基于MAXQ分层强化学习的有人机/无人机协同路径规划研究

程先峰[1]，严勇杰[2]
（1. 南京莱斯信息技术股份有限公司，南京，210007；
2. 空中交通管理系统与技术国家重点实验室，南京，210014）

摘　要：针对有人机/无人机混合运行的复杂任务和环境下，很难为无人机协调行为设计合适的控制策略和控制参数的问题，文章设计了基于对策论的无人机强化学习模型与算法。针对无人机/有人机协调的特点，结合MAXQ分层强化学习和Multi-agent的优点，采用了一种基于MAXQ的Multi-agent分层强化学习的无人机协调方法，增强了无人机在混合运行复杂环境下适应环境和自协调的能力。

关键词：有人机/无人机协同；分层强化学习；协同路径规划
中图分类号：TN915.03

0　引　言

随着无人机(Unmanned Aerial Vehicle，UAV)系统在物流、作战等领域的广泛使用，有人机/无人机协同执行任务的问题逐渐受到了国外研究机构和学者的普遍关注。在过去很长一段时间里，由于无人机“感知-避让”性能上的缺陷，国内外对无人机的管制一般都采取划设隔离空域的方式运行，这样就避免了无人机对有人机的安全干扰。但随着无人机技术的发展，无人机的飞行高度覆盖从近地面的几十米到几万米高空，航程可达上万公里，无人机早已具备了全空域飞行的能力[1]。以往划设“隔离空域”的方式无法再满足无人机的空域使用需求，无人机飞出“隔离空域”进入“非隔离空域”与有人机共同飞行已成为必然趋势。

有人机/无人机混合飞行是一个复杂过程，要达到出色的协同效果，涉及各个环节的良好配合，这包括协同态势感知与评估技术、协同任务分配技术、协同路径规划技术、编队飞行与跟踪控制技术等。其中，有人机/无人机协同路径规划是指在满足飞行任务、单机性能以及空中环境等各种约束条件下，在协同任务规划方案基础上规划各机可行有效的飞行路径，满足多机在空间和时间上的协调一致关系，避免飞行冲突，具有高度不确定性和突发性的特点。为使无人机进入“非隔离空域”运行，美国国防部(United States Department of Defense)与联邦航空局(Federal Aviation Administration，FAA)分别对军用无人机与民用无人机制定了详细的空域集成计划，以促进无人机尽快融入美国国家空域系统。其中，“感知-避让”系统是其研究重点，但主要基于协同路径规划技术对动态与静态的威胁源的避让，总体上还不够完善，这主要的原因就在于无人机“感知-避让”这一过程取决于很多的组合因素，需要进一步加以细化研究。从国内的研究现状来看，大量研究都使用群智能算法作为无人机路径规划的基本模型，虽然在路径规划质量上都取得了较好的效果，实时性要求上还很难适应快速多变的空域环境。

路径规划本身是一个约束条件多且相互耦合的多目标优化与决策问题，需要综合利用运筹学、智能计算以及计算几何等理论，而有人机/无人机混合系统协同路径规划问题更加复杂，十分具有挑战性。近十几年来，关于机器学习的研究越来越受到人们的关注。强化学习在单个及多个智能体行为的学习研究中取得了成功，为解决多个智能体协调所涉及的学习空间“维数灾难问题[2]”和多智能体系统的非马尔可夫问题[3]，Littman[4]采用对策论作为Multi-agent系统的形式框架，将Multi-agent系统的强化学习模型转化为马尔可夫对策强化学习，解决随机变化环境的问题；为了克服维数灾难问题，越来越多的研究者把注意力集中到分层强化学习(Hierarchical Reinforcement Learning，HRL)上。HRL方法引入抽象机制实现状态空间降维，将强化学习任务分解到抽象内部和抽象间的不同层次上分别实现，从而每层上的学习任务仅需在低维空间中进行。近

收稿日期：2019-11-18

年来，HRL 在解决维数灾难问题中取得了显著进展，典型的成果有 Sutton 提出的基于选件学习的 Option算法[5]，Parr 提出的基于分层局部策略的学习 HAM 的算法[6]和 Dietterich 提出基于子任务学习的 MAXQ 算法[7]。

本文将无人机/有人机的协同转化为多智能体协同问题进行考虑，在基于马尔可夫的 Multi-agent 强化学习基础上，结合 MAXQ 学习的优点和 Multi-agent 的特点，将 MAXQ 分层强化学习方法用到对有人机/无人机协同运行的多机路径规划中，并进行了仿真实现。

1　分层强化学习的基础理论

1.1　强化学习的思想

强化学习的过程是一个试探、评价的过程。首先，强化学习系统感知环境状态，采取一个动作作用于环境；环境接受该动作后状态发生变化，同时给出一个强化信号（奖励或惩罚），反馈给强化学习系统；强化学习系统根据强化信号和环境当前状态，以使学习系统增大受到正强化（奖励）的概率为原则，选择下一个动作。强化学习系统的原理如图 1 所示。

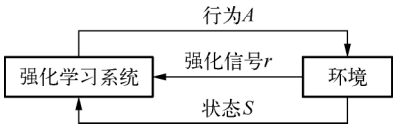


图 1　强化学习基本原理示意图

强化学习系统的主要组成要素有动作策略、奖赏函数和值函数。

1）动作策略

动作策略是指学习系统在一个给定的状态下产生动作的方法。针对系统状态集合 S 中的每一个状态 s，学习系统选择动作集合 A 中的一个动作 a，记动作策略为 π，则 $\pi: S \rightarrow A$，是一个从状态空间到动作空间的映射函数。

2）奖赏函数

奖赏函数，记为 $r(s)$或 $r(s,a)$，根据当前系统状态和所选择的动作，产生一个奖赏信号。奖赏函数是从状态（或状态-动作对）到一个奖赏信号的映射。奖赏信号 r 是对所选择动作的作用效果的一种评价。奖赏信号通常是一个标量，常用正数表示奖励，用负数表示惩罚，强化学习过程的目标就是使系统最终得到的总的奖赏值最大。

3）值函数

奖赏函数是对一个状态（或状态-动作对）即时的评价，而值函数则是从长远的角度考虑一个状态（或状态-动作对）的好坏。值函数又称为评价函数。状态 s_t 的值定义为从状态 s_t 开始，按动作策略 π 选择后续动作，直到最终到达目标期间所得到的累计奖赏的期望，记为 $V_\pi(s_t)$：

$$V_\pi(s) = E_k\{R_t \mid s_t = s\} = E_k\left(\sum_{k=0}^{\infty}\gamma^k r_{t+k+1} \mid s_t = s\right) \quad (1)$$

式中，$r_t = R(s_t, a_t)$为 t 时刻的奖赏；γ 为后续动作奖赏的折扣率，$\gamma \in [0,1]$。

动作 a 的值定义为在状态 s 选择动作 a，接着按动作策略 π 选择后续动作，直到最终到达目标，这期间所得到的累计奖赏的期望，记为 $Q_\pi(s,a)$，即为 Q-学习方法：

$$Q_\pi(s,a) = E_\pi\{R_t \mid s_t = s, a_t = a\} = E_\pi\left\{\sum_{k=0}^{\infty}\gamma^k r_{t+k+1} \mid s_t = s, a_t = a\right\} \quad (2)$$

1.2　半马尔柯夫决策过程（SMDP）模型

分层强化学习（HRL）的实质是通过在强化学习的基础上增加“抽象”机制，把整体任务分解为不同层次上的子任务，使每个子任务在规模较小的子问题空间中求解，并且求得的子任务策略可以复用，从而加快问题的求解速度。与强化学习不同，在 HRL 中的决策时间间隔是一个变化的量，这与达到某个特殊状态或完成某个子任务的时间有关，可以采用半马尔柯夫决策过程（Semi-Markov Decision Processes，SMDP）对其进行模型化。

半马尔可夫决策过程是对马尔可夫决策过程的扩展，它允许动作在多个时间步内完成。系统状态可以在动作执行时连续变化，而不是像马尔可夫过程一样，状态变化由动作决定。在分层强化学习中，所建立的模型都是以半马尔可夫决策过程为基础的。

定义：一个半马尔可夫决策过程可以描述成一个五元组$\langle S, A, P, R, I\rangle$。其中，$S$，$A$ 分别是有限状态和动作的集合；$P: S \cdot N \cdot S \cdot A \rightarrow [0,1]$是多步转移概率函数，$P(s', N|s, a)$表示采取动作 a，在 N 步内系统状态由 s 转移到 s' 的概率；$R: S \cdot A \rightarrow R$ 是奖赏函数，$r(s,a)$是系统在状态 s 选择动作 a 后期望获得的总的奖赏值，它包含了分析半马尔可夫决策过程获得的报酬的所有必要信息；I 是初始的状态分布。

对于 SMDP 中每一次状态转变，到下一个决策过程期望的时间步可以定义为：

$$y(a,s)=E[N\mid s,a]=\sum_{N\in N}N\sum_{s'\in S}P(s',N\mid s,a) \tag{3}$$

与 MDP 一样，在 SMDP 中，我们的目的是找到一个最优策略使获得的奖赏值最大。SMDP 中动作在一定时间步内完成，对任意状态 s 按照策略 π 获得的奖赏可以表示为：

$$V^{\pi}(s)=E[r(s_0,\pi(s_0))+\gamma^{N_0}r(s_1,\pi(s_1))+\gamma^{N_0+N_1}r(s_2,\pi(s_2))+L\mid s_0=s,\pi] \tag{4}$$

则基于 SMDP 的值函数 Bellman 方程和状态-动作对值函数 Bellman 方程分别为：

$$V^{\pi}(s)=r(s,\pi(s))+\sum_{s'\in s,N\in N}\gamma^{N}P(s',N\mid s,\pi(s))V^{\pi}(s') \tag{5}$$

$$Q^{\pi}(s,a)=r(s,a)+\sum_{s'\in s,N\in N}\gamma^{N}P(s',N\mid s,a)Q^{\pi}(s',\pi(s')) \tag{6}$$

其 Bellman 最优方程分别为：

$$V^{*}(s)=\max_{a\in A_s}\Big[r(s,a)+\sum_{s'\in s,N\in N}\gamma^{N}P(s',N\mid s,a)V^{*}(s')\Big] \tag{7}$$

$$Q^{*}(s,a)=r(s,a)+\sum_{s'\in s,N\in N}\gamma^{N}P(s',N\mid s,a)\max_{a'\in A_{s'}}Q^{*}(s',a') \tag{8}$$

2　MAXQ 分层强化学习模型与算法

2.1　分层结构模型

设给定任务为 M，并将它分解成一系列子任务的集合 $\{M_0,M_1,\cdots,M_n\}$，习惯上，定义 M_0 为根任务，每个子任务可以定义为一个三元组 $\langle T_i,A_i,\widetilde{R}_i\rangle$。

（1）T_i 为终止判据，将状态集 S 分成活动状态集 S_i，和终止状态集 T_i，子任务 M_i 的策略仅当当前状态 $s\in S_i$ 时才被执行。当 M_i 被执行时，一旦达到 T_i 中的一个状态，则 M_i 立即终止。

（2）A_i 为用于完成子任务 M_i 的动作集，这些动作可以是来自 MDP 基本动作集 A 的基本动作，也可以是其他子任务的动作。

（3）$\widetilde{R}_i$ 为伪报酬函数，它表示从状态 $s\in S_i$ 到终止状态 $s'\in T_i$ 的每一步转移所获的伪报酬，反映了此子任务对每个终止状态的期望。一般来说，赋予目标终止状态的伪报酬为 0，非目标终止状态的伪报酬为一负值。伪报酬仅仅用于学习过程中。

下面以无人机运动协调为例进行分析，假设有人机的优先级高于无人机，即无人机路径规划时尽量减少对正常有人机飞行路径的影响。基于无人机不同层次的能力，在 MAXQ 结构的基础上提出了一个分层的自底向上的多无人机学习模型用于无人机与有人机的协调规划问题，如图 2 所示。

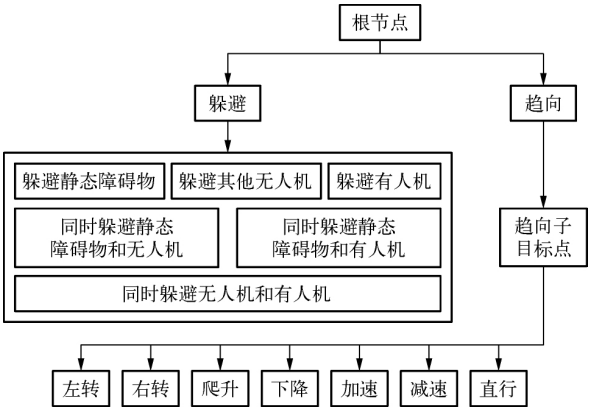


图 2　分层强化学习模型

基本动作包括左转、右转、直行、加速、减速、爬升和下降 7 个，动作的执行需要满足无人机自身性能参数的要求，将整个任务分解为子任务如下：

（1）根节点：表示整个协调规划任务；

（2）趋向子目标点：使无人机朝着目标点移动；

（3）躲避静态障碍物：当检测到固定障碍物时，让无人机躲避与之发生碰撞，静态障碍物包括禁区、危险区、建筑物、山峰等；

（4）躲避其他无人机：当检测到一定范围内有其他无人机存在时，避免两者发生碰撞；

（5）躲避有人机：当检测到一定范围内有有人机存在时，避免两者发生碰撞；

（6）同时躲避静态障碍物和无人机：当检测到同时存在静态障碍物和其他无人机时，让无人机避免与两者发生碰撞；

（7）同时躲避静态障碍物和有人机：当检测到同时存在静态障碍物和有人机时，让无人机避免与两者发生碰撞；

（8）同时躲避无人机和有人机：当检测到同时存在其他无人机和有人机时，让无人机避免与两者发生碰撞。

整个分层模型中，每个子任务是由子目标定义的，当子目标达到时子任务结束。假设对每一个子任务用一个策略来完成，可以将完成这些子任务的

策略看作是一个个子程序，上层子任务的完成就可以看成调用下一层子任务程序的过程。若具备每个子任务的策略，则可以获得整个任务策略。如路径规划任务是通过调用躲避子任务策略或趋向目标点子任务策略程序来实现的，而躲避子任务策略又调用子层的3个子任务策略程序等，称这些策略的集合为一个分层策略。在这个分层策略中，执行每一个子程序直到进入该子任务的终止状态。将图2用MAXQ图来描述，如图3所示。

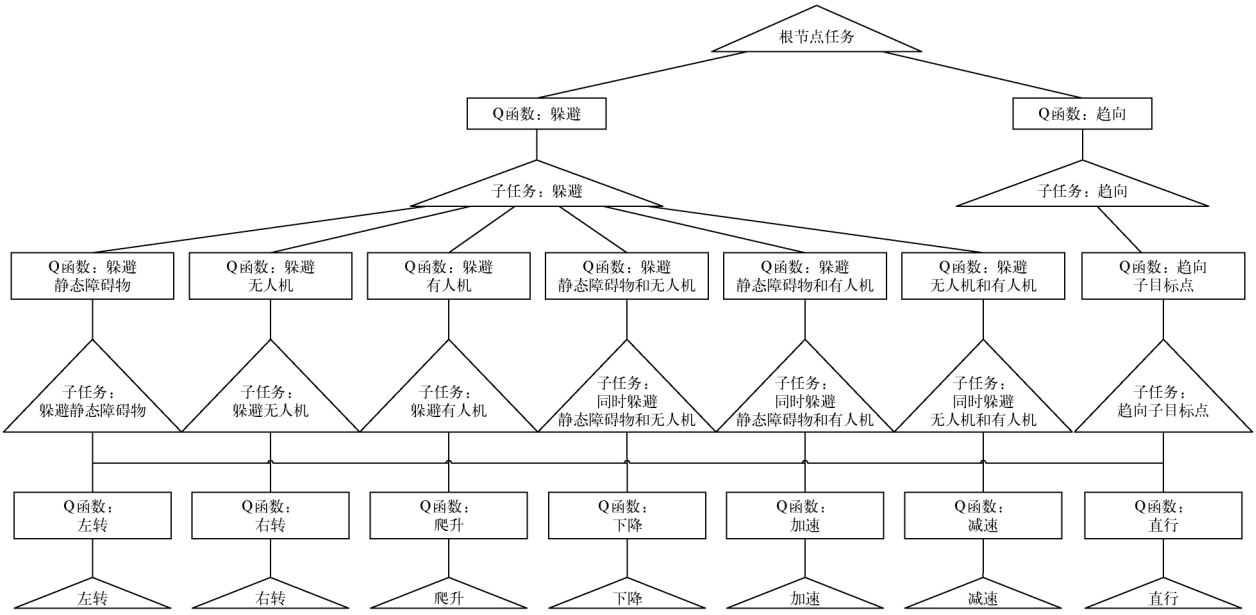


图3　多有人机/无人机运动协调的MAXQ图

为便于设计，Dietterich定义了MAXQ图来表示任务的分解结构。图3中包括两类节点：Max节点（以三角形表示）和Q节点（以方框表示）。Max节点负责任务分解后的子任务，每一个基本动作对应一个Max节点，每一个子任务也对应一个Max节点，包括根节点。其中每一个基本Max节点i保存一个$V^{\pi}(i,s)$值。Q节点对应每个子任务可获得的动作，每个Q节点保存状态s下父代任务i的子代动作为a的完成函数$C^{\pi}(i,s,a)$。

2.2　评价函数分解结构

分层结构中，各个子任务是基于半马尔可夫模型的。SMDP与MDP模型的区别在于在状态s下动作a被执行多个时间步，这个时间步数记为N。给定一个任务分解图$\{M_0,\cdots,M_n\}$，一个分层策略π，每个子任务定义为一个SMDP模型，则其状态集$\boldsymbol{S}$，动作集$\boldsymbol{A}$，状态转移函数$P_i^{\pi}(s',N\mid s,a)$和期望报酬函数$\tilde{r}(s,a)=V^{\pi}(a,s)$，其中，$V^{\pi}(a,s)$为在状态$s$下执行子代任务$M_a$的项目评价函数，若$a$为基本动作，则

$$V^{\pi}(a,s)=\sum_{i}P(s'\mid s,a)r(s'\mid s,a) \tag{9}$$

假设子任务策略π_i选择的第一个动作a，调用动作a子程序，并被执行N次后系统终止于状态s'，其状态转移概率为$P_i^{\pi}(s',N\mid s,a)$。则Bellman方程为：

$$V^{\pi}(i,s)=V^{\pi}(\pi_i(s),s)+\sum_{s',N}P_i^{\pi}(s',N\mid s,\pi_i(s))V^{\pi}(i,s') \tag{10}$$

为了获得评价函数的分层分解，将上式转换成Q函数表达式，令$Q^{\pi}(i,s,a)$表示为完成子任务M_i，从状态s出发根据分层策略π执行动作a直到子任务M_i终止的累积报酬期望：

$$Q^{\pi}(i,s,a)=V^{\pi}(a,s)+\sum_{s',N}P_i^{\pi}(s',N\mid s,a)\gamma^{N}Q^{\pi}(i,s',\pi(s')) \tag{11}$$

定义完成函数$C^{\pi}(i,s,a)$为在状态s下经调用子任务M_a子程序而完成子任务M_i的累积折扣报酬期望：

$$C^{\pi}(i,s,a)=\sum_{s',N}P_i^{\pi}(s',N\mid s,a)\gamma^{N}Q^{\pi}(i,s',\pi(s')) \tag{12}$$

则Q函数重写为：

$$Q^{\pi}(i,s,a)=V^{\pi}(a,s)+C^{\pi}(i,s,a) \tag{13}$$

这样在一般情况下，在状态s下执行子任务i的评价函数$V^{\pi}(i,s)$可表示为：

$$V^{\pi}(i,s)=\begin{cases}Q^{\pi}(i,s,a) & \text{若 } i \text{ 是复合任务}\\ \sum_{s'}P(s'\mid s,i)r(s'\mid s,i) & \text{若 } i \text{ 是基本动作}\end{cases} \tag{14}$$

式(12)，(13)和(14)表达了 MAXQ 分层结构的某一分层策略评价函数的分解方程，这些方程递归地将根层的项目评价函数 $V^{\pi}(0,s)$ 分解成单个子任务 $\{M_1,\cdots,M_n\}$ 的项目评价函数和完成函数 $C^{\pi}(j,s,a),j=1,\cdots,n$。表达评价函数分解的最基本内容包括所有非基本动作子任务的 C 函数和所有基本动作的 V 函数。

根据上述方法，图 3 中，在状态 s_1 下计算项目评价函数 $V^{\pi}(\text{Root},s_1)$。Root 根据它的策略 $\pi_{\text{Root}}(s_1)$，获得其策略动作为避障 Root，于是调用 QAvoid 计算 $Q^{\pi}(\text{Root},s_1,\text{Avoid})$，为完成 Root 任务而执行完 Root 子任务获得完成函数 $C^{\pi}(\text{Root},s_1,\text{Avoid})$，这仅仅是完成子任务 Avoid 获得的报酬，还要估计执行 Avoid 自身的期望报酬，搜索 Avoid 的策略获得子任务为躲避静态障碍物 AvoidObstacle，于是调用 QAvoidObstacle 计算 $Q^{\pi}(\text{Avoid},s_1,\text{AvoidObstacle})$，以及子任务 AvoidObstacle 的完成函数 $C^{\pi}(\text{Avoid},s_1,\text{AvoidObstacle})$，并调用 MaxAvoidObstacle，搜索 AvoidObstacle 子任务的策略获得子任务为左转，于是调用 Qleft 计算 $Q^{\pi}(\text{AvoidObstacle},s_1,\text{left})$ 和 $C^{\pi}(\text{AvoidObstacle},s_1,\text{left})$，并调用子任务 left，由于子任务是基本动作，因此执行完基本动作后终止，计算执行此基本动作的评价函数 $V^{\pi}(\text{left},s_1)$。这样，可得任务的项目评价函数为：

$$\begin{aligned}V^{\pi}(\text{Root},s_1)=&V^{\pi}(\text{left},s_1)+C^{\pi}(\text{AvoidObstacle},s_1,\\&\text{left})+C^{\pi}(\text{Avoid},s_1,\text{AvoidObstacle})\\&+C^{\pi}(\text{Root},s_1,\text{Avoid})\end{aligned} \tag{15}$$

一般可以将评价函数分解为以下形式：

$$\begin{aligned}V^{\pi}(0,s)=&V^{\pi}(a_m,s)+C^{\pi}(a_{m-1},s,a_m)+\\&L+C^{\pi}(a_1,s,a_2)+C^{\pi}(0,s,a_1)\end{aligned} \tag{16}$$

式中，$a_1,a_2,\cdots,a_m$ 是根据分层策略从任务到基本动作的节点“路径”。

2.3　算法描述

评价函数经分解后，对 V 函数和 C 函数的学习可以用标准的强化学习算法进行学习，这里用简单的 Q 学习算法。只是由于子任务模型是基于 SMDP 模型的，对 C 函数的学习也须是基于 SMDP 模型的更新规则。假设在状态 s 下智能体要完成任务 i，根据其子任务策略选择并执行了子代任务 a，经过 N 个时间步子代任务终止并导致状态转移到了 s'，则基于 SMDP 模型的完成函数的学习更新规则为：

$$\begin{aligned}C_{t+1}(i,s,a)\leftarrow&[1-\alpha_t(i)]C_t(i,s,a)+\alpha_t(i)\\&\gamma^N[C_t(i,s',a^*)+V_t(a^*,s')]\end{aligned} \tag{17}$$

式中，$\alpha_t(i)$ 为学习率。这个完成函数被父代任务用来计算评价函数 $V(i,s)$，即从状态 s 出发执行子任务 i 的期望报酬，与伪报酬函数无关。而为了搜索子任务 i 的局部最优策略，需要计算另一个节点内部的完成函数 $\widetilde{C}(i,s,a)$，这个函数同时与即时报酬函数 $r(s'|s,a)$ 和伪报酬函数 $\widetilde{r}_t(s)$ 相关：

$$\begin{aligned}\widetilde{C}_{t+1}(i,s,a)\leftarrow&[1-\alpha_t(i)]\widetilde{C}_t(i,s,a)+\alpha_t(i)\\&\gamma^N[\widetilde{r}_t(s')+\widetilde{C}_t(i,s,a^*)+V_t(a^*,s)]\end{aligned} \tag{18}$$

一个分层策略是每一个子任务策略的集合：$\pi=\{\pi_0,\cdots,\pi_n\}$，分层项目评价函数 $V^{\pi}(s)$ 是在状态 s 下，从根任务层出发执行层策略 π 的评价函数。以分层模型分解为 $\{M_0,\cdots,M_n\}$ 的一个 MDP 模型 M，其回归最优策略(Recursively Optimal Policy)是一个分层策略 $\pi=\{\pi_0,\cdots,\pi_n\}$，其中每个子任务 M_i 对应的策略 π_i 是基于 SMDP 模型最优的，此 SMDP 模型是由状态集 S_i、动作集 A_i、状态转移概率函数 $P^{\pi}(s',N|s,a)$ 以及由基本报酬函数 $r(s'|s,a)$ 和伪报酬函数 $\widetilde{r}_t(s')$ 之和组成的报酬函数定义的。

将算法步骤描述如下：

(1) 令 seq=()为当执行子任务 i 时所遍历的状态序列；

(2) 若 i 为基本动作节点，转(3)，否则，转(6)；

(3) 执行子任务 i，观察新状态 s' 和报酬 r_t；

(4) 更新：$V_{t+1}(i,s)\leftarrow[1-\alpha_t(i)]V_t(i,s)+\alpha_t(i)r_t$；

(5) 将 s 压入序列 seq 的开始处，转(14)；

(6) 若子任务 i 终止，转(14)，否则，转(7)；

(7) 根据当前探索策略 $\pi_t(i,s)$ 选择动作 a；

(8) 令 chileseq 为执行动作 a 所遍历的状态序列；

(9) 观察新状态 s'；

(10) 令 $a^*=\text{argmax}_{a'}[\widetilde{C}_t(i,s',a')+V_t(a',s')]$　$N=\text{length}(\text{childseq})$；

(11) 对 childseq 中每一状态 s 进行更新：

$\widetilde{C}_{t+1}(i,s,a)\leftarrow[1-\alpha_t(i)]\widetilde{C}_t(i,s,a)+\alpha_t(i)$

$\gamma^N[\tilde{r}_t(s')+\tilde{C}_t(i,s,a^*)+V_t(a^*,s)]$

$C_{t+1}(i,s,a)\leftarrow[1-\alpha_t(i)]C_t(i,s,a)+\alpha_t(i)\gamma^N[C_t(i,s',a^*)+V_t(a^*,s')]$

$N\leftarrow N-1$；转(11)直到 $N=0$；

(12) 将 childseq 序列添加到 seq 序列前面；

(13) $s\leftarrow s'$，转(6)；

(14) 若不满足系统终止条件，转(1)。

3 仿真实验与分析

3.1 仿真环境及假设

由于本文旨在对基于强化学习的多无人机/有人机协调问题进行研究，在初级阶段将实验平台建立在计算机仿真环境下。在一个 200 km×200 km 的环境中散落着几个不同形状和位置的障碍物，要求无人机从已知的起点向已知的终点以路径最短原则运动。下面对仿真环境和无人机做几点说明：

(1) 假设环境中的障碍物是山峰等固定障碍物；

(2) 假设环境比较开阔，不存在只允许一个无人机通行的狭窄地带，因此可将无人机假设为一个质点；

(3) 模拟的无人机性能参数：机长 5.5 m，翼展 8.0 m，最大速度 256 km/h，巡航速度 150～170 km/h，巡航高度 5 000 m，最大升限 7 000 m；

(4) 无人机要主动躲避有人机，避免碰撞。

3.2 强化信号

强化信号的作用是对学习行为的优劣作出评价。将强化信号分为 4 个部分：目标强化信号、躲避固定障碍物强化信号、躲避其他无人机强化信号和躲避有人机强化信号，分别定义如下：

$$R_{\text{goal}}=\begin{cases}1 & \text{趋向目标}\\ -1 & \text{远离目标}\\ -2 & \text{发生碰撞}\\ 0 & \text{其他}\end{cases}$$

$$R_{\text{avoidObs}}=\begin{cases}1 & \text{远离固定障碍物}\\ -2 & \text{靠近固定障碍物}\\ 0 & \text{其他}\end{cases}$$

$$R_{\text{avoidUAV}}=\begin{cases}1 & \text{远离其他无人机}\\ -2 & \text{靠近其他无人机}\\ 0 & \text{其他}\end{cases}$$

$$R_{\text{avoidAirS}}=\begin{cases}1 & \text{远离有人机}\\ -2 & \text{靠近有人机}\\ 0 & \text{其他}\end{cases}$$

3.3 动作选择策略

在学习过程中，无人机如果要获取较高的强化值，则在每个状态下都必须选择具有最高 Q 值的动作(即贪婪策略)。但是，在学习的初始阶段，Q 值并不能准确地反映正确的强化值。选择最高 Q 值的动作往往导致无人机总是采用相同的高 Q 值动作，不能探索其他动作，所以不能发现更好的动作组合。为了使无人机能够有效地探索和比较各种动作，在无人机选择动作时引入一定的随机性。本文采用基于动作概率的 Boltzmann 分布探索方法，在环境 s 下选择动作 a_i 的概率定义为：

$$P(a_i)=\frac{e^{Q(s,a_i)/T}}{\sum_{a_k\in A}e^{Q(s,a_k)/T}} \tag{19}$$

式中，T 是温度系数，T 越大，随机选择性越强，T 可以采用近似模拟退火的方法确定。

3.4 移动航空器威胁模型

根据《飞行间隔规定》的要求，在雷达管制的条件下，航空器在巡航阶段的间隔应保持在 10 km 以上。因此，可以借鉴 Reich 改进模型建立碰撞模型，将无人机看成质点，而将其他周围的航空器看成是半径为 10 km 的圆，当无人机质点进入航空器的移动圆内，则认为发生了飞行冲突，存在较大的相撞风险。威胁的计算公式为：

$$J_a=f_2/\min_{x,y\in\rho}(\sqrt{(x-x_f)^2+(y-y_f)^2}) \tag{20}$$

式中，(x_f,y_f)为其他航空器的位置坐标；f_2 为常数。

3.5 实验结果

3 个无人机采用分层强化学习算法所获得的路径如图 4 所示。

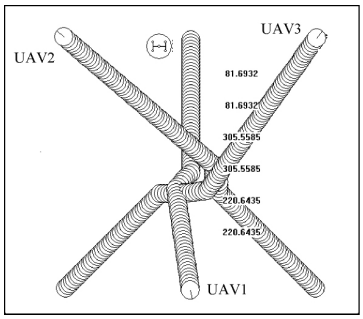
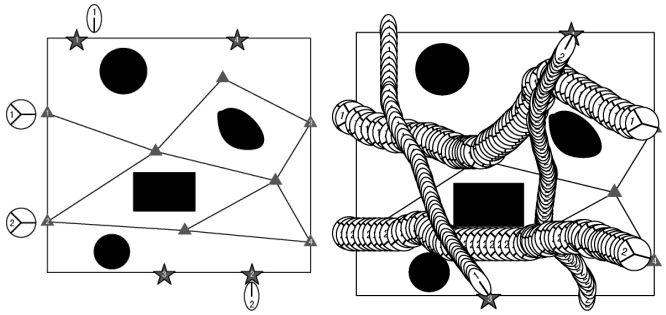


图 4　3 个无人机学习后的路径规划与避碰

图 4 仿真了多个无人机通过强化学习后规划出的无碰路径。2 个有人机和 2 个无人机经过学习后规划的路径如图 5 所示。



(a) 规划的路径示意图　　(b) 路径推演示意图

图 5　2 个有人机和 2 个无人机经过学习后规划的路径

图 5 为 2 个有人机和 2 个无人机运动协调结果，图 5(a)圆形的有人机，分别从▲1 去往目标点▲3、▲2 去往目标点▲4，有人机按照划设的航线飞行；椭圆形的是无人机，分别从★1 去往目标点★3、★2 去往目标点★4。从图 5(b)中看出，在存有多个障碍物的环境中，通过一段时间的学习后，无人机可以从给定的起始点避开有人机，顺利到达终点，而且获得了较好的优化路径。

4　结束语

本文针对无人机/有人机混合运行协调的特点，结合 MAXQ 分层强化学习和 Multi-agent 的优点，采用了一种基于 MAXQ 的 Multi-agent 分层强化学习的无人机/有人机协同路径规划方法；将无人机的任务分为不同层次，把子任务又分为协作子任务和非协作子任务，无人机在不同层次采用不同的协调策略，通过层层学习增强了无人机在复杂环境下适应环境和自协调的能力。从仿真结果看，这种分层强化学习算法获得了较好的优化结果，而且学习速度和效率与原有的分层强化学习相比有了改善。

参　考　文　献

[1] 蔡志浩，杨丽曼，王英勋，等. 无人机全空域飞行影响因素分析[J]. 北京航空航天大学学报，2011，37(02)：175-179.

[2] Kondo T，Ito K. A reinforcement teaming with evolutionary state recruitment strategy for autonomous mobile robots control[J]. Robotics and Autonomous Systems，2004，46(2)：111-124.

[3] Piao，Hong B S. Fast reinforcement learning approach to cooperative behavior acquisition in multi-agent system. Proceedings of the 2002 IEEEIRSJ Int[C]l. Conference on Intelligent Robots and Systems. 2002：871-875.

[4] Littman M L. Markov games as a framework for multiagent reinforcement learning. Proceedings of the 11'th International Conference on Machine Learning [C]. 1994：157-163.

[5] Sutton R S，Precup D，Singh S P. Between MDPs and semi-MDPs：a framework for temporal abstraction in reinforcement learning[J]. Artificial Intelligence，1999，112(1-2)：181-211.

[6] Parr R. Hierarchical control and learning for markov decision processes. Ph. D dissertation[R]，University of California，Berkeley，1998.

[7] Dietterich T G. Hierarchical reinforcement learning with the MAXQ value function decomposition[J]. Journal of Artificial Intelligence Research，2000(13)：227-303.

程先峰(1979—　)，男，高级工程师，主要研究方向为空管系统总体技术研究等。

Research on the Collaborative Path Planning of Manned/Unmanned Aerial Vehicles Based on MAXQ Hierarchical Reinforcement Learning

Cheng Xianfeng[1]，**Yan Yongjie**[2]

(1. Nanjing LES Information Technology Co.，Ltd.，Nanjing 210007，China；
2. State Key Laboratory of Air Traffic Management System and Technology，Nanjing 210014，China)

Abstract：Aiming at the complex task and environment of manned/unmanned aerial vehicles (MAV/UAV)，it is difficult to design appropriate control strategies and control parameters for the coordinated behavior of UAV. This paper designs a UAV reinforcement learning model and algorithm based on game theory. Aiming at the characteristics of MAV/UAV coordination，combining the advantages of MAXQ hierarchical reinforcement learning and Multi-agent，a UAV coordination method based on MAXQ Multi-agent hierarchical reinforcement learning was adopted to enhance the environment adaptive and self-coordination ability of UAV in a complex environment of mixed operation.

Key words：manned/unmanned aerial vehicles (MAV/UAV) coordination；hierarchical reinforcement learning；collaborative path planning