http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2018.0132

基于多目标优化与强化学习的空战机动决策

杜海文[1,*]，崔明朗[1]，韩统[1]，魏政磊[1]，唐传林[2]，田野[3]

(1. 空军工程大学 航空工程学院，西安 710038； 2. 94782 部队，杭州 310004；
3. 福州大学 物理与信息工程学院，福州 350108)

摘 要：为了解决无人机自主空战中的机动决策问题，提出了一种将优化思想与机器学习相结合的机动决策模型。采用多目标优化方法作为决策模型核心，既解决了传统优化方法需要为多个优化目标设置权重的困难，又提高了决策模型的可拓展性；同时在多目标优化的基础上通过强化学习方法训练评价网络进行辅助决策，解决了决策模型在对抗时博弈性不足的缺点。为了测试决策模型的性能，以近距空战为背景，设计了 3 组仿真实验分别验证多目标优化方法的可行性、辅助决策网络的有效性以及决策模型的总体性能，仿真结果表明，决策模型可以对有机动的敌机进行有效的实时机动对抗。

关 键 词：自主空战；机动决策；多目标优化；强化学习；神经网络
中图分类号：V279
文献标识码：A **文章编号**：1001-5965(2018)11-2247-10

随着无人机技术的不断发展，无人作战飞行器(UCAV)的作用与地位也在不断升高，在战场上的意义越来越重要[1]；由于不必考虑人身体条件限制，UCAV 可以完全发挥出飞行器的性能，做出有人机难以做出的大过载机动，可以预见 UCAV 必将成为未来空中战场的主角。而要实现高强度的空中对抗，UCAV 必须脱离地面控制，具备自主空战的能力，本文结合传统优化模型以及机器学习方法，建立了基于多目标优化的机动决策模型，用于解决 UCAV 自主空战时的机动决策问题。

关于空战机动决策问题有很多研究成果(包含有人机与无人机)，总的来说大致可以分为 3 类：①基于各类基本战术动作库的机动决策，文献[2]最早对建立机动动作库进行了系统的研究和总结，文献[3-4]分别就机动动作库的设计、控制应用以及基于动作库的机动动作识别等问题进行了研究，详细阐述了基于动作库的机动决策中存在的各类问题。②基于优化方法的机动决策，该类方法的共同点在于通过各类态势评估方法将机动决策问题转化为标准的优化模型，文献[5-6]基于各类不同的智能算法来求解优化模型，文献[7]基于各类态势分析方法建立了隐马尔可夫模型，并使用维比特算法进行求解。③基于机器学习方法的机动决策，机器学习方法研究在近年得到了极大的发展，采用各类机器学习方法研究机动决策也越来越多，文献[8]应用深度置信网络来进行态势评估，文献[9]采用了强化学习方法研究空战智能决策。

然而，以上方法在处理无人机空战机动决策时都存在一些弊端：机器学习方法在处理类似对抗博弈问题时效果很好，但不同于有人机的空战决策，无人机空战基本不存在有学习价值的样本；而各种基于动作库的方法虽然是建立在大量空战

收稿日期：2018-03-15；录用日期：2018-06-15；网络出版时间：2018-07-27 18:17
网络出版地址：kns.cnki.net/kcms/detail/11.2625.V.20180727.1038.001.html
基金项目：国家自然科学基金(61601505)；陕西省自然科学基金(2017JM6078)
* 通信作者. E-mail：18191856512@163.com

引用格式：杜海文，崔明朗，韩统，等. 基于多目标优化与强化学习的空战机动决策[J]. 北京航空航天大学学报，2018，44(11)：2247-2256. DU H W，CUI M L，HAN T，et al. Maneuvering decision in air combat based on multi-objective optimization and reinforcement learning[J]. Journal of Beijing University of Aeronautics and Astronautics，2018，44(11)：2247-2256(in Chinese).

经验之上，但灵活性较差，且现在的空战经验都是有人机的经验，无法确定其用在无人机上是否可靠。相比之下，传统的优化方法原理是基于对态势分析的寻优，反而可以根据不同的飞行器性能和空战环境得出实时性与灵活度都较强的决策，但是传统优化方法在整合不同态势参数时缺少严谨的方法，且其决策结果随着模型的确立就已经确定下来，无法体现出对抗博弈的思想。基于上述分析，本文依然使用优化模型作为决策的核心思想，采用多目标优化方法取代单目标优化，并通过强化学习方法建立辅助决策网络，建立了具备实时对抗性的无人机空战机动决策模型。

1 机动决策模型

1.1 UCAV 运动模型

在对 UCAV 近距空战进行机动决策与仿真时，采用三自由度质点模型描述 UCAV 的运动状态，模型参数定义如图 1 所示。

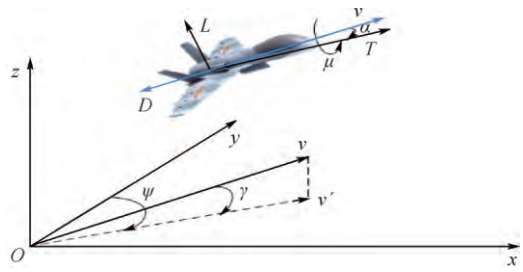

图 1 UCAV 三自由度质点模型

Fig. 1 UCAV three-degree-of-freedom particle model

1.1.1 模型假设

对建立 UCAV 运动、动力学模型作如下假设：

1) 假设 UCAV 为一个刚体。

2) 假设地球为惯性坐标系(将地面坐标系看作惯性坐标，忽略地球自转及公转影响)。

3) 忽略地球曲率。

1.1.2 UCAV 质点模型

在考察 UCAV 运动时，将 UCAV 视为质点。在惯性坐标系下，其质点运动方程为

$$\begin{cases} \dot{x} = v_{\mathrm{u}}\cos\gamma\cos\psi \\ \dot{y} = v_{\mathrm{u}}\cos\gamma\sin\psi \\ \dot{z} = v_{\mathrm{u}}\sin\gamma \end{cases} \tag{1}$$

相同惯性坐标系下，UCAV 的质点动力学方程为

$$\begin{cases} \dot{v}_{\mathrm{u}} = \dfrac{T\cos\alpha - D}{m} - g\sin\gamma \\ \dot{\gamma} = \dfrac{(L + T\sin\alpha)\cos\mu}{mv_{\mathrm{u}}} - \dfrac{g}{v_{\mathrm{u}}}\cos\gamma \\ \dot{\psi} = \dfrac{(L + T\sin\alpha)\sin\mu}{mv_{\mathrm{u}}\cos\gamma} \end{cases} \tag{2}$$

式中：v_{u} 为速度；γ 为航迹倾角；ψ 为航向角；μ 为滚转角；α 为迎角；m 为质量；T 为发动机推力；D 为空气阻力；L 为升力；g 为重力加速度。

飞行过程中，UCAV 所受升力 L 与空气阻力 D 计算公式如下：

$$\begin{cases} L = \dfrac{1}{2}\rho v_{\mathrm{u}}^2 S C_L \\ D = \dfrac{1}{2}\rho v_{\mathrm{u}}^2 S C_D \end{cases} \tag{3}$$

式中：ρ 为空气密度；S 为 UCAV 参考横截面积；C_L 和 C_D 分别为升力和阻力系数。

UCAV 发动机推力 T 计算公式如下：

$$T = \delta T_{\max} \tag{4}$$

式中：$T_{\max}$ 为发动机最大推力；δ 为油门，取值范围为 $[0,1]$。

在控制量的选择上，仿照有人机中飞行员的驾驶方式，采用迎角 α、油门 δ、滚转角 μ 三个控制量来控制 UCAV 进行机动。

1.2 多目标优化方法

基于优化方法的机动决策模型具有较高的决策效率与良好的实时性，但在寻优过程中需要对多个目标参数进行合并，这样的合并过程往往使用层次分析法、专家打分法等主观性较强的方法来确定权值，缺少严格的证明过程，其决策结果难以使人信服。

事实上，在不同的空战环境下，对于各个态势参数的需求程度也是不同的，所以将不同态势参数加权求和后进行优化的方法本身就具有很大的局限性。为了避免这种局限性，本文结合多目标优化思想，建立了多目标优化机动决策模型。

1.2.1 多目标优化思想

首先简要介绍一些多目标优化问题中的概念，在多目标优化中，采用 Pareto 支配[10]关系来判断解的优劣程度，Pareto 支配关系的定义如下。

定义 1 对于可行域内任意 2 个解 $\boldsymbol{x}_1$ 与 $\boldsymbol{x}_2$，假设在最小化问题 $f(f_1,f_2,\cdots,f_k)$ 中，当且仅当式(5)成立时称 $\boldsymbol{x}_1$ 对 $\boldsymbol{x}_2$ 形成 Pareto 支配：

$$[\forall i \in \{1,2,\cdots,k\}, f_i(\boldsymbol{x}_1) \geqslant f_i(\boldsymbol{x}_2)] \cap [\exists i \in \{1,2,\cdots,k\}, f_i(\boldsymbol{x}_1) \geqslant f_i(\boldsymbol{x}_2)] \tag{5}$$

$\boldsymbol{x}_1$ 支配 $\boldsymbol{x}_2$ 表示解 $\boldsymbol{x}_1$ 优于解 $\boldsymbol{x}_2$，一般记作

$\boldsymbol{x}_1 \succ \boldsymbol{x}_2$。

由定义 1 可知，求解多目标优化问题的本质就是在全部可行解中找到所有不被任何一个其他可行解所支配的解的集合。将这个集合称之为多目标优化问题的 Pareto 边界，具体定义如下。

定义 2　设多目标优化问题 f 的可行解集为 X，则其 Pareto 边界为

$$P_f = \{\boldsymbol{x} \in X \mid \nexists \boldsymbol{x}_i \in X, \boldsymbol{x}_i \succ \boldsymbol{x}\} \tag{6}$$

多目标优化的目的就是求出优化问题的 Pareto 边界。

1.2.2　优化目标

使用优化模型必然需要构建优化目标参数，采用速度、高度、距离、角度[11-12] 4 个量作为优化目标是最为常用的方法之一，但这些量的具体战术意义还不够明确，本文将基于空战实际将这些参量进行耦合后提出了如下优化目标参数。

1）基于武器攻击区的威胁参数

空战的最终目的就是击落敌方与保护己方，进行机动也正是为了使己方构成武器发射条件和避免使对方构成武器发射条件，故本文基于机载武器攻击区的概念，结合与之相关的角度、距离等常规评估参数，提出了一种新的威胁参数 η_{A} 作为一个优化目标，参数模型以双方携带弹药类型为基础，具体定义如下。

① 常规条件下。制导武器一般以空空导弹为主，现在的空空导弹的攻击区大致如图 2 所示。

假设图 2 中攻击区为我机携带的第 i 枚导弹的攻击区，则该型导弹对敌机威胁参数为

$$\eta_{ai} = \begin{cases} 1 & R \leqslant R_{\mathrm{g}} \\ 0 & R_{\mathrm{g}} = 0 \\ \mathrm{e}^{-\frac{(R-R_{\mathrm{g}})^2}{R_{\mathrm{g}}^2}} & R > R_{\mathrm{g}} > 0 \end{cases} \tag{7}$$

式中：R_{g} 为该导弹沿视线角 α_{u} 方向上的最远攻击距离，由于部分导弹不具备全向打击能力，故若在当前 α_{u} 下 R_{g} 为 0，则定义此刻 $\eta_{ai}=0$。

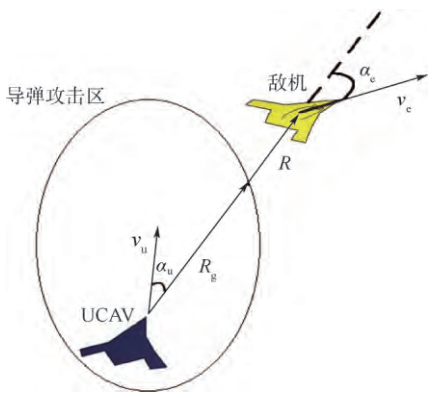

图 2　态势参数定义

Fig. 2　Definition of situation parameters

根据上述方法计算出我机携带所有空空导弹对敌机威胁参数（$\eta_{\mathrm{a1}}, \eta_{\mathrm{a2}}, \cdots, \eta_{an}$）后，取其中的最大值 η_{amax}，即为我机当前对敌机的威胁参数 η_{a}；采用相同方法计算出敌机对我机威胁参数后，取两者之差即为总威胁参数值 η_{A}：

$$\eta_{\mathrm{A}} = \eta_{\mathrm{a(ucav)}} - \eta_{\mathrm{a(enemy)}} \tag{8}$$

② 仅使用非制导武器时。空对空作战非制导武器一般指航炮，由于航炮的发射条件比较苛刻，一般只在形成尾追时才能构成发射条件，故直接使用双方角度参数与距离参数进行耦合来定义其态势参数：

$$\eta_{\mathrm{A}} = \begin{cases} \dfrac{\pi - \alpha_{\mathrm{u}} - \alpha_{\mathrm{e}}}{\pi} & R \leqslant R_{\mathrm{a}} \\ \dfrac{\pi - \alpha_{\mathrm{u}} - \alpha_{\mathrm{e}}}{\pi} \cdot \mathrm{e}^{-\frac{(R-R_{\mathrm{a}})^2}{R_{\mathrm{a}}^2}} & R > R_{\mathrm{a}} \end{cases} \tag{9}$$

式中：R_{a} 为航炮射程。

2）能量参数

能量理论[13] 是近期提出的一种空战机动理论，该理论的核心在于：在空战中首先寻求获得能量上的优势，然后将能量优势转化为态势上的优势。能量理论随着飞机性能的提升愈发受到重视，现在的飞机性能可以支持完成各种大过载机动、过失速机动等非常规动作，这使得飞机可以有更多方式扭转不利的态势。即使在常规的机动对抗中，能量也是一个不可忽略的条件，因为所有机动动作都是以消耗能量为前提，高能量就意味着更多的机会与选择。故本文设置能量参数 η_{W} 作为一个优化目标，计算公式如下：

$$\begin{cases} \eta_{\mathrm{W}} = \Delta W / W_{\mathrm{st}} \\ \Delta W = \Delta W_p + \Delta W_k \\ W_{\mathrm{st}} = 10\,000\, m_{\mathrm{u}} \end{cases} \tag{10}$$

式中：W_p 和 W_k 分别为重力势能与动能；W_{st} 为能量标准化参数；m_{u} 为我方 UCAV 质量。

1.2.3　多目标优化机动决策模型

决策模型结构如图 3 所示。

目前有很多种多目标算法可供使用，由于上述模型复杂度不高且机动决策对实时性有较高要求，考虑到灰狼算法在处理维数较低问题时收敛速度快，本文在仿真时采用多目标灰狼算法（MOGWO）[14-15]。

事实上，多目标优化模型具有良好的可拓展性，在实际应用时，可以根据实际空战环境在以上 2 种优化目标的基础上添加其他新的优化目标（如雷达性能、电子战等），添加时只需将新的目标参数模型加入原优化目标集即可，不需要对决

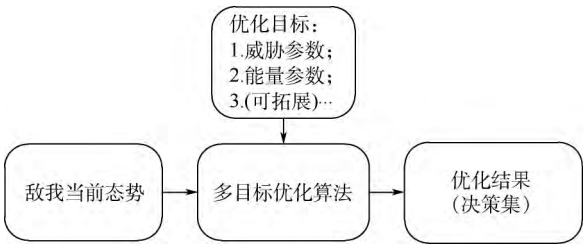

图 3 多目标优化机动决策模型结构

Fig. 3 Structure of multi-objective optimization model for maneuver decision

策模型中的其他结构进行任何变化。

1.3 基于强化学习方法的辅助决策

1.2 节提出了多目标优化思想并建立了优化参数的模型,但多目标优化模型仍存在以下缺点:

1) 多目标优化的结果是一个决策集,并没有给出从决策集中的选择具体决策的方法,如果不采用其他辅助决策方法,则只能从决策集中随机选取决策。

2) 多目标优化的本质依然是优化模型,未体现出空战博弈的思想。

强化学习方法[16]在处理类似的对抗博弈决策中取得了很大成果,但由于无人机空战决策问题的复杂度太高而难以实现。然而,如果在多目标优化的基础上进行强化学习,强化学习任务的探索空间将大大减少,故本文以多目标优化为决策基础,使用强化学习方法训练评价网络,用于对决策集中的决策进行评价与选择,从而解决上述 2 点不足。

1.3.1 蒙特卡罗强化学习

强化学习任务通常用马尔可夫决策过程(Markov Decision Process,MDP)来描述,任务对应了四元组 $E = <\mathrm{St},\mathrm{Ac},\mathrm{Pr},\mathrm{Re}>$,St 为状态空间,是机器所感知到的环境描述的集合,Ac 为系统能够采取的行为的集合,Pr 指定了不同环境下采用各种行为的概率,用以在探索过程中为行为选择提供参考,Re 指定了奖赏,通过反馈来调整 Pr 中的概率。

虽然对于空战中态势评估的研究已较为完善,但考虑到空战过程中机动动作往往是一系列的连续动作,即在完整的机动决策中,并非每一时刻都是为了追求最优态势。所以要设置符合要求的奖赏函数并不容易,而蒙特卡罗方法可以解决这个问题。

蒙特卡罗强化学习[17]的思路是采用多次"采样"求平均奖赏的方式来近似对行为的评价,即系统从起始状态下开始探索环境直至结束,将整个过程的奖赏作为过程中经历的每一个状态 st 的一次累积奖赏,在多次采样后,对每一个状态 st 的累积奖赏取均值得到其奖赏值 re。

就效率而言,蒙特卡罗强化学习比其他强化学习方法相去甚远,在实践中对蒙特卡罗方法的应用也不是很广泛,但本文模型的决策核心还是多目标优化,强化学习任务只需对多目标优化的决策结果进行评价与选择,即强化学习的行为空间 A 为一个已经经过筛选的较小空间,故收敛速度必然大大提升,从而使蒙特卡罗方法具备了可行性。

1.3.2 基于神经网络的值函数近似

初始的强化学习方法都是针对离散的状态-动作空间来进行的,但对于空战而言,其状态空间与动作空间都是连续的高维空间,进行离散化处理显然不是合适的方法。在类似的高维连续空间强化学习中,往往采用值函数近似的方法来进行连续空间的强化学习。

值函数近似[18]指的是通过一个函数 ϕ 建立从状态 St 到状态奖赏值的映射: ϕ: St→Re。考虑到空战决策问题的复杂性,最终的近似函数必然是复杂非线性函数,而神经网络在拟合复杂非线性函数时具备较好的性能,Hornik 等[19]在 1989 年就证明了只需一个隐层的 BP 神经网络可以逼近任何闭区间的连续函数,故本文将训练一个三层的 BP 神经网络来拟合值函数,用以对多目标决策集进行评价。隐层节点数将依照以下经验公式进行设计:

$$l_{\mathrm{no}} = \sqrt{n_{\mathrm{no}} + m_{\mathrm{no}}} + a_{\mathrm{no}} \tag{11}$$

式中: l_{no}为隐层节点数; n_{no}和 m_{no}分别为输入和输出节点数; a_{no}为 1 ~ 10 之间的调节常数。

结合对辅助决策网络功能的需求,网络具体设置如下:

1) 将空战态势(即状态量)作为输入层,利用 si{ $R,\alpha_{\mathrm{u}},\alpha_{\mathrm{e}},\Delta h,\Delta v$} 5 个参数来描述空战态势,即网络输入层节点数为 5, Δh 为两机高度差, Δv 为速度差。

2) 网络输出为对输入态势下我机获胜期望的预测值 v([0,1] 之间的数, v 值越大代表获胜期望越大),输出层节点数为 1。

3) 神经网络训练采用 LM(Levenberg-Marquardt)方法,其中 BP 误差计算类似于时序差分(TD)方法[20]误差计算公式,但由于模型采用蒙特卡罗方法,只能使用每次仿真结果作为本次仿真所经历状态的统一奖赏值,具体计算公式如下:

$$\begin{cases} \Delta V(\boldsymbol{x}_i) = \alpha_{\mathrm{RL}}(r + \gamma_{\mathrm{RL}} V(\boldsymbol{x}_{i+1}) - V(\boldsymbol{x}_i)) \\ r = (r_{\mathrm{end}} - V(\boldsymbol{x}_i))/(n - i) \end{cases} \tag{12}$$

式中：α_{RL}为学习率，一般根据训练次数确定；γ_{RL}为折扣率，本文取 $\gamma_{RL}=0.4$；r 为奖赏值，r 值由仿真结果 r_{end}给出，r_{end}取值为 0、0.5 或 1（对应失败、平局或胜利）；n 为本次仿真经历的总步数；i 为当前状态步数。

4）结合隐层节点数经验公式，通过实际仿真效果，选择隐层节点数为 12。

1.3.3　辅助决策模型

辅助决策网络的强化学习模型训练步骤如下：

步骤 1　初始化辅助决策网络。

步骤 2　随机产生敌我双方初始位置状态，开始仿真模拟。

步骤 3　记录下当前敌我态势关系 si_i，由多目标决策模型得出决策集（敌机可以采用与我机相同策略进行机动，或根据实际需求预先设置其轨迹）。

步骤 4　预测每种决策后敌我态势关系，进而通过辅助决策网络得出对应的获胜期望$\{\nu_1,\nu_2,\cdots,\nu_n\}$。

步骤 5　从决策集中随机选取出最终执行的决策，每种决策的被选取概率为

$$P_i=\frac{\nu_i}{\sum_{j=1}^{n}\nu_j} \tag{13}$$

步骤 6　执行决策后判断是否达到空战结束条件，若未达到，返回步骤 3；若已达到，进入步骤 7。

步骤 7　对本次仿真所经历的所有状态 si，通过式（12）计算 BP 误差返回辅助决策网络用于网络更新。

步骤 8　判断是否达到最大训练次数，若未达到，返回步骤 2。

注意：训练过程中，若敌机采用相同的决策模型，则双方数据均可通过步骤 7 中的网络更新；若敌机采用预先设置好的其他机动方法，则只有我方数据可用于网络更新。

1.4　机动决策模型整体框架

结合 1.2 节和 1.3 节描述的多目标决策模型与辅助决策网络，机动决策模型整体框架如图 4 所示。

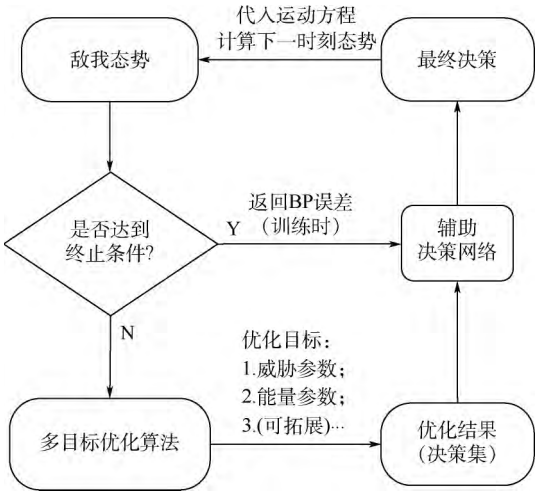

图 4　机动决策模型结构

Fig. 4　Structure of maneuvering decision model

2　仿真实验

由于本文机动决策涉及模型较多，故在仿真时将针对各个模型予以验证，仿真环境及参数设置如下。

仿真时，敌我飞行器采用相同的参数，质量 $m=14\,680$ kg，参考截面积 $S=49.24$ m^2，高度限制为 $h\in[1,12]$ km，速度限制为 $v\in[80,400]$ m/s，迎角限制为 $\alpha\in[-10°,30°]$；发动机采用 F-4 涡喷发动机数据[21]，其最大推力采用式（14）拟合：

$$T_{max}=\begin{bmatrix}1\\ \bar{v}_u\\ \bar{v}_u^2\\ \bar{v}_u^3\\ \bar{v}_u^4\end{bmatrix}^{T}\begin{bmatrix}30.21 & -0.668 & -6.877 & 1.951 & -0.1512\\ -33.80 & 3.347 & 18.13 & -5.865 & 0.4757\\ 100.80 & -77.56 & 5.441 & 2.864 & -0.3355\\ -78.99 & 101.40 & -30.28 & 3.236 & 0.1089\\ 18.74 & -31.60 & 12.04 & -1.785 & 0.09417\end{bmatrix}\begin{bmatrix}1\\ h\\ h^2\\ h^3\\ h^4\end{bmatrix} \tag{14}$$

式中：$\bar{v}_u$ 为 UCAV 飞行马赫数；h 为飞行高度，单位为 10 000 ft（即 3 048 m）；T_{max}的单位为 1 000 lb（即 4 436.26 N）。

升力系数和阻力系数采用式（15）拟合[22]：

$$\begin{cases}C_L=(-0.0434+0.1369\alpha)\sin\alpha+(0.131+3.0825\alpha)\cos\alpha\\ C_D=(0.0434-0.1369\alpha)\cos\alpha+(0.131+3.0825\alpha)\sin\alpha\end{cases} \tag{15}$$

考虑到本文未针对探测能力设置优化函数，故训练时设置双方均只使用航炮进行近距空战(近距格斗时电子战作用较小，但对机动决策模型有较高要求)，所有仿真中决策步长为 1 s；训练过程中，判定相互脱离距离为 15 km；攻击条件设置为 $\alpha_u \in [-20°, 20°]$ 且 $R < 2.5$ km(参数定义见图 2)，满足攻击条件 3 s 视为进行有效攻击；任意一方进行有效攻击或双方脱离则仿真结束。

仿真实验在 Matlab 2013a 下进行，运行环境为 Inter(R) Core(TM) i5-2310 处理器，3.40 GB 内存。

2.1 多目标优化可行性验证

1) 时间可行性

由于辅助网络在决策时的耗时远小于多目标优化，故首先验证多目标优化方法的实时性，本文采用 MOGWO 作为求解模型的算法，随机产生 100 组敌我态势并使用算法寻优，仿真时灰狼种群与外部种群数均设置为 30，迭代次数为 3 次。

采用 MATLAB 自带的计时功能记录了 100 次决策时间，决策平均时长 $t = 0.286\,541$ s，远小于决策步长 1 s，故决策模型具有良好的实时性。

为了展示寻优效果，图 5 记录了上述实验过程中的一次寻优的结果，其中红点为算法寻优结果，蓝点为可行域的大致范围(通过穷举法得出)，可以看出 MOGWO 可以在上述条件下找到基本完整、均匀的 Pareto 边界。

2) 决策可行性

验证通过多目标优化方法决策集的可行性，仿真时随机产生 100 组初始态势，敌机按初始态势做匀速直线运动，我机在不使用辅助网络的情况下进行机动，即决策时从多目标优化的决策集中按等概率随机选取最终决策，每组仿真模拟 75 s 的空战情形(若在 75 s 内达到结束条件则提前结束仿真)。

图 6 记录了 100 组仿真中我方优化目标函数在每秒的平均值(提前结束的组自结束起至 75 s 的目标函数值均按结束时的目标函数值记录)。

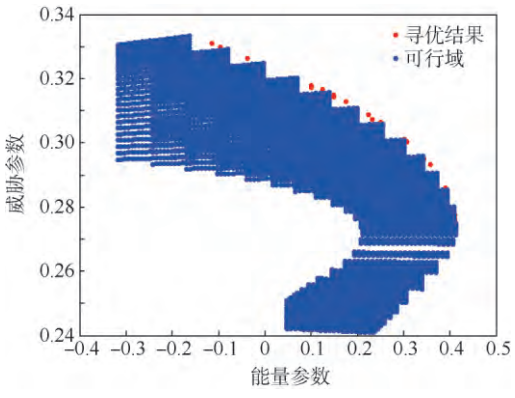

图 5 算法寻优效果

Fig. 5 Algorithm optimization result

通过图 6 可知，我机态势在多目标优化方法的决策下明显优于初始时刻，且过程中威胁参数基本始终保持递增，而能量参数仅出现一次大幅下降后同样保持递增(初始的大幅机动必然会导致能量损失)。

为了更直观地展示多目标优化的性能，图 7 记录了在相同初始条件下进行 2 次重复实验的结果(其中红色为我方轨迹，蓝色为敌方轨迹)，在

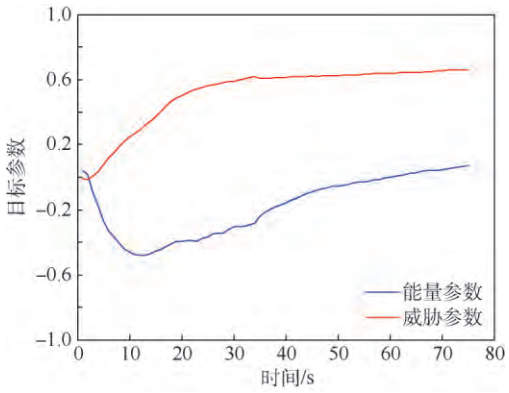

图 6 目标函数变化趋势

Fig. 6 Change trend of objective function

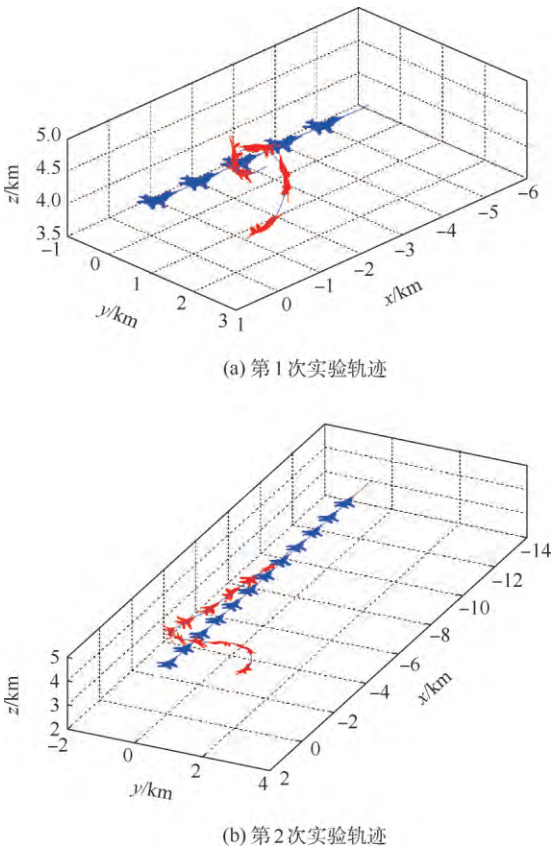

图 7 仿真轨迹(相同初始条件)

Fig. 7 Simulation trajectory map (the same initial conditions)

初始条件相同的 2 次仿真中，我方做出了 2 次不同但均有效的机动。

2.2　辅助决策模型有效性验证

为了验证辅助决策模型的有效性，按照 1.3 节中的强化学习模型训练辅助决策网络，训练总次数为 20 000 次，训练中双方战机采用相同的决策模型。

为了实时反映训练效果，每进行 200 次训练就对辅助模型性能进行一次检测；检测方法类似 2.1 节中验证决策可行性的实验方法，但我机在决策时采用辅助网络（即使用 1.4 节决策模型），为了节省时间，每次检测重复 200 次且仅记录最终结果，测试结果如图 8 所示。

由测试结果可以看出，随着训练次数的增加，我方的获胜次数明显得到了提升，获胜概率从 25% 左右提升到 50% 左右，说明在辅助网络的帮助下，模型可以给出更为有效的决策。此外，由于双方初始位置为随机产生，每次测试中必然会出现少数极端不利的初始条件，故测试结果中一直存在一定的失败次数。

为了进一步体现辅助网络的效果，使用带辅助网络的决策模型与仅使用多目标优化方法的决策模型进行对抗仿真（将仅使用优化方法的一方视为敌机），为了使仿真结果更具代表性，初始态势将在一定范围内随机产生，具体约束条件如下。

初始距离 d 范围为（3，10） km 之间，初始速度差 $\Delta v=\sqrt{2g\Delta h}$（保证双方初始能量相同），并根据 $\alpha_e+\alpha_u$ 的值分为 3 种情形：

1）初始有利（$\alpha_e+\alpha_u\in[0°,90°]$）。

2）初始均势（$\alpha_e+\alpha_u\in(90°,270°)$）。

3）初始不利（$\alpha_e+\alpha_u\in[270°,360°]$）。

在 3 种情形下各进行 100 次对抗仿真，结果如表 1 所示。

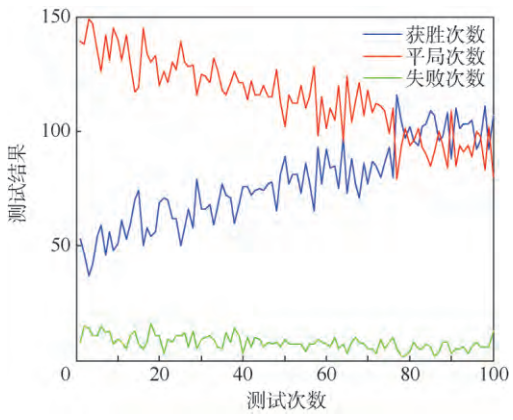

图 8　辅助决策模型性能测试结果

Fig. 8　Test results of auxiliary decision model's performance

表 1　对抗仿真结果（使用辅助网络一方）

Table 1　Confrontation simulation results (the side with auxiliary network)

初始条件	获　胜	平　局	失　败
初始有利	59	41	0
初始均势	32	51	17
初始不利	11	39	50

通过仿真结果可知，在使用了辅助网络后，决策模型可以做出更高效、更具有对抗性的决策，平均获胜概率提升了 11.7%。

2.3　机动决策模型性能仿真

为了体现本文机动决策模型的性能，设置了 2 种情形下的空战环境，其中敌机采用的机动均为经典的战术动作，我机采用基于多目标优化的机动决策模型，辅助决策网络采用 2.2 节仿真实验中训练出来的神经网络。

1）情形 1 中，我机初始处于较优的态势环境，敌机采用“S 型”机动进行规避，仿真结果如图 9所示。图中红色为我方，蓝色为敌方，轨迹上的飞机模型表示飞机当前姿态，相邻 2 个模型时间间隔为 4 s。

图 10 分别给出了空战过程中双方攻击判定条件（视线角与距离，我机视线角 α_u 的定义见图 2，敌机视线角即为 $\pi-\alpha_e$）以及我方决策得出的控制量的实时变化情况。

通过仿真数据可知，初始条件下我方占据较大优势，决策模型根据敌方位置调整我方视线角以形成攻击条件；但由于我方速度较大且敌方采取“S 型”机动，在 20 s 左右我方基本完成转向后存在超越敌方的风险；决策模型采用了类似异面机动的原理，先适当俯冲再拉起机头以避免战机冲前，在拉起机头的过程中再次调整视线角；从第 44 s 开始对敌方形成有效攻击条件并保持，47 s 时达到仿真结束条件，我方获胜。

2）情形2中，我方初始处于不利条件，但由于

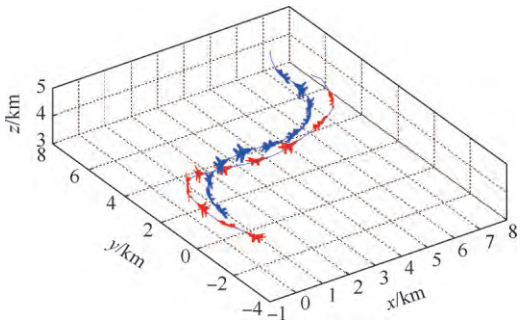

图 9　仿真轨迹（初始有利）

Fig. 9　Simulation trajectory map (favorable initial conditions)

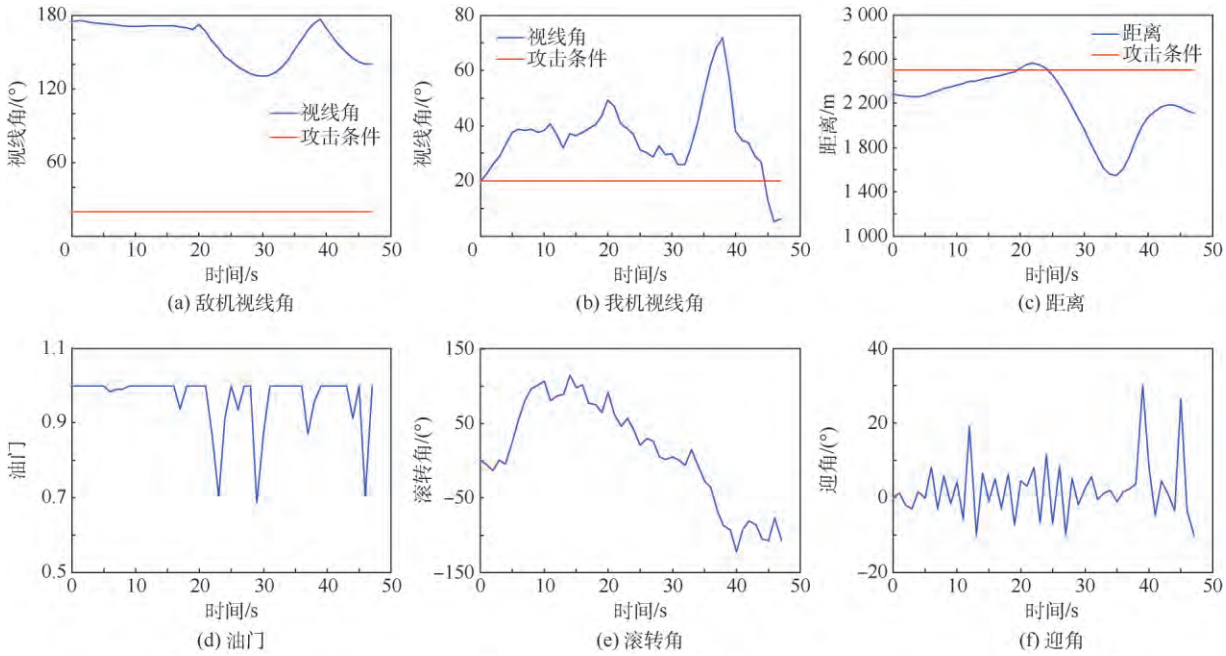

图 10 态势关系与控制量(初始有利)

Fig. 10 Situation relationship and control quantity (favorable initial conditions)

距离敌机较远,故存在机动规避的空间;敌机采用“纯跟踪”的方法试图接近并攻击我方,仿真结果如图 11 所示(图中标记同情形 1)。

图 12 分别给出了空战过程中双方攻击判定条件(视线角与距离)以及我方决策得出的控制量的实时变化情况。

通过仿真数据可知,初始我方处于不利态势,决策模型选择在向右机动规避的同时拉起机头;爬升的过程必然会损失动能,故双方距离逐渐缩小,15 s左右时,敌方开始右转以保持态势优势;

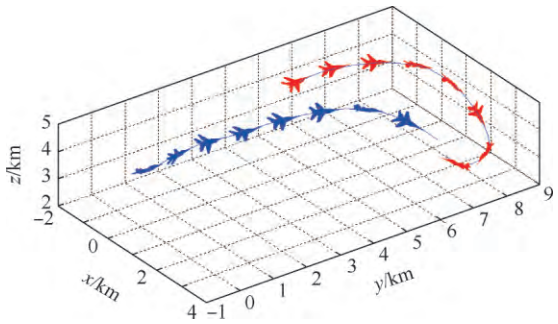

图 11 仿真轨迹(初始不利)

Fig. 11 Simulation trajectory map (adverse initial conditions)

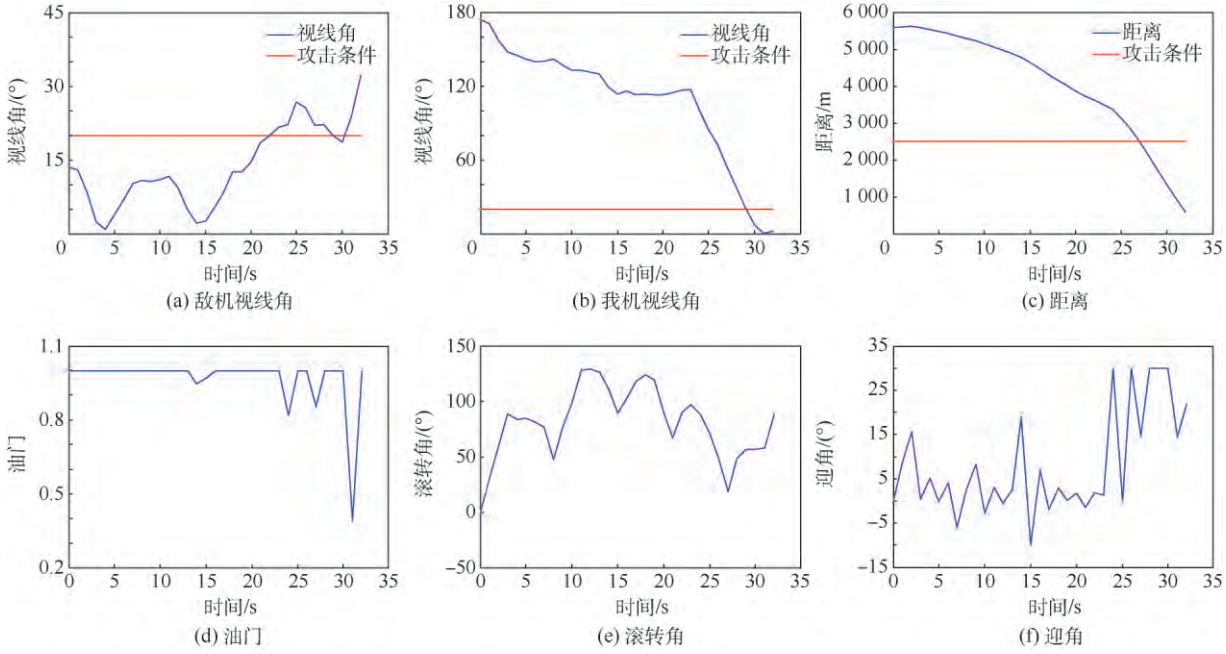

图 12 态势关系与控制量(初始不利)

Fig. 12 Situation relationship and control quantity (adverse initial conditions)

由于此时我方速度较低，具有更小的转弯半径，决策模型选择向右下方急转接敌，并在第 29 s 抢先形成攻击条件并保持，32 s 时达到仿真结束条件，我方获胜。

3　结　论

本文提出了多目标优化与强化学习相结合的机动决策模型，模型融合了传统优化方法与机器学习方法的优点：

1）多目标优化方法解决了传统优化方法中处理目标函数权重的问题，增加了决策模型的可信度和可拓展性。

2）多目标优化方法继承了传统优化方法的优点，可以进行实时有效的机动决策。

3）多目标优化的决策集直接给出了足够的可执行决策，极大程度上简化了动作空间，使强化学习任务具备了可行性。

4）通过强化学习建立辅助决策网络，从而可以在多目标优化决策集中做出更好的选择，弥补了优化方法在对抗、博弈问题上的不足。

由于本文重点在于结合传统优化方法和机器学习方法，在设置优化目标时仅针对较为理想的仿真环境设置了 2 个目标，设置在复杂电磁环境下新的目标函数模型是下一步的改进方向。

参考文献（References）

[1] 黄长强，唐上钦. 从“阿法狗”到“阿法鹰”——论无人作战飞机智能自主空战技术[J]. 指挥与控制学报，2016，2(3)：261-264.
HUANG C Q，TANG S Q. From Alphago to Alphaeagle：On the intelligent autonomous air combat technology for UCAV [J]. Journal of Command and Control，2016，2(3)：261-264(in Chinese).

[2] 朱可钦，董彦非. 空战机动动作库设计方式研究[J]. 航空计算技术，2001，31(4)：50-52.
ZHU K Q，DONG Y F. Study on the design of air combat maneuver library [J]. Aeronautical Computer Technique，2001，31(4)：50-52(in Chinese).

[3] 钟友武，柳嘉润，杨凌宇，等. 自主近距空战中机动动作库及其综合控制系统[J]. 航空学报，2008，29(s1)：114-121.
ZHONG Y W，LIU J R，YANG L Y，et al. Maneuver library and integrated control system for autonomous close-in air combat [J]. Acta Aeronautica et Astronautica Sinica，2008，29(s1)：114-121(in Chinese).

[4] 钟友武，柳嘉润，申功璋. 自主近距空战中敌机的战术动作识别方法[J]. 北京航空航天大学学报，2007，33(9)：1056-1059.
ZHONG Y W，LIU J R，SHEN G Z. Recognition method for tactical maneuver of target in autonomous close-in air combat [J]. Journal of Beijing University of Aeronautics and Astronautics，2007，33(9)：1056-1059(in Chinese).

[5] 张涛，于雷，周中良，等. 基于混合算法的空战机动决策[J]. 系统工程与电子技术，2013，35(7)：1445-1450.
ZHANG T，YU L，ZHOU Z L，et al. Decision-making for air combat maneuvering based on hybrid algorithm [J]. Systems Engineering and Electronics，2013，35(7)：1445-1450(in Chinese).

[6] SU M C，LAI S C，LIN S C，et al. A new approach to multi-aircraft air combat assignments [J]. Swarm & Evolutionary Computation，2012，6：39-46.

[7] 冯超，景小宁，李秋妮，等. 基于隐马尔可夫模型的空战决策点理论研究[J]. 北京航空航天大学学报，2017，43(3)：615-626.
FENG C，JING X N，LI Q N，et al. Theoretical research of decision-making point in air combat based on hidden Markov model [J]. Journal of Beijing University of Aeronautics and Astronautics，2017，43(3)：615-626(in Chinese).

[8] 张彬超，寇雅楠，邬蒙，等. 基于深度置信网络的近距空战态势评估[J]. 北京航空航天大学学报，2017，43(7)：1450-1459.
ZHANG B C，KOU Y N，WU M，et al. Close-range air combat situation assessment using deep belief network [J]. Journal of Beijing University of Aeronautics and Astronautics，2017，43(7)：1450-1459(in Chinese).

[9] 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[J]. 航空学报，2017，38(10)：321168.
ZUO J L，YANG R N，ZHANG Y，et al. Intelligent decision-making in air combat maneuvering based on heuristic reinforcement learning [J]. Acta Aeronautica et Astronautica Sinica，2017，38(10)：321168(in Chinese).

[10] PARETO V. Cours d’economie politique [M]. Lausanne，Paris：F. Rouge，1896.

[11] 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.
GUO H F，HOU M Y，ZHANG Q J，et al. UCAV robust maneuver decision based on statistics principle [J]. Acta Armamentarii，2017，38(1)：160-167(in Chinese).

[12] HUANG C，DONG K，HUANG H，et al. Autonomous air combat maneuver decision using Bayesian inference and moving horizon optimization [J]. Journal of Systems Engineering and Electronics，2018，29(1)：86-97.

[13] VEERASAMY N. A high-level mapping of cyberterrorism to the OODA loop [C] // Proceedings of 5th European Conference on Information Management and Evaluation. Red Hook，NY：Curren Associates Inc.，2011：352-360.

[14] MIRJALILI S，SAREMI S，MIRJALILI S M，et al. Multi-objective grey wolf optimizer：A novel algorithm for multi-criterion optimization [J]. Expert Systems with Applications，2016，47：106-119.

[15] 崔明朗，杜海文，魏政磊，等. 多目标灰狼优化算法的改进策略研究[J]. 计算机工程与应用，2018，54(5)：156-164.
CUI M L，DU H W，WEI Z L，et al. Research on improved strategy for multi-objective grey wolf optimizer [J]. Computer Engineering and Applications，2018，54(5)：156-164(in Chinese).

[16] 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.
MA Y F,GONG G H,PENG X Y. Cognition behavior model for air combat based on reinforcement learning [J]. Journal of Beijing University of Aeronautics and Astronautics,2010,36(4):379-383(in Chinese).

[17] BOUZY B,CHASLOT G. Monte-Carlo go reinforcement learning experiments [C]//2006 IEEE Symposium on Computational Intelligence and Games. Piscataway,NJ: IEEE Press,2007: 187-194.

[18] 左磊. 基于值函数逼近与状态空间分解的增强学习方法研究[D]. 长沙: 国防科学技术大学,2011.
ZUO L. Research on reinforcement learning based on value function approximation and state space decomposition [D]. Changsha: National University of Defense Technology,2011(in Chinese).

[19] HORNIK K,STINCHCOMBE M,WHITE H. Multilayer feedforward networks are universal approximators [J]. Neural Networks,1989,2(5):359-366.

[20] SUTTON R S. Learning to predict by the method of temporal differences [J]. Machine Learning,1988,3(1):9-44.

[21] WILLIAMS P. Three-dimensional aircraft terrain-following via real-time optimal control [J]. Journal of Guidance,Control and Dynamics,2007,30(4):1201-1206.

[22] WILLIAMS P. Aircraft trajectory planning for terrain following incorporating actuator constraints [J]. Journal of Aircraft,2005,42(5):1358-1361.

作者简介:

杜海文 男,硕士,教授,硕士生导师。主要研究方向: 机载武器系统应用工程。

崔明朗 男,硕士研究生。主要研究方向: 无人飞行器武器作战系统与技术。

韩统 男,博士,副教授,硕士生导师。主要研究方向: 机载武器系统应用工程。

Maneuvering decision in air combat based on multi-objective optimization and reinforcement learning

DU Haiwen[1,*], CUI Minglang[1], HAN Tong[1], WEI Zhenglei[1], TANG Chuanlin[2], TIAN Ye[3]

(1. College of Aeronautics and Astronautics, Air Force Engineering University, Xi'an 710038, China;
2. Unit 94782 of PLA, Hangzhou 310004, China; 3. College of Physics and Information Engineering, Fuzhou University, Fuzhou 350108, China)

Abstract: To solve the problem of maneuvering decision in the autonomous air combat of unmanned combat aerial vehicle, the existing research achievements are analyzed and a maneuvering decision model that combines optimization idea with machine learning is proposed. The multi-objective optimization method is used as the core of decision model, which solves the problem of setting weight for multiple optimization targets and improves the extensibility of decision model. On the basis of multi-objective optimization, an evaluation network is trained by reinforcement learning and used for auxiliary decision-making to enhance the antagonism of decision model. In order to test the performance of decision model, with the background of short-range air combat, three simulation experiments are designed to test the feasibility of multi-objective optimization method, the effectiveness of auxiliary decision network and the overall performance of decision model. The simulation results show that the maneuvering decision model can be used in real-time confrontation with the maneuvering enemy aircraft.

Keywords: autonomous air combat; maneuvering decision; multi-objective optimization; reinforcement learning; neural network

Received: 2018-03-15; **Accepted:** 2018-06-15; **Published online:** 2018-07-27 18:17

URL: kns.cnki.net/kcms/detail/11.2625.V.20180727.1038.001.html

Foundation items: National Natural Science Foundation of China (61601505); Natural Science Foundation of Shaanxi Province of China (2017JM6078)

* **Corresponding author**. E-mail: 18191856512@163.com