多智能体强化学习飞行路径规划算法

第 16卷 第 10期
2009年 10月

电 光 与 控 制
Electronics Optics & Control

Vol 16 No 10
Oct 2009

多智能体强化学习飞行路径规划算法

李东华， 江 驹， 姜长生
(南京航空航天大学自动化学院，南京 210016)

摘 要：为了减轻现代空战中大量信息处理给飞行员带来的负担，同时为了实现无人机航路自主规划，提出了一种基于多智能体强化学习理论的飞行路径规划算法。该算法采用多智能体强化学习的方法，采用两个功能不同的智能体，分别对应局部和全局路径规划。该算法对状态和动作空间进行划分和抽象，有效地减少了状态的数量，解决了强化学习维数灾难的问题。最后用 Matlab对此算法进行了数字仿真，验证了算法的可行性，仿真实验结果显示该算法收敛速度快，能够解决飞行路径规划的任务。

关键词：多智能体系统；强化学习；路径规划；无人机；自主规划
中图分类号：V279 文献标志码：A 文章编号：1671 - 637X(2009) 10 - 0010 - 05

概述

提出一种基于智能体强化学习理论的飞行路径规划算法.

1. 算法对状态空间进行划分和抽象, 避免了状态数量, 解决了强化学习维数灾难的问题
2. MATLAB对算法进行数字仿真, 验证算法可行性, 显示结果收敛快, 可解决飞行规划的问题

问题描述

> 本文考虑的问题是从某一点如何寻找一条安全的路径到达目标 。如图 1所示是一个简单的飞机进入敌方防御阵地的任务规划示意图 。图中飞机出发点用五角星表示，目标点用三角表示，地形 、雷达 、 导弹 、 高炮 、气候威胁等威胁源用圆圈表示，圆圈半径即为威胁源的威胁半径 。

强化学习

对所希望的结果奖励, 不希望的结果惩罚

算法介绍

全局智能体状态和动作的划分

此处状态划分不再以网格为依据，而是以威胁源作为划分的基准 。即每个威胁源所在区域 、 飞机出发点和目标点构成状态空间 。这样总的状态数量就是n + 2 个，其中 n 为威胁源数量 。图 1中所示是一个有两个威胁源的状态划分示意图 。图中威胁源周围的虚线框表示的威胁源的状态区域，比如威胁源 1虚线框内的区域都表示状态 2(状态 2的状态域)，图中前点表示状态域的进口，终点表示状态域的出口 。全局智能体的动作共有 n + 1个，这里的动作不是某一个具体的动作，而是指明下一步要转移到的状态 。

比如在上图中，在出发点选择动作 2, 表示下一步就是要向状态 2转移 。如果此动作可以执行，则转移到状态 2的状态域 。

局部智能体状态和动作的划分

局部智能体的任务主要是寻找能绕过威胁源的路径 。由于局部智能体处于状态域中，一般一个状态域中只有一个威胁源，可以按照网格的方法进行状态和动作的划分，即一个网格对应着一个状态 。当智能体选择动作后进入到威胁源时，奖赏 - 1, 反之如果更接近状态域的出口时，则奖赏 1。局部智能体状态划分如图 3所示。

为了能进一步减少状态，加快学习速度，在这里还可以使用相关强化学习[14]中状态和动作划分方法对状态和动作进行划分 。相关联强化学习同时把相关状态和动作进行表述 。强化学习提供了一个通用的框架和一组方法 。这一方法可以使智能体最优化它们在随机环境中的动作 。但是强化学习对状态和动作的描述使得它在应用到复杂的现实世界中时非常困难 。在很多领域中，状态和动作更多的是以相关联的形式表述的，比如在做饭的过程，在每道工序中采用什么动作是一定的，而不需进行更多的考虑 。

在这里把所有状态划分为 4个相关联状态 。前点和后点把整个威胁源划分为两个半圆，按照各个半圆中是否有其他威胁源接触进行相关联状态划分 。如图4所示 。

算法的描述本文中，采用了两个智能体，对路径规划任务进行划分：一个是全局智能体，负责全局路径的搜索；一个是局部智能体，负责状态区域内的路径搜索，找到绕过威胁源的路径 。每个智能体内部 Q 表的更新都采用标准的 Q 学习算法 。改进算法的过程伪码见表 1。

仿真结果及分析

这里先以 7个威胁源的地形图为例进行路径规划，然后威胁源的个数增加到 17个，分别对这两种情况进行路径规划，结果如图 5 ~图 8所示 。

> 7个威胁源时的参考路径

由图可知，这一路径并不是实际中最优的路径，强化学习在13 第 10期李东华等： 多智能体强化学习飞行路径规划算法 实际路径规划中寻找的并不是现实中的最优，而是强化学习定义的 “最优 ”。虽然这不是实际的最优但是已经很接近了 。 当然，可以通过定义新的奖惩函数的表达式，将强化学习定义的最优与通常概念下的最优联系起来 。

> 7个威胁源时的收敛速度

> 17个威胁源时的参考路径

> 17个威胁源时的收敛速度

总结

本方法使得飞行器能够完成给定的任务 、适应不同环境的要求，提高其生存率和任务完成率 。此算法中采用多智能体的方法，将研究对象进行了分层和抽象，解决了强化学习在复杂问题中维数灾难的困难 。实验仿真证明这一方法是可行的，并且具有较快的收敛速度

深度强化学习在变体飞行器自主外形优化中的应用

第 38 卷 第 11 期
2017 年 11 月

宇 航 学 报
Journal of Astronautics

Vol. 38 No. 11
November 2017

深度强化学习在变体飞行器自主外形优化中的应用

温 暖，刘正华，祝令谱，孙 扬
（北京航空航天大学自动化科学与电气工程学院，北京 100191）

摘 要：基于深度强化学习策略，研究了一类变体飞行器外形自主优化问题。以一种抽象化的变体飞行器为对象，给出其外形变化公式与最优外形函数等。结合深度学习与确定性策略梯度强化学习，设计深度确定性策略梯度（DDPG）学习步骤，使飞行器经过训练学习后具有较高的自主性和环境适应性，提高其在战场上的生存、应变和攻击能力。仿真结果表明，训练过程收敛较快，训练好的深度网络参数可以使飞行器在整个飞行任务过程中达到最优气动外形。

关键词：变体飞行器；深度强化学习；气动外形优化
中图分类号：V249.1 文献标识码：A 文章编号：1000-1328(2017)11-1153-07
DOI：10.3873/j.issn.1000-1328.2017.11.003

概述

基于深度强化学习策略，研究了一类变体飞行器外形自主优化问题。以一种抽象化的变体飞行器为对象，给出其外形变化公式与最优外形函数等。结合深度学习与确定性策略梯度强化学习，设计深度确定性策略 梯度(DDPG) 学习步骤，使飞行器经过训练学习后具有较高的自主性和环境适应性，提高其在战场上的生存、应变 和攻击能力。仿真结果表明，训练过程收敛较快，训练好的深度网络参数可以使飞行器在整个飞行任务过程中达 到最优气动外形。

关键词: 变体飞行器; 深度强化学习; 气动外形优化

变体飞行器外形模型

基于深度确定性策略梯度的变体飞行器外形优化学习

对于本文的变体飞行器，强化学习的目标是通 过大量的学习训练使飞行器对于特定的飞行状态 F 能够根据经验策略自主的控制电压 Vy 与 Vz ，从而 在整个飞行包线内处于最优的气动外形 Sy 和 Sz 。

考虑到上述动作空间的连续性问题，本文采用 的是强化学习中的确定性策略梯度算法以实现连续控制问题。针对单纯的确定性策略无法探索环境这 个缺陷，可以利用Actor-Critic(AC)学习框架实现异 策略学习方式，即行动策略与评估策略不是同一个 策略方法。行动策略为随机策略，以保证充足的探 索。而评估策略为确定性策略，其可以通过梯度计 算来实现累计奖赏 J 的最大化。

DDPG 的算法步骤

1) 随机初始化 Critic 深度神经网络 Q(s，a | θQ) 的权重θQ 和Actor的深度神经网络μ(s| θμ) 的权 重θμ。
2)初始目标网络Q- 与μ- 的权重θQ- 与θμ- 。
3) 初始化经验回放的缓存区 R 。
4) 重复每一幕。
5) 初始化随机过程 N 以用于行动策略的探索。
6) 初始观测得到状态 s1 。

7)重复步骤8) ~16)。

8) 根据当前的策略和随机探索选择动作: $a_t = \mu(s_t \mid \theta^{\mu}) + N_t$

9) 执行动作 a 从而得到奖励 r 和新的状态 st+1。

10) 将 (st，at，rt，st+1) 存储在缓存区 R 中。

11) 在 R 中随机选取一组数量为 M 的 (si，ai，ri，si+1) 。

12) 设定 $y_i = r_i + \gamma Q^{-}(s_{i+1}, \mu_{\theta^-}(s_{i+1} \mid \theta^{\mu^-}) \mid \theta^{Q^-})$

13)更新Critic的网络参数使得 $L = \frac{1}{N}\sum_{i=1}^{N}(y_i - Q(s_i, a_i \mid \theta^{Q}))^2$ 最小

14) 利用所选取样本的策略梯度更新 Actor 的网络参数

$$\nabla_{\theta^{\mu}} J = \frac{1}{M}\sum_{i=1}^{M}(\nabla_a Q^{\pi}(s, a \mid \theta^{Q})\mid_{s=s_i, a=\mu(s_i)} \nabla_{\theta^{\mu}}\mu(s \mid \theta^{\mu})\mid_{s=s_i})$$

15) 更新目标网络 $\begin{cases}\theta^{\mu^-} = \tau\theta^{\mu} + (1-\tau)\theta^{\mu^-} \\ \theta^{Q^-} = \tau\theta^{Q} + (1-\tau)\theta^{Q^-}\end{cases}$

16) 直到最大步数和最大幕数。

仿真校验

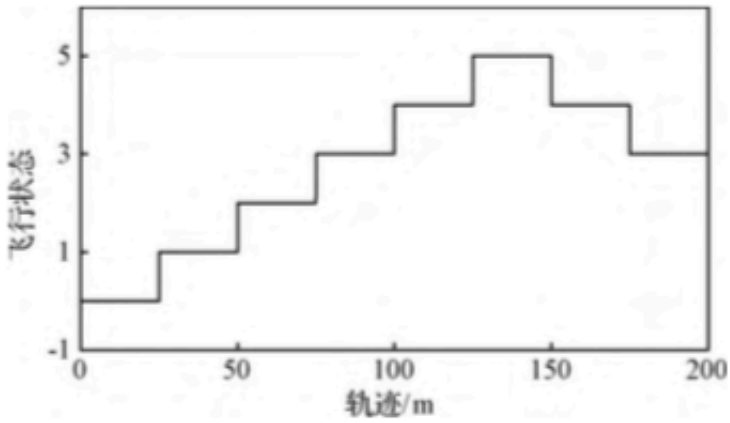

图5　飞行状态与飞行轨迹关系图

Fig. 5　Flight condition at various flight path locations

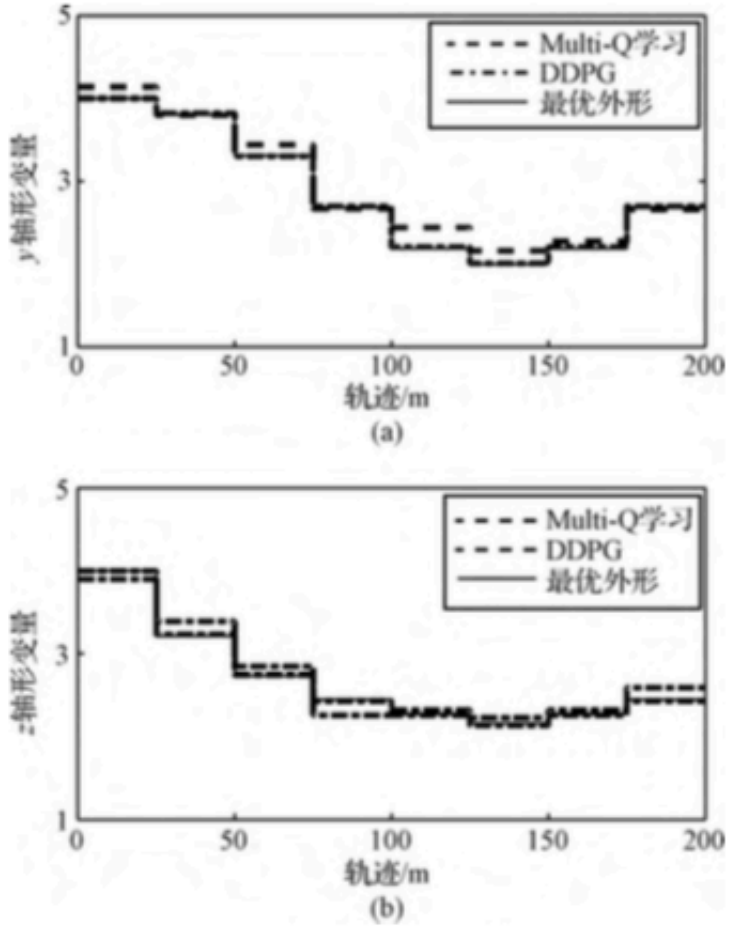

图6　最优外形与学习效果对比图

Fig. 6　Comparison between optimal shape and learned shape

总结

本文针对变体飞行器的外形优化问题，应用近 几年较为热门的深度强化学习算法使飞行器通过训 练学习具有了自主优化外形的能力，将人工智能方 法拓展到飞行器策略优化领域。为了解决传统的强 化学习框架不适用于连续控制这个问题，结合确定 性策略梯度算法与 Actor-Critic 框架进行强化学习 过程，并将深度神经网络替代原来传统的 Actor 函 数与 Critic 函数结构，以实现更好的学习效果。仿 真结果表明，整个学习过程收敛较快，并且利用训练 好的深度网络参数，可以使后期飞行过程中的外形 优化

效果大幅度提高。

飞行器强化学习多模在轨控制

2020 年 4 月 第 47 卷 第 2 期 西安电子科技大学学报 JOURNAL OF XIDIAN UNIVERSITY Apr. 2020 Vol. 47 No. 2

doi:10.19665/j.issn1001-2400.2020.02.011

飞行器强化学习多模在轨控制

张 英[1,2,3]，韦闽峰[2,3,4]，王世会[2,3]，陶磊岩[5]，
曹 健[1]，张 兴[1]

(1. 北京大学 软件与微电子学院，北京 100871；
2. 北京航天自动控制研究所，北京 100854；
3. 宇航智能控制技术国家级重点实验室，北京 100854；
4. 北京理工大学 自动化学院，北京 100081；
5. 北京遥感设备研究所，北京 100854)

摘要：为了提高飞行器控制系统长期在轨飞行的可靠性，提出了一种基于强化学习的多模式控制系统方案。该系统包括传感器模块、控制模块和执行模块。其中，传感器模块用于向控制模块实时输入飞行器敏感的飞行数据，该数据分为可供飞行器控制直接使用的具有历史相关性的多维结构化浮点数据以及某特定传感器独有的物理表征量；控制模块使用实时并行化决策机制，分为输入层、特征抽取层和全连接层；执行模块用于接收控制模块实时输出的驱动数据，包括用于决策的状态最优值和用于评价的动作输出值。系统根据用于决策的回报最优值决定使用哪些具体的执行模块，而某个被选定的具体执行模块的输出值取决于用于评价的动作输出值。该系统使飞行器在多模式输入输出状态下具备 15ms 快响应，5.23GOPs/sec/W(性能功耗比单位)性能功耗比的能力。

关键词：飞行器；控制系统；多模式；强化学习

中图分类号：TN 911.22 文献标识码：A 文章编号：1001-2400(2020)02-0075-08

概述

传感器模块用于向控制模块实时输入飞行器敏 感的飞行数据，该数据分为可供飞行器控制直接使用的具有历史相关性的多维结构化浮点数据以及某特定 传感器独有的物理表征量;控制模块使用实时并行化决策机制，分为输入层、特征抽取层和全连接层;执行模 块用于接收控制模块实时输出的驱动数据，包括用于决策的状态最优值和用于评价的动作输出值。系统根 据用于决策的回报最优值决定使用哪些具体的执行模块，而某个被选定的具体执行模块的输出值取决于 用 于 评 价 的 动 作 输 出 值 。

系统框架

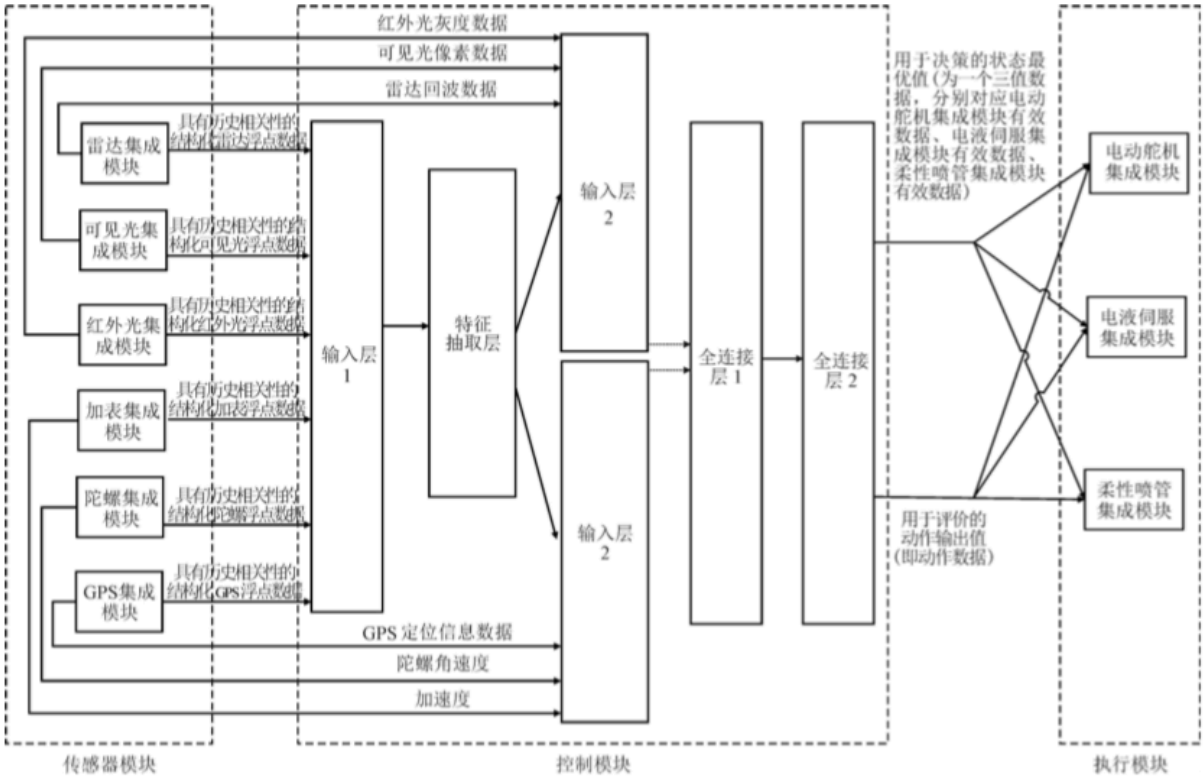

图 1　系统组成框图

网络结构:

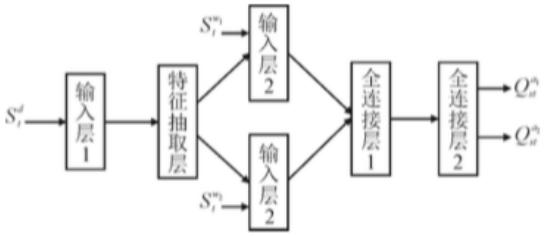

图 2　控制模块基于 Deep Q-Learning 网络结构图

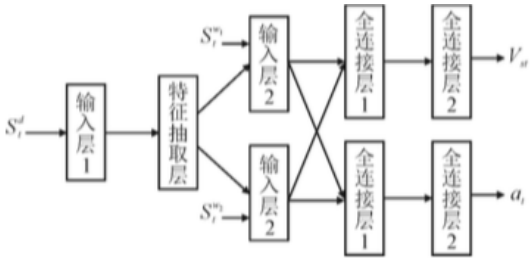

图 3　控制模块基于 A^{3}C 的网络结构图

仿真及结果

表 1　功耗及性能对比

内容	主服务器		FPGA2015[21]	FPGA2016[22]	ZynqNet[23]	文中
	CPU	GPU				
功率/W	69.00	142.00	18.61	25.8	12	1.65
效率/(GOPs/sec/W)	103.31×10^{-6}	64.70×10^{-6}	3.31	4.57	0.56	5.23
比例	1.97×10^{-5}	1.24×10^{-5}	0.63x	0.87	0.11x	1.00x

总 结

通过使用6路异构传感器模块、基于强化学习算法的控制模块和3路异构执行模块，完成了飞行器长期在轨多模式控制。通过使用可控的两个并行输入层和两个串行连接层结构，可准确实时判别多个模块健康 度，提前选择更优模块执行控制操作。飞行器多控制模型在不同场景下的控制效果不一样，因此控制模块的 强化学习算法可以使飞行器通过与环境的不断交互试错，自主学习动作策略，在多控制模型的飞行器决策方 法中选择较优模块，达到控制优化的决策目的。在控制器出现故障失效的情形下，强化学习

也可根据当前的 环境模型和状态空间感知出故障的发生，并且快速地做出决策。