Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2024, 13(1), 66-71
Published Online February 2024 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2024.131008

Hans汉斯

分层强化学习在无人机领域应用综述

杨永祥[1]，王念杰[2]，胡涵川[3]

[1]贵州师范大学数学科学学院，贵州 贵阳
[2]日照市岚山区行政审批服务局，山东 日照
[3]贵州师范大学大数据与科学学院，贵州 贵阳

收稿日期：2024年1月3日；录用日期：2024年2月23日；发布日期：2024年2月29日

摘　要

分层强化学习是强化学习领域的一个重要分支。基于分而治之的思想，将一个复杂问题分解成多个子问题，最终解决整个问题。近年来，由于传感器能力的提高和人工智能算法的进步，基于分层强化学习的无人机自主导航成为研究热点。本篇文章对国内外发表的具有代表性的文章进行概述，首先分析无人机和分层强化学习的含义，其次重点研究了分层强化学习在无人机轨迹规划和资源分配的优化问题上的应用。

关键词

分层强化学习，无人机，人工智能

A Review of the Application of Hierarchical Reinforcement Learning in the Field of Drones

Yongxiang Yang[1], Nianjie Wang[2], Hanchuan Hu[3]

[1]School of Mathematical Sciences, Guizhou Normal University, Guiyang Guizhou
[2]Rizhao Lanshan District Administrative Approval Service Bureau, Rizhao Shandong
[3]School of Big Data and Science, Guizhou Normal University, Guiyang Guizhou

Received: Jan. 3rd, 2024; accepted: Feb. 23rd, 2024; published: Feb. 29th, 2024

Abstract

Hierarchical reinforcement learning is an important branch in the field of reinforcement learning. Based on the idea of divide and conquer, a complex problem is decomposed into multiple

文章引用: 杨永祥, 王念杰, 胡涵川. 分层强化学习在无人机领域应用综述[J]. 人工智能与机器人研究, 2024, 13(1): 66-71. DOI: 10.12677/airr.2024.131008

sub-problems and finally the entire problem is solved. In recent years, due to the improvement of sensor capabilities and the advancement of artificial intelligence algorithms, autonomous drone navigation based on hierarchical reinforcement learning has become a research hotspot. This article provides an overview of representative articles published at home and abroad. First, it analyzes the meaning of UAVs and hierarchical reinforcement learning. Secondly, it focuses on the application of hierarchical reinforcement learning in UAV trajectory planning and resource allocation problems.

Keywords

Hierarchical Reinforcement Learning, Drone, Artificial Intelligence

1. 引言

强化学习在无人机领域开始广泛的适用。近年来，无人驾驶飞行器(以下简称无人机)因灵活性高、成本低等优点。在军事和民用领域得到了广泛的应用。包括探查和搜索、环境侦测、救援任务等。当无人机在处理具体问题时，例如传统的无人机导航方法需要通过构建精确的环境或依赖专家经验人为设置规则。或者作为通信中站传统的方法很难解决优化问题。不同于传统的方法，基于强化学习的方法在于它能够通过不断的试错和学习，优化无人机的决策策略，使其能够更好地适应不同的环境和任务。其次，强化学习可以使无人机具备自主学习和适应能力，而无需人工预先编程所有可能的情况和行为。这种自主学习能力使得无人机更灵活、智能，并能够应对未知或复杂的环境。特别是近年来深度强化学习取得了快速发展，利用深度学习强大的感知与拟合能力学习高维环境状态到控制动作之间的映射，从而能够获得更好的策略。

分层强化学习可以解决强化学习中稀疏奖励的问题。强化学习中，在一个复杂的环境中学习一项任务，其中稀疏奖励是一个问题，这对人工智能来说是一个重大挑战，与一般的优化固定结构系统参数的强化学习方法不同，分层强化学习通过优化系统结构来解决稀疏奖励的问题。如果能在多个时空抽象层次上表示所学知识，或者提供确切的奖励信号和及时的反馈，来指导智能体有效地朝着目标学习。在强化学习术语中，这就演变成多目标结构指导的分层学习过程，从而产生了分层强化学习的概念。因此，分层强化学习本质上迎合了稀疏奖励问题，并方便地适应了一类涉及多个任务的问题，这些问题具有良好的定义。和一般的优化固定结构系统参数的强化学习方法不同，分层强化学习通过优化系统结构来提高算法性能。

2. 无人机

近年来，无人机因灵活性高、成本低等优点。在军事和民用领域取得了广泛的应用。其独特的设计和卓越的性能，正在引领着航空领域的新潮流。与传统的飞行器相比，无人机不仅摆脱了人类驾驶员的限制，还通过先进的自主导航系统和高度灵敏的传感器技术，实现了更为灵活和精准的飞行。现在无人机已经广泛应用于各个领域，从军事侦察到商业航拍，再到紧急救援和科学研究。其多功能性使其成为执行多种任务的理想选择。与此同时，无人机的出现不仅提高了工作效率，降低了成本，还拓展了人类对于遥远或危险区域的探索能力。在过去的几十年内，无人机发展迅速，并且取得了实质性的突破。

3. 分层强化学习理论基础

3.1. 强化学习理论基础

强化学习是机器学习领域中的重要分支，其目的是智能体在与环境的交互中学习如何获得最大奖励。强化学习是除了监督学习和非监督学习之外的第三种的机器学习方法。而与监督学习不同的是，强化学习不需要任何的带标签的输入输出，而是通过与环境的交互来积累经验。目前大部分强化学习研究都是建立在马尔可夫决策过程理论[1]的基础上。使用四元组 $\langle S, A, P, R\rangle$ 来定义控制过程。其中 S 是环境中的状态空间。某个状态 $s \in S$ 表示环境中的一些情况。A 是智能体可以采取的一组动作。动作对状态的影响由状态转移函数 $P\left(s_{t+1} \middle| s_t, a_t\right)$ 来表述。R 是奖励函数。策略 π 表状态到动作的映射函数 $\pi: S \to A$，当给定一个策略 π 时，在每个时间步数 t 时，智能体会观察当前状态 $s_t \in S$ 并根据策略 π 执行动作 $a_t \in A$。从而获得奖励 $r_{t+1} = R\left(s_t; a_t\right)$。重复执行下去直到任务结束就会得到一个序列。对于确定性环境，我们直接可以用累计奖励 G_t 作为策略是否好坏的评判标准。然而由于实际情况的不同，状态转移概率 p 的存在，环境往往具有不确定性。智能体从初始状态到任务结束状态可能会存在很多条序列。在这种情况下，我们使用回报的期望值作为策略好坏的评判标准。策略 π 所得到的期望回报为状态价值函数：

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \middle| s_0 = s\right] \tag{1}$$

若智能体在状态 s 处执行动作 a，而后根据策略 π 所得到的期望回报为状态动作价值函数：

$$Q^{\pi}\left(s, a\right) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \middle| s_0 = s, a_0 = a\right] \tag{2}$$

强化学习的核心目标就是学习满足以下最优策略 π^*：

$$\pi^* = \arg\max_{\pi} Q^{\pi}\left(s, a\right), \forall s \in S, \forall a \in A \tag{3}$$

当处理MDP问题时，智能体的动作都是在单位时间内完成的，Sutton等人[2]提出了跨时序性的SMDP理论。在SMDP下，智能体可以采取序列动作，构建分层架构。通过单次执行多步动作从而到达更远的状态。

3.2. 分层强化学习

从上述公式可以看出，强化学习的目标就是寻找一个最优策略，使得智能体根据遵循该策略执行的各种可能的序列的累计奖励期望最大化。然而，当状态空间和动作空间都很大，序列范围很长时，使用标准的强化学习在高维连续的状态空间做出决策导致训练的复杂性和低效性。很难取得理想的效果。分层强化学习出现的原因主要是为了应对复杂任务中的探索和决策问题。通过在决策过程中引入多个层次的控制策略，从而提高学习的数据效率和性能减少样本数。这种方法的目标是通过将任务分解成更简单的子任务，并对每个子任务使用单独的控制层次，从而简化整个学习问题。使得智能体能够更高效地学习和执行复杂任务，提高了学习的速度和性能。

3.3. 经典的分层强化学习架构

分层强化学习提供了一种方法，将复杂的强化学习问题进行分解，将复杂性的问题分解为更简单的子任务来执行任务。在这样的层次结构中，高层策略通常被用来选择子任务或者子目标。将其训练为根据子任务的顺序执行。在层次结构的较低层次上，由较高层次选择的子任务本身就是一个强化学习问题。低层策略学习使用与其相关的内部奖励来执行子任务(也可以添加主任务奖励)。以下介绍三种比较经典的

分层强化学习架构。包括 MAXQ、Options 和 Feudal。

MAXQ：Dietterich 在 2000 年提出一种分层强化学习方法 MAXQ 值函数分解[3]。其目的是提高分层强化学习的效率。MAXQ 通过将一个大任务分解成一系列更小的子任务，将给定的任务 M 分解为一组子任务{M0，M1，M2，…，Mn}。将，将 M0 作为根子任务。解决 M0 意味着也就解决了 M。每个子任务有自己的学习目标，来实现这一目标。学习任务形成以 M0 为根任务的层次结构。从下到上逐层解决子任务，直到解决根任务。分层任务模型中的节点不是按顺序排列的。相反，由上一层的任务决定下一层的哪个子任务应该首先执行。MAXQ 通过解决任务中的所有子任务来完成任务。由于子任务需要较少的状态动作空间，因此可以快速求解。

Options：Sutton 等人[4]在 1999 年基于时序抽象法重新定义了智能体得基本动作并提出了 option 架构。Options Framework 实现了层次结构和宏动作的方法。宏动作是指持续一段时间的动作序列，可以看作是高层次的抽象动作。从而使智能体能够更灵活地应对各种情境。Option 框架一般由两个层组成：高层策略根据当前状态选择某一个选项，然后按照这一选项的策略 π 选择一个动作或者其他选项，执行动作或者进入新的 option，继续循环选择或者终止。

FeUdal：Dayan & Hinton [4]在 1992 年提出一种封建等级制度。被称为封建强化学习。封建强化学习的大致描述为：模型分为管理者和工作者，管理者设置一个子任务，子任务将由工作者完成。受封建强化学习的启发，Vezhnevets [5]提出封建分层网络。其中称为“管理者”的高级网络在学习的潜在子目标空间中采样子目标。子目标可以是潜在空间中的一个点，也可以是代表潜在空间方向的单位向量。子目标被称为“工作者”的低级网络作为输入，它必须学习一种策略来实现子目标，使用到子目标的距离作为奖励。

4. 分层强化学习在无人机领域中的应用

4.1. 轨迹规划

无人机轨迹是无人机导航中的核心问题。目的是寻找从初始点到目标点最优的路径。除了寻找最短的路径之外，还需要确保无人机在飞行过程中避免碰撞。基于计算的轨迹规划有助于引导无人机规避障碍，快速抵达目的地。

目前比较普遍的算法是群体智能算法。虽然在轨迹跟踪方面表现良好，但是需要无人机飞行过程的全局环境。在处理比较复杂的环境中，群体智能算法在轨迹规划中表现不佳。如何在复杂的环境中进行路径规划是当前领域的热点问题之一。近年来，许多学者将机器学习引入轨迹规划问题中，试图通过机器学习的角度来探索更好的解决方案。大量的研究表明，深度学习和强化学习可以应对复杂多变的环境。深度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“维数灾难问题和多智能体系统的非马尔可夫问题。越来越多的研究者把注意力集中到分层强化学习。分层强化学习将复杂的决策任务分解为一系列连续的简单子目标，解决了强化学习的维数突变问题。下面介绍一下研究者为了解决轨迹跟踪这个问题。将分层强化学习应用到无人机中。

Kouk [6]提出一种基于期权的分层强化学习模型架构。并设计了高层策略模型和低层策略模型两层架构。高层策略模型为行为选择模型，低层策略模型由两个模型组成，分别是避障控制模型和目标驱动模型。行为选择模型以视觉传感器所采集的图像数据作为输入，学习并输出两个离散动作和，分别对应两种模型:避障控制模型和目标逼近模型，其中用数值较高的动作激活相应的行为策略。从而实现移动机器人行为的自动选择。仿真结果表明，该方法能有效解决复杂环境下自主控制策略效果差的问题。

ALP [7]提出了一种基于封建的分层强化学习模型架构。并且设计了由元控制器和控制器组成的两层架构。元控制器使用多项逻辑回归模型来充当管理者，并且使用状态转移图来进行训练。控制器使用了

DQN 模型。需要注意的是，元控制器目标必须预先指定，虽然训练过程带来了更好的性能，但限制了方案的适用性。

与单无人机轨迹规划相比，多无人机轨迹跟踪不光要考虑路径最优的问题，还要考虑协同飞行的问题。在飞行过程中，成群的无人机在飞行过程中不相互影响。基于多无人机协同飞行的问题。程[8]将MAXQ 分层强化学习方法用到多无人机轨迹规划中，在 MAXQ 的结构基础上提出了一种分层的由下而上的多无人机学习模型。将基本动作分为左转、右转、直行、加速、减速、爬升和下降七个。Cheng [9]在 MAXQ 方法的基础上，提出基于模拟退火的 MAXQ 分层强化学习(SA-MAXQ)算法。仿真结果表明，基于 SA-MAXQ 算法可以避免路径规划陷入局部最优解。

4.2. 资源分配

随着无线通信技术和移动计算的发高速发展，移动设备和移动应用越来越普及。很多当下的应用对算力都是要求比较高的。普通用户的移动设备往往受限于算例，导致效率低下。同时由于云计算难以满足的低延迟需求，移动边缘计算(MEC)已经成为一种有效的补充手段。得益于无人机的优点，无人机作为一种被提出并设想的新技术，来协助军事和民用应用中的移动边缘计算。举例来说，当网络基础设施无法使用(灾难救助情况下)或者设备端数量超过网络服务能力(重大比赛直播)。无人机可以作为通信中站或者边缘计算平台。无人机辅助移动边缘计算可以获得许多好处，例如减少网络开销，降低执行延迟，提高体验质量等。尽管有许多优点，但在无人机辅助 MEC 计算是具有挑战性的，因为合理安排无人机的飞行轨迹和任务负载。现有的研究大多集中在静态环境下无人机辅助 MEC 调度的研究。最近，一些研究试图通过强化学习制定调度策略，在动态环境下实现无人机辅助的实时 MEC。强化学习探索 MEC 的动态环境，并从经验中学习有效的调度策略，从而解决优化问题。然而，当无人机或移动设备数量增加时，MEC 的系统状态空间和动作空间将呈指数级增长。极大提高了任务学习的复杂度。因此，考虑用分层强化学习的角度来分析。为了解决此问题，Ren [10]提出 HT30 架构用于大规模无人机辅助 MEC 的可扩展调度方法。将优化问题分层分解为两个子问题(位置优化问题和和负载优化问题)。两个子问题分别对应着两个优化器。在任务执行阶段，位置优化器在外部阶段负责位置处理位置信息。卸载优化器在内部阶段确定卸载调度变量，以使所有移动设备的平均延迟最小。

随着物联网的发展，人类部署了大量的传感器节点来监测环境，然而，传感器电池更换和充电费用很高，维持如此大规模的网络是相当困难的。新兴的反向散射通信技术被认为是解决物联网设备电池问题的一种有前途的解决方案(功耗)。该技术使传感器能够通过反射射频信号来传输数据。与传统的无线传感器网络相比，反向散射功耗极低。遗憾的是，反向散射传感器节点的传输范围不是很远，这给数据采集带来了困难。为了克服传感器通信距离短的缺点，采用灵活易控制的无人机辅助数据采集并用无人机作为无线传感器网络的移动数据采集器是延长网络寿命的一种节能方式。为了解决可充电的无人机场景下的数据收集问题，Zhang [11]提出了种基于选项的分层强化学习方法来解决该问题。将子任务分为收集数据、飞到充电站充电、飞到起点告知任务完成等。通过选择选项来优化问题。其中无人机可以飞到靠近后向散射传感器节点(BSN)的地方激活它，然后收集数据，在收集任务完成后，使可充电无人机的总飞行时间最小。在数据采集过程中，当无人机能量不足以完成任务时，无人机可以返回充电站自行充电。

5. 结语

综上所述，常用的深度强化学习方法可能无法解决无人机的复杂问题。分层强化学习方法引入抽象机制实现状态空间降维，通过学习在不同级别的时间抽象上进行操作来解决“维数灾难”的问题，可以明显的看出分层强化学习可以解决无人机资源分配和轨迹规划的问题。但是目前实际应用可能还存在以

下方面的问题。

1. 需要计算资源的需求、模型的泛化能力等。分层强化学习在未来可能会出现技术创新，例如算法优化、学习效率的提高等来提高效率。

2. 现阶段的绝大多数分层强化学习算法都需要人为的设计分层。而人为的设计分层很难发现其内在联系。因此自动学习层次结构将是分层强化学习的重要发展趋势。

3. 分层强化学习是否可以与其他领域(如人工智能、机器人技术、物联网等)的融合？或许融合会带来新的解决方案和改善现有技术。

参考文献

[1] Bellman, R. (1954) The Theory of Dynamic Programming. *Bulletin of the American Mathematical Society*, **60**, 503-515. https://doi.org/10.1090/S0002-9904-1954-09848-8

[2] Sutton, R.S., Precup, D. and Singh, S. (1999) Between MDPs and Semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning. *Artificial Intelligence*, **112**, 181-211. https://doi.org/10.1016/S0004-3702(99)00052-1

[3] Dietterich, T.G. (2000) Hierarchical Reinforcement Learning with the MAXQ Value Function Decomposition. *Journal of Artificial Intelligence Research*, **13**, 227-303. https://doi.org/10.1613/jair.639

[4] Dayan, P. and Hinton, G.E. (1992) Feudal Reinforcement Learning. *Advances in Neural Information Processing Systems*, **5**, 272-278.

[5] Vezhnevets, A.S., Osindero, S., Schaul, T., *et al*. (2017) Feudal Networks for Hierarchical Reinforcement Learning. *International Conference on Machine Learning*, Sydney, 6 August 2017, 3540-3549.

[6] Kou, K., Yang, G., Zhang, W., *et al*. (2022) Autonomous Navigation of UAV in Dynamic Unstructured Environments via Hierarchical Reinforcement Learning. 2022 *International Conference on Automation, Robotics and Computer Engineering* (*ICARCE*), Wuhan, 16-17 December 2022, 1-5. https://doi.org/10.1109/ICARCE55724.2022.10046655

[7] Alpdemir, M.N. (2023) A Hierarchical Reinforcement Learning Framework for UAV Path Planning in Tactical Environments. *Turkish Journal of Science and Technology*, **18**, 243-259. https://doi.org/10.55525/tjst.1219845

[8] 程先峰, 严勇杰. 基于 MAXQ 分层强化学习的有人机/无人机协同路径规划研究[J]. 信息化研究, 2020, 46(1): 13-19. https://doi.org/CNKI:SUN:DZGS.0.2020-01-004

[9] Cheng, Y., Li, D., Wong, W.E., *et al*. (2022) Multi-UAV Collaborative Path Planning Using Hierarchical Reinforcement Learning and Simulated Annealing. *International Journal of Performability Engineering*, **18**, 463-474. https://doi.org/10.23940/ijpe.22.07.p1.463474

[10] Ren, T., Niu, J., Dai, B., *et al*. (2021) Enabling Efficient Scheduling in Large-Scale UAV-Assisted Mobile-Edge Computing via Hierarchical Reinforcement Learning. *IEEE Internet of Things Journal*, **9**, 7095-7109. https://doi.org/10.1109/JIOT.2021.3071531

[11] Zhang, Y., Mou, Z, Gao, F., *et al*. (2020) Hierarchical Deep Reinforcement Learning for Backscattering Data Collection with Multiple UAVs. *IEEE Internet of Things Journal*, **8**, 3786-3800. https://doi.org/10.1109/JIOT.2020.3024666