV 的鲁棒性及对动态环境的适应能力，对提升无人机自主作战效能具有重要的意义。

文献[3]使用 RBF 神经网络建立了航空兵认知行为模型,一定程度上克服了空间离散方法的使用困境，但 RBF 网络中隐层节点数固定不变且离散动作集有限,在面对快速变化的空战态势时算法的泛化能力明显不足;文献[4-7]针对连续状态空间强化学习问题进行了详细的研究,提出了一系列用于解决连续状态空间下强化学习应用问题的方法，但其涉及的控制动作控制变量为离散形式,控制情形简单,要应用于复杂的空战机动决策中还有待改进。需要指出的是,现有机动决策的研究成果虽能够解决连续状态空间问题,但均以建立战术动作集与基本动作集为基础[8],将动作空间作离散化处理以实现机动决策。一方面,离散的动作集虽然能够快速生成决策序列,但针对剧烈变化的空战态势缺乏灵活有效的调整方法[9],无法有效应对非典型空战态势;另一方面,离散化粒度粗糙的动作集会使飞行轨迹不够平滑，增大控制系统负担;而离散化粒度精细的动作集会提高问题维度,增大计算负担。采用连续动作集强化学习的机动决策方法建立状态与动作值的映射关系,直接输出连续的动作控制变量,在这种情况下机动动作将不受任何限制,同时控制量形式简单,极大地提升了 UCAV 的机动灵活性，较好地克服了离散动作集的缺点,对提升 UCAV 自主控制水平与作战能力而言非常必要。

针对上述问题，本文采用强化学习中的 Actor-Critic 构架,充分发挥了其解决连续状态空间问题的优势，通过共用隐层的自适应 NRBF 神经网络,实现对不同态势下真实值函数与最优策略的逼近，在输出效用值的同时输出连续动作控制变量;结合熵理论设计了 NRBF 神经网络隐层节点的自适应调整方法;通过构建高斯分布的随机动作控制变量平衡学习中“探索 - 利用”的关系。最后通过不同态势下的仿真分析了决策方法在空战中的对抗能力,并通过对比实验验证了连续动作控制变量集对提升 UCAV 控制性能的有效性。

1　基本模型

1.1　无人机状态与动作描述参数

设红方 UCAV 为我机,蓝方 UCAV 为敌机。状态变量是 UCAV 对态势的描述，输入状态变量越多,则对态势的描述越全面,但会增大网络学习的计算量,选取基本的状态描述变量 x_0 为

$$x_0=[q_r,q_b,V_r,h_r,d,\beta,\Delta h,\Delta V]$$

式中,V_r 与 h_r 分别是我机和敌机 UCAV 的速度和高度;q_r 与 q_b 分别表示我机和敌机 UCAV 速度矢量与其质心的夹角;ΔV 与 Δh 分别表示我机和敌机的速度差、高度差;β 表示我机和敌机 UCAV 速度的矢量夹角;d 表示我机和敌机 UCAV 的相对距离。

为了充分发掘态势信息,x_0 经过数据处理后才能作为输入变量，其中包含了对 q_r,q_b,V_r 的微分运算,实际输入 x 为

$$x=[q_r,\dot{q}_r,q_b,\dot{q}_b,V_r,\dot{V}_r,h_r,d,\beta,\Delta h,\Delta V]$$

这种选取方法侧重于无人机的位置与速度状态,适合描述双方的空战态势。

状态变量与双方在地面坐标系下的换算关系为

$$\begin{aligned}
q_r&=\arccos\{[(x_b-x_r)\cos\psi_r\cos\tau_r+\\
&\quad(y_b-y_r)\sin\psi_r\cos\tau_r+(z_b-z_r)\sin\tau_r]/d\}\\
q_b&=\arccos\{[(x_r-x_b)\cos\psi_b\cos\tau_b+\\
&\quad(y_b-y_r)\sin\psi_b\cos\tau_b+(z_r-z_b)\sin\tau_b]/d\}\\
d&=\sqrt{(x_b-x_r)^2+(y_b-y_r)^2+(z_b-z_r)^2}\\
\beta&=\arccos(\cos\psi_r\cos\tau_r\cos\psi_b\cos\tau_b+\\
&\quad\sin\psi_r\cos\tau_r\sin\psi_b\cos\tau_b+\sin\tau_r\sin\tau_b)
\end{aligned}$$

式中(x_r,y_r,z_r)、(x_b,y_b,z_b)分别为我机 UCAV 与敌机 UCAV 在地面坐标系下的坐标;ψ_r、ψ_b分别为双方 UCAV 的航向角；τ_r、τ_b分别为双方 UCAV 的航迹角。

为简化分析过程,这里选用三自由度质点运动模型,UCAV 动力学方程可表示为

$$\begin{aligned}
\frac{\mathrm{d}v}{\mathrm{d}t}&=g(n_x-\sin\tau)\\
\frac{\mathrm{d}\tau}{\mathrm{d}t}&=\frac{g}{v}(n_z\cos\gamma_x-\cos\tau)\\
\frac{\mathrm{d}\psi}{\mathrm{d}t}&=\frac{g}{v\cos\tau}n_z\sin\gamma_x
\end{aligned}\tag{1}$$

式(1)中,n_x 为切向过载,用于表示 UCAV 的切向加速度;n_z 为法向过载，用于表示 UCAV 的法向加速度;γ_x 为滚转角。

根据 UCAV 的动力学方程,可以得到 UCAV 在地面坐标系下各个方向上的加速度,表示为

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}t}&=v\cos\tau\cos\psi\\ \frac{\mathrm{d}y}{\mathrm{d}t}&=v\cos\tau\sin\psi\\ \frac{\mathrm{d}z}{\mathrm{d}t}&=v\sin\tau\end{aligned}\tag{2}$$

通过对式(1)、式(2)进行运算，可以得到 UCAV 的速度 v，航迹角 τ，航向角 ψ 与 UCAV 在地面坐标系下的坐标；此外，通过改变 n_x、n_z 与 γ_x，可以控制 UCAV 在任意方向上的加速度，因此，可以选择 n_x、n_z 与 γ_x 3 个参数来描述 UCAV 的动作控制量 a，记为

$$a=[n_x,n_z,\ \gamma_x]\tag{3}$$

1.2　Actor-Critic 强化学习模型

在 UCAV 机动决策中应用强化学习方法，首先必须解决连续状态空间的表示问题。Actor-Critic 构架将动作生成与策略评价部分分离，分别进行策略学习与值函数的逼近，能够较好地解决连续状态空间下的决策问题，其基本结构如图 1 所示。

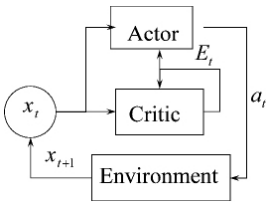

图 1　Actor-Critic 结构

其中，x_t 为 t 时刻的输入状态变量，a_t 为当前策略下 t 时刻生成的动作控制变量。

Critic 部分用于对状态进行评估，是从状态 x_t 到效用值 $V(x_t)$ 的映射关系，表示为

$$x_t\mapsto V(x_t)\tag{4}$$

其中，$V(x_t)$ 是瞬时奖赏 r 的折扣累积，是状态与动作的评价指标。在真实值函数未知的情况下，由时序差分学习方法[10]，设折扣率为 γ，$V(x_t)$ 的更新公式为

$$V(x_t)=r+\gamma V(x_{t+1})\tag{5}$$

同时 Critic 部分存储了过去的效用值，用于生成误差函数 E_t。

Actor 部分用于实现状态 x_t 到动作 a_t 的映射关系，这种映射关系是策略的具体体现，表示为

$$x_t\mapsto a_t\tag{6}$$

对于空战机动决策问题，式(4)、式(6)一般由复杂的非线性映射关系表示，本文采用归一化 RBF 神经网络实现这种映射。

图 1 的运行流程为：状态 x_t 分别输入 Actor 与 Critic 部分，分别输出当前策略下的动作 a_t 与误差函数 E_t；误差 E_t 用于更新 Actor 与 Critic 中的参数；动作 a_t 作用于环境产生下一个状态变量 x_{t+1}。

2　基于连续动作强化学习的机动决策方法

2.1　NRBF 神经网络的值函数逼近

以 NRBF 神经网络作为 Actor 与 Critic 中的非线性逼近器，实现从状态到效用值与动作控制量的映射，基本结构如图 2 所示。无人机获取的状态变量输入后，网络输出为当前状态的效用值 $V(x)$。通过不断地修正网络中的参数，可以实现对真实效用值的逼近。

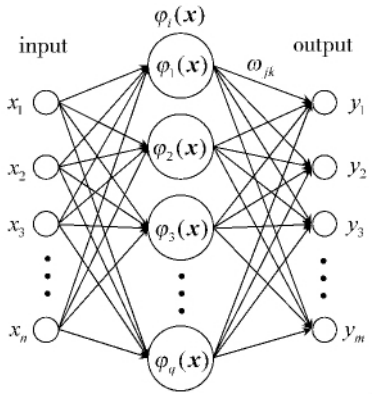

图 2　RBF 神经网络结构

用于实现非线性映射的基函数采用高斯函数，第 j 个 RBF 函数如式(7)所示。

$$\varphi_j(x)=\exp\left(-\sum_{i=1}^{n}\frac{\left\|x_i-c_{ij}\right\|^2}{2\sigma_{ij}^2}\right),i=1,2,\cdots,n\tag{7}$$

其中，x 为输入状态变量，c_{ij} 为第 i 个基函数的中心，σ_{ij} 为基函数中心的宽度。

输出层的线性映射表示为

$$y_k=\sum_{i=1}^{q}w_{ik}\varphi_j(x),k=1,2,\cdots,m\tag{8}$$

其中，w_{ik} 表示各个连接节点的权重，m 表示输出节点的个数。

NRBF 神经网络提高了网络的插值性能，降低了网络对高斯基函数中参数的敏感度[11]，这些特性均有助于提高网络的泛化性。将 RBF 神经网络的输出进行归一化处理，即得到 NRBF 神经网络，表示为

$$\phi_k(x)=\frac{\varphi_k(x)}{\sum_{i=1}^{q}\varphi_k(x)}\tag{9}$$

其中，$\phi_i(x)$ 为归一化基函数。

网络采用 TD 方法逼近真实值函数，则 TD 误差表示为

$$\delta_t=r_t+\gamma V(x_{t+1})-V(x_t)\tag{10}$$

其中，r_t 为瞬时奖赏，γ 为折扣率。

2.2　状态空间自适应调整方法

NRBF 神经网络中,隐层节点个数 q、中心 c_i、宽度 σ_i 与输出层的连接权重 w_{ij} 直接决定了网络的性能,对于本例中的无样本学习情况,并没有办法事先确定完整抽象空战态势特征的隐层节点个数,较好的方法是令网络起始节点数为零,在训练过程中自行判断是否需要添加新的节点,故为其设计自适应调整的优化方法。

2.2.1　结构学习

由于强化学习相当于采用无标签数据学习,对每一个输入 x_t,仅能得到瞬时奖赏 r_t,因此,采用 TD 误差 δ_t 作为增加节点的准则,表示为

$$|\delta_t|>\delta_{\min} \tag{11}$$

其中,$\delta_{\min}$ 为 TD 的误差阈值。

同时计算输入状态变量到每一个基函数中心的距离和 d,如果新输入向量满足

$$d>d_{\min}\ 且\ |\delta_t|>\delta_{\min} \tag{12}$$

则增加一个节点。该节点增加规则基于这样一个推论:如果新输入状态变量到每一个基函数中心的距离和越大,那么它与所有基函数的不相关性就越高,现有基函数对该状态变量的表达能力就越弱。

距离 d 的计算方法决定了节点增加方法的性能,通常采用的欧式距离公式的问题在于:该公式是距离的几何度量,而在空战态势评估中,两个相近的状态变量可能代表完全不同的态势,这里结合熵原理,以相对熵作为距离的评判标准。

定义 对输入状态变量 x_t,它与任意基函数中心的相对熵为

$$\begin{aligned}I(x_t,c_i)=\sum_{j=1}^{n}(x_{tj}\log\frac{x_{tj}}{1/2(x_{tj}+c_{ij})}+\\(1-x_{tj})\log\frac{1-x_{tj}}{1-1/2(x_{tj}+c_{ij})})\end{aligned} \tag{13}$$

则相对熵距离可表示为

$$d(x_t,c_i)=I(x_t,c_i)+I(c_i,x_t) \tag{14}$$

式(14)的意义在于保证 $d(x_t,c_i)=d(c_i,x_t)$。

当输入状态变量到其他所有基函数中心的相对熵距离和小于 $d_{\min}$,表明该状态变量包含的相对信息量较小,不应增加新的节点;反之,表明该状态变量包含的信息量较大,现有基函数的表达能力不足,需要增加新的节点。

对于新增基函数的中心,直接选取当前输入样本,新增基函数的宽度为

$$\sigma_{new}=\tau\min\|x_t-c_i\| \tag{15}$$

其中,τ 为重叠系数。

2.2.2　参数学习

基于 TD 误差,误差函数表示为

$$E(t)=\frac{1}{2}\delta_t^2 \tag{16}$$

由式(5),使用当前估计的值函数代替真实值函数,采用梯度下降法对权重向量 w 的误差求负导数,可以得到对参数 w 的更新规则为

$$w_{jk}=w_{jk}+\alpha_w\delta_t\phi_k(x_t) \tag{17}$$

其中,α_w 为对连接权值向量的学习率。同理分别对 c_i 与 σ_i 的误差求负导数,可得到基函数中心与宽度的更新公式为

$$c_{ij}=c_{ij}+\alpha_c\delta_t w_{jk}\phi_k(x_t)\frac{x_i-c_{ij}}{\sigma_{ij}^2} \tag{18}$$

$$\delta_{ij}=\delta_{ij}+\alpha_\delta\delta_t w_{jk}\phi_k(x_t)\frac{(x_i-c_{ij})^2}{\sigma_{ij}^3} \tag{19}$$

其中,α_c 与 α_δ 为对应参数的学习率。

2.3　无人机动作选择策略

通过网络输出的动作控制变量本质上是过去所有瞬时奖赏的折扣累积,因此,输出层的权重更新方法体现了对过去经验的“利用”,为了平衡网络对于未知动作的探索,有必要设计随机动作选择方法。

设对状态 x_t,网络输出动作值为 a_t,表示为

$$a_t=[a_{1t},a_{2t},a_{3t}]^{\mathrm{T}}$$

以纵向过载 n_x 为例,网络输出的 a_{1t} 并不能直接作为动作值,需要将其转换为符合无人机控制特性的 a_{1t}',设 $n_x\in[n_{x\min},n_{x\max}]$,有

$$a_{1t}'=\frac{1}{1+e^{-a_{1t}}}\times\Delta n_x-\frac{\Delta n_x}{2} \tag{20}$$

其中,$\Delta n_x=n_{x\max}-n_{x\min}$。

设当前状态下的输出为 $a_{1t}'=n_{xt}$,服从 $N(n_{xt},\sigma(k))$的高斯分布,k 表示网络运行次数,其标准差为 k 的函数,设

$$\sigma(k)=b_1e^{-(\frac{k}{b_2})^2} \tag{21}$$

其中,b_1,b_2 为参数。

受实际情况的约束,3 个控制变量均有固定的取值范围,将 a_{1t}' 的概率分布修改为

$$p(a_{1t}')=\frac{f(a_{1t}')}{\int_{a_{1\min}}^{a_{1\max}}f(a_{1t}')} \tag{22}$$

$$f(a_{1t}')=\frac{1}{\sqrt{2\pi}\sigma(k)}e^{-\frac{1}{2}\left(\frac{a_{1t}'-n_{xt}}{\sigma(k)}\right)^2} \tag{23}$$

其中,$a_{1t}'\in[n_{x\min},n_{x\max}]$。

对每一个 a_{1t}'，按 $p(a_{1t}')$ 的概率在 $[n_{x\min}, n_{x\max}]$ 内随机选取动作值，当参数确定时，选取动作值的概率分布仅由 k 决定。

令 $a_{1t}'=n_{xt}$，可得选择网络输出动作值的概率为

$$p=\frac{1}{[\Phi(\frac{n_{x\max}-n_{xt}}{\sigma(k)})-\Phi(\frac{n_{x\min}-n_{xt}}{\sigma(k)})]}$$

由式(21)，$\lim\limits_{k\to\infty}\sigma(k)=0^{+}$，对 $a_{1t}'\in[n_{x\min}, n_{x\max}]$，有 $\lim\limits_{k\to\infty}p=1$，即随着网络运行的次数增多，选择网络输出动作值的概率会趋向于 1。同理可得 a_{2t}' 与 a_{3t}' 的概率分布函数。

图 3 列出了在 $a_1'\in[-10,10]$，b_1=20，b_2=1 000，n_{xt}=0 情况下的动作值概率分布情况。

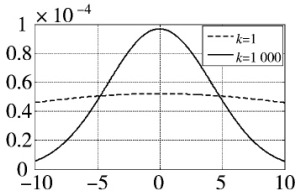

图 3　动作控制变量概率密度

可见在 k=1 时，概率分布曲线平缓，近似于均匀分布，表明在初次试探时，无人机更注重对未知动作的探索；当 k=1 000 时，概率分布集中于 0 附近，表明此时无人机更倾向于选择与网络实际输出相近的动作值；同时，如果网络的后期训练效果不够理想，可以通过人为减小 k 值以增强对未知动作的探索，且 k 值的改变并不直接作用于网络的参数更新过程，有效提高了网络的训练效果。

2.4　奖赏函数设计

对空战其中一方而言：当达成导弹发射条件时，获得最大奖赏；当态势有利时，获得一般奖赏；当态势不利时，获得负奖赏。依据该原则，设计奖赏函数如下。

2.4.1　角度奖赏

式(24)中：φ 为本机速度矢量与目标线的夹角；q 为目标速度矢量与目标线的夹角；$\varphi_{\max}$ 为导弹最大发射偏角。

$$r_A=\begin{cases}\dfrac{|\varphi-\varphi_{\max}|}{2\varphi_{\max}}\exp(\dfrac{q-\pi}{\pi}), & \varphi\leqslant\varphi_{\max}\\ -\dfrac{|\varphi-\varphi_{\max}|}{2\varphi_{\max}}\exp(\dfrac{q-\pi}{\pi}), & \varphi>\varphi_{\max}\end{cases}\tag{24}$$

2.4.2　速度奖赏

$$r_v=\exp(-\frac{V_r-V_0}{V_0})\tag{25}$$

其中，V_0 为最佳空战速度，表示为

$$V_0=V_b+(V_{\max}-V_b)(1-\exp(\frac{d_{\max}-d}{d_{\max}}))\tag{26}$$

其中，$V_{\max}$ 表示无人机最大飞行速度，$d_{\max}$ 表示雷达最大搜索距离。

2.4.3　距离奖赏

$$r_d=\begin{cases}1, & d\leqslant d_{\max}\\ \exp(-\dfrac{d_{\max}-d}{d_{\max}}), & d>d_{\max}\end{cases}\tag{27}$$

2.4.4　高度奖赏

$$r_h=\begin{cases}\exp(-\dfrac{\Delta h-\Delta h_0}{\Delta h_0}), & \Delta h\geqslant\Delta h_0\\ \exp(\dfrac{\Delta h-\Delta h_0}{\Delta h_0}), & \Delta h<\Delta h_0\end{cases}\tag{28}$$

其中，Δh_0 为最佳空战高度差。

2.4.5　达成发射条件奖赏

设计导弹发射条件为：$v_r\leqslant$1.8 Ma，$|\varphi|\leqslant\pi/3$，$\pi/3\leqslant q\leqslant5\pi/3$，$d\leqslant$10 000 m。当本机对目标达到上述态势时，变量 con=1(否则为 0)，获得奖赏

$$r_c=10$$

当敌机形成上述态势时，本机获得奖赏

$$r_c=-10$$

2.4.6　环境奖赏

设置环境奖赏的主要目的是使无人机能够较好地适应环境。实验发现：在网络训练的初期，网络动作选择策略的探索主要是为了适应飞行环境，避免危险的飞行动作，如：飞行高度低于最低安全高度或飞行速度超过最大允许速度。此过程耗时较长，且此时网络对空战态势奖赏不敏感，对提升无人机机动决策能力的贡献较小，因此，为了提升网络训练效率，环境奖赏的力度要远大于其他奖赏，促使网络迅速生成符合环境约束的动作策略。

设环境奖赏为 r_e，表示为

$$r_e=\begin{cases}-10, h_r<200\\ -10, h_r>20\ 000\\ -10, v_r>300\\ -10, v_r<50\\ 0, \text{else}\end{cases}\tag{29}$$

综上所述，总奖赏函数表示为

$$r=\begin{cases}r_A+r_v+r_d+r_h+r_e, & \text{con=0}\\ r_c+r_e, & \text{con=1}\end{cases}\tag{30}$$

2.5　机动决策方法结构

基于以上的分析，机动决策强模型如下页图 4 所示。

输出为 3 个动作控制量与当前状态的效用值

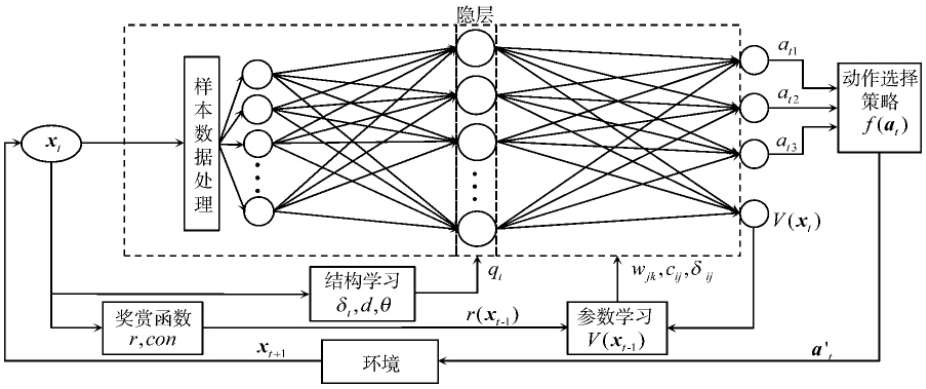

图 4　基于强化学习的机动决策方法

$V(x_t)$,参数学习模块通过结合存储的 $V(x_{t-1})$对隐层与输出层进行参数更新;动作控制量首先经动作选择策略模块选择要采取的动作值, 再作用于环境,产生下一个状态变量 x_{t+1}。

算法流程为:

Step 1:数据初始化;

Step 2:读取输入变量并处理生成 x_t,产生 3 个动作控制量与对当前状态的估计效用值 $V(x_t)$;

Step 3:计算瞬时奖赏 r_t,并按式(11)与式(14)计算输入状态变量到每一个隐层节点的相对熵距离与 TD 误差 δ_t,如果满足式(12),则增加一个节点,否则节点数不变;

Step 3:动作选择策略对网络输出动作控制量进行处理,生成作用于环境的动作控制量;

Step 4:结合缓存的 $V(x_{t-1})$,按式(17)~ 式(19)更新网络参数;

Step 5:判断是否达到最大学习步数,是则结束;否则将动作控制量作用于环境,生成 x_{t+1},转到 Step 2。

3　仿真实验

3.1　空战对抗实验

为了加快计算速度,仿真设置敌方无人机按固定航迹飞行,起始点为(0,0,0),匀速 200 m/s,共有直线飞行、蛇形机动、盘旋飞行 3 个情形,其状态参数事先以数组形式存储,按时间输出作为环境的反馈信号。3 种情形下的我方 UCAV 参数设置如表 1 所示。

这里选取了训练后期的 3 条飞行轨迹与最终收敛的飞行轨迹,并给出了不同训练次数下法向过载控制量的变化曲线,结果如图 5~ 图 13 所示。

其中蓝色轨迹为敌机飞行轨迹。

图 5~ 图 13 表明:前期策略学习侧重于对环境的适应,由于本例中环境奖赏力度较大,导致起始累积奖赏较低;训练进行到 3 000 次时,3 种情形下

表 1　仿真参数设置

	情形 1	情形 2	情形 3
起始点	(-1 000,-2 000,2 000)	(1 000,0,2 000)	(2 000,0,2 000)
速度(m/s)	300	300	300
攻角(rad)	0	0	0
滚转角(rad)	0	0	0
学习率	0.8	0.8	0.8
奖赏折扣	0.95	0.95	0.90
$d_{\min}$	0.5	0.5	0.5
$\delta_{\min}$	0.01	0.01	0.01

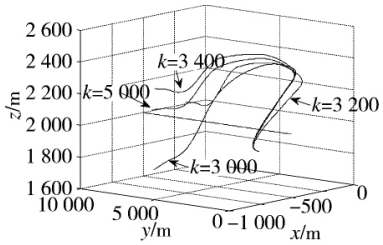

图 5　敌机直线飞行的训练轨迹

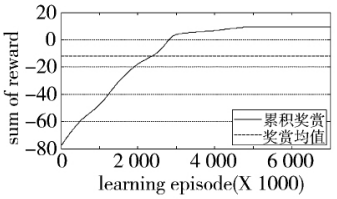

图 6　累积奖赏变化曲线

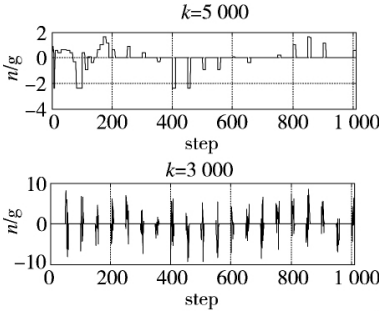

图 7　法向过载控制量变化曲线

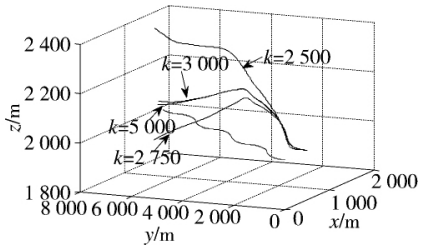

图 8　敌机蛇形机动的训练轨迹

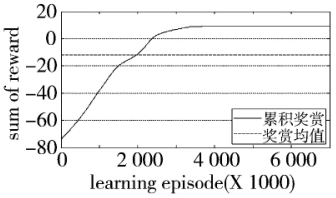

图 9　累积奖赏变化曲线 1

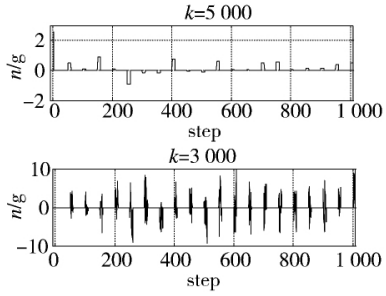

图 10　法向过载控制量变化曲线 1

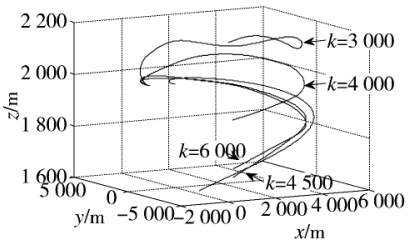

图 11　敌机盘旋飞行的训练轨迹

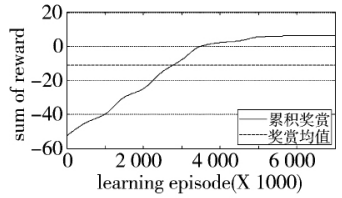

图 12　累积奖赏变化曲线 2

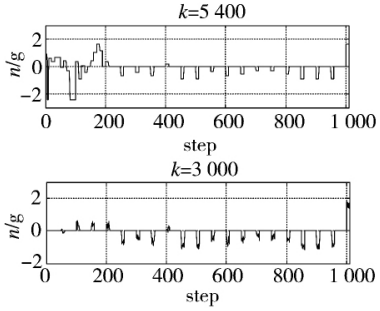

图 13　法向过载控制量变化曲线 2

的无人机动作选择策略均可以较好地适应飞行环境的约束，未出现早期训练过程中超出范围限制的危险飞行动作。此时无人机已经能够对敌机的机动动作做出一定的反应，但整体航迹上机动决策的质量并不高，末端仍然存在错过敌机的现象。控制量曲线的变化速率较快，表明此时的动作选择策略中仍然包含了较大的随机因素；继续进行训练，无人机与敌机航迹末端的距离明显减小，基本能够分辨出无人机要采取的机动动作，累积奖赏值趋于收敛，表明动作选择策略已接近最优；最终，无人机的动作选择策略会使航迹收敛，动作值中随机因素基本消失，控制量曲线趋于平滑。

以上 3 个情形的仿真表明：无人机的动作选择策略在经训练后能够依据态势输出合理的连续动作控制量，具备空战能力。

3.2　动作选择策略性能分析

采用文献[12]中的基本机动动作集实现情形 3 中的无人机飞行轨迹。主要采取的动作与控制量如表 2 所示。

表 2　基本机动动作操纵指令

操纵指令	n_x	n_z	γ_x
匀速前飞	0	1	0
俯冲	0	$-n_{z\max}$	0
右俯冲	0	$-n_{z\max}$	$\pi/4$
右转弯	0	$n_{z\max}$	$-\arccos(1/n_{z\max})$
左转弯	0	$n_{z\max}$	$-\arccos(1/n_{z\max})$

按相同的仿真步长对无人机分别施加控制量，得到控制量 n_z 与 γ_x 的变化曲线如图 14、图 15 所示。

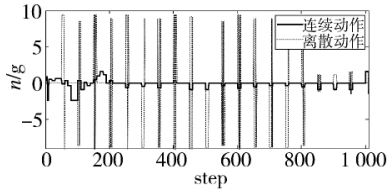

图 14　法向过载控制量变化曲线 3

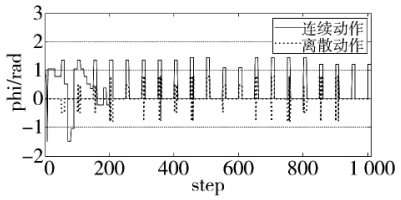

图 15　滚转角控制量变化曲线

图 14、图 15 表明：由于法向过载 n_z 的取值仅为 1，$n_{z\max}$ 或 $-n_{z\max}$，采用基本机动动作集在每一步长上只能施加固定的控制量，而情形 3 中无人机盘旋过

程中的法向过载保持在 2 g 以内，无论起始采取右转弯或右俯冲均使控制量产生严重的跳变，如果考虑控制系统的延迟特性，还会进一步降低控制效果；而连续动作情况下产生的控制量连续性较强，整体上控制量分布较均匀。因此，采用连续动作控制量更有利于无人机的控制。

3.3　自适应节点构建策略性能分析

3.1 节中情形 3 下的奖赏值变化曲线如图 16 所示，其中奖赏 1 与均值 1 曲线为采用相对熵距离的奖赏与均值；奖赏 2 与均值 2 曲线为采用欧式距离的奖赏与均值。图 17 记录了第一次训练过程中 1 010 个仿真步长（40 s）内 RBF 神经网络隐层节点数目的变化情况，可以看出：在训练开始阶段，采用相对熵距离公式与欧式距离公式的网络节点增长速率相同，表明在训练开始阶段，由于网络隐层节点数少，对新状态变量的表达能力较弱，因此，此时节点增加速度较快；在训练后期，两者产生了显著的差异，采用相对熵距离的网络隐层节点数最终为 616 个，而采用欧式距离的隐层节点数为 671 个，通过观察图 16 中效用值变化曲线，采用相对熵距离与欧式距离计算方法的效用值的均值分别为 −11.113 2 与 −11.925 9，最终效用值保持在 6.387 0 与 6.106 3，表明两者的动作选择策略性能相近。基于相对熵距离的网络用较少的隐层节点数，达到了与采用欧式距离的网络相同的训练效果，证明基于相对熵距离的自适应节点构建方法更加高效。

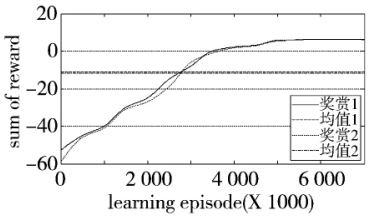

图 16　奖赏值变化曲线

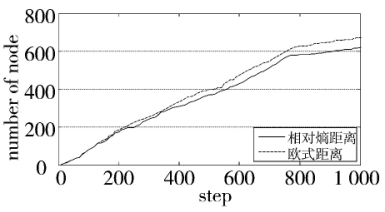

图 17　隐层节点数量变化曲线

4　结论

本文设计了基于连续动作控制变量的无人机机动决策方法，采用共用隐层的 NRBF 神经网络结构分别逼近效用值与动作控制变量，提出了基于相对熵距离的神经网络隐层节点自适应构建方法，并设计了基于高斯分布的随机动作选择策略。仿真实验结果表明，所设计的算法能够满足无人机机动决策的需求，相对于采用有限离散动作集的机动决策算法，该方法产生的航迹更为平滑，无人机的机动也更加灵活；同时，该自适应节点构建策略有效降低了 RBF 神经网络的隐层节点数，提高了算法运行效率。

参考文献：

[1] 国海峰，侯满义，张庆杰，等. 基于统计学原理的无人作战飞机鲁棒机动决策[J].兵工学报，2017，38(1)：160-167.

[2] 李俊涛，毛红保，张鹏，等. 基于多优化策略 RRT 的无人机实时航线规划 [J]. 火力与指挥控制，2017，42(12)：115-119.

[3] 马耀飞，龚光红，彭晓源. 基于强化学习的航空兵认知行为模型 [J]. 北京航空航天大学学报，2010，36(4)：379-383.

[4] 朱圆恒，赵冬斌. 概率近似正确的强化学习算法解决连续状态空间控制问题 [J]. 控制理论与应用，2016，33(12)：1603-1613.

[5] 刘慧霞，马丽娜，李大健，等. 无人机多机协同侦察系统关键技术[J]. 火力与指挥控制，2017，42(12)：1-4.

[6] LIN C T，JOU C P，LIN C J. GA- based reinforcement learning for neural networks [J]. International Journal of Systems Science，1998，29(3)：233-247.

[7] POTJANS W，MORRISON A，DIESMANN M. A spiking neural network model of an actor-critic learning agent[J]. Neural Computation，2009，21(2)：301-339.

[8] ROGER W S，ALAN E B. Neural network models of air combat maneuvering [D]. New Mexico：New Mexico State University，1992.

[9] 周思羽，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.

[10] SUTTON R S，BARTO A G. Reinforcement learning：an introduction[M]. Cambridge MA：MIT Press，1998.

[11] 刘全，肖飞，傅启明，等. 基于自适应归一化 RBF 网络的 Q-V 值函数[J]. 计算机学报，2015，38(7)：1386-1395.

[12] BREITNER M H，PESCH H J，GRIMM W. Complex differential games of pursuit-evasion type with state constraints，part 2：numerical computation of optimal open-loop strategies [J]. Journal of Optimization Theory and Applications，1993，78(3)：443-463.