引用格式：王天孜, 史红权, 张思洁, 等. 基于次优策略的动态分步强化学习路径规划算法 [J]. 电子科技大学学报, 2025, 54(5): 671-675.
WANG T Z, SHI H Q, ZHANG S J, et al. Dynamic stepwise reinforcement learning path planning algorithm based on suboptimal policy[J]. Journal of University of Electronic Science and Technology of China, 2025, 54(5): 671-675.

基于次优策略的动态分步强化学习路径规划算法

王天孜，史红权*，张思洁，陈爱国

（电子科技大学 计算机科学与工程学院，成都 611731）

摘要：强化学习允许智能体在未知环境中进行路径规划，智能体能够使用与环境交互所得到的信息自主调整策略并找到最优路径。然而，大多数基于强化学习的路径规划任务都面临着稀疏奖励的问题，即获取外部奖励以及有效训练样本的难度大，这使得算法迭代缓慢，甚至难以收敛。为此，提出了一种基于次优策略的动态分步强化学习路径规划算法，该算法通过动态分步的方法将次优策略引入到强化学习框架下，并设计内在奖励鼓励智能体探索优于次优策略的新策略。实验结果表明，与基线算法相比，该算法有着更好的表现，智能体系统获得的奖励更高，策略收敛速度更快。

关键词：动态分步；路径规划；强化学习；次优策略

中图分类号：TP391 **文献标志码：**A DOI: 10.12178/1001-0548.2024034

Dynamic stepwise reinforcement learning path planning algorithm based on suboptimal policy

WANG Tianzi, SHI Hongquan*, ZHANG Sijie, and CHEN Aiguo

(School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)

Abstract: Reinforcement learning equips agents with the capability to successfully complete path planning task in an unknown environment, where the agent uses the information derived from its interaction with the environment to autonomously adjust its policy and find the optimal path. However, most path planning tasks face the problem of sparse rewards. Within path planning tasks characterized by these sparse rewards, the process of obtaining external rewards and acquiring valid training data becomes notably challenging, which makes the algorithm iterate slowly and even difficult to converge. To this end, a dynamic stepwise reinforcement learning path planning algorithm based on suboptimal policy is proposed. The suboptimal policy is introduced into the reinforcement learning framework through dynamic stepwise methods and the designs intrinsic reward is designed to encourage agent to explore better than suboptimal ones. Experimental results show that proposed algorithm has better performance compared with the baseline algorithm. The agent obtains higher rewards, and the policy convergence speed is faster.

Key words: dynamic stepwise; path planning; reinforcement learning; suboptimal policy

随着计算机技术、人工智能技术及自动化控制技术的不断发展，移动机器人的智能化程度不断提高，路径规划作为实现移动机器人自主导航的一项关键技术备受关注[1]，越来越多的研究人员致力于优化路径规划算法[2]。强化学习是一种与环境交互，利用交互所得的信息寻找最优策略的方法，与传统路径规划算法相比，强化学习允许智能体在未知环境中进行路径规划，因此将强化学习用于路径规划问题成为目前移动机器人路径规划的研究热点[3]。

在基于强化学习的路径规划任务中，算法的收敛依赖于学习大量奖励可变的经验数据。对于较复杂的路径规划任务问题，通常只有当智能体到达目的地后才能获得正向奖励，因此在收集的经验数据中，拥有有效奖励的经验数据占比很少，稀疏的奖励信号会导致算法训练缓慢甚至失败[4]。

收稿日期：2024−02−20
基金项目：国家自然科学基金（U19A2059）；四川省科技计划（206999977）
作者简介：王天孜，主要从事计算机软件与理论方面的研究。
*通信作者 E-mail: shihongquan72@sina.com

引入先验策略是解决稀疏奖励问题的方式之一[5]，即让智能体对环境信息建立一个初步的判断，再利用强化学习的探索能力不断改善策略。文献 [6] 使用基于领域知识的专家先验策略解决稀疏奖励问题，但在实际应用中，收集专家先验策略成本高昂，代价大。文献 [7] 使用次优先验策略引导智能体探索，有效缓解了稀疏奖励问题，但次优策略中含有一些不良信息，盲目使用会使智能体的学习陷入天花板效应[8]，即策略网络收敛到局部最优解。

针对以上问题，本文提出了一种基于次优策略的动态分步强化学习路径规划算法（dynamic stepwise reinforcement learning path planning algorithm based on suboptimal policy, DSQSP）。首先，将次优策略转为一系列带有标签的里程碑。其次，为每个里程碑赋予特定的奖励值，当智能体到达里程碑时会获得对应的奖励，该奖励值能够鼓励智能体探索优于初始次优策略的新策略。在智能体第一次到达目的地后，丢弃只在初始次优策略上的里程碑，并在新策略上新增一定数量的里程碑。最后，设计了两个路径规划任务进行实验。实验结果表明，本文提出的算法能完成复杂场景的路径规划任务，并具有较高的学习效率。

1 相关工作

过去的许多研究都是基于已知环境进行路径规划[9]，而基于强化学习的路径规划算法能够在复杂的未知环境中实现有效避障，因此强化学习在路径规划中的使用越来越广泛。如文献 [10] 将强化学习算法应用到路径规划中，无须知道环境模型就可以保证收敛。文献 [11] 提出一种基于深度强化学习的机械手动态避障路径规划方法，能够有效避开环境中的移动障碍物。

在强化学习中，奖励起着引导智能体学习的作用。在越来越复杂的实际场景中，智能体难以获得有效的奖励信号，稀疏的奖励信号导致收敛缓慢甚至难以收敛。引入外部引导信息是解决稀疏奖励的方法之一，许多研究通过将次优策略引入强化学习中来解决稀疏奖励问题[12]。如文献 [13] 提出的 DQfD（deep Q-learning from demonstrations）算法用先验策略对 Q 函数进行预训练，缓解了智能体的盲目探索。文献 [14] 提出的 T-REX（trajectory-ranked reward extrapolation）算法能够从先验策略中学习到高质量奖励函数，从而引导智能体的学习。文献 [15] 提出的 JSRL（jump-start reinforcement learning）算法使用先验策略创建探索策略的初始状态，利用课程学习的思想，将稀疏奖励任务分解成容易实现的子任务。但是，次优策略中包含大量干扰信息，这些干扰信息会给智能体带来错误的指导，降低智能体的学习上限，导致使用次优策略的智能体陷入“天花板”效应。尽管强化学习的随机特性使“天花板”的上限有可能会提高，但提高程度有限。

为此，一些学者开始研究如何打破次优策略带来的“天花板”问题。文献 [16] 提出了融合多次优策略的强化学习路径规划算法（multiple suboptimal policies integrated reinforcement learning algorithm for path planning, SPQRP），该算法把次优策略转换为内在奖励，能够引导智能体在决策的同时跳出次优策略中的错误部分，有效缓解了“天花板”效应，但该算法并没有充分利用次优策略，学习效率有待提高。

2 基于次优策略的动态分步强化学习路径规划算法

本文提出了一种分步使用次优策略的强化学习路径规划算法，算法框架如图 1 所示。该算法使用次优策略构建可以提供内在奖励的子目标，这些子目标为里程碑。在训练开始之前，利用次优策略得到初始次优路径，在初始次优路径上设置可以提供内在奖励的里程碑，用于引导智能体探索。在第一次探索到新路径后，保留既在新路径又在初始次优路径上的里程碑，并在新路径上增设一定数量的里程碑，提高次优策略的利用效率。

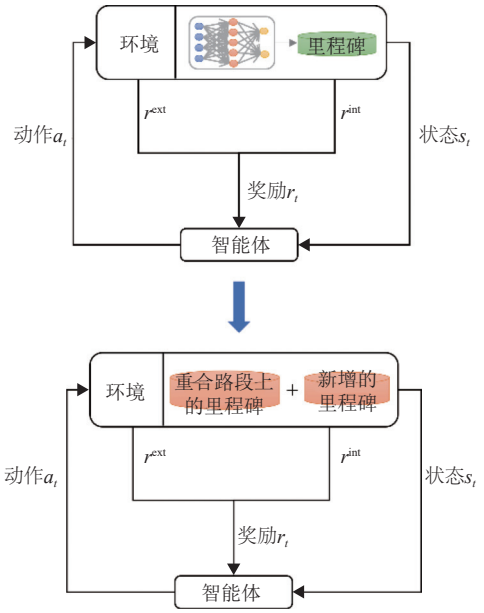


图 1 算法框架

将次优策略引导下的路径规划任务定义为马尔可夫决策过程模型$G=<S,A,R^{\text{ext}},R^{\text{int}},t>$，其中$S$是智能体在规定任务中所有状态的集合，$A$是智能体执行当前任务所有可能的动作，$R^{\text{ext}}$是来自环境的外部奖励，$R^{\text{int}}$是内部奖励集合，即每个里程碑所对应的奖励值的集合。在时刻t，智能体得到当前的环境状态$s_t\in S$，依据策略选择动作$a_t\in A$，根据状态–动作对$<s_t,a_t>$，智能体收到一个数值化收益$r=r^{\text{ext}}+r^{\text{int}}\left(r^{\text{ext}}\in R^{\text{ext}},r^{\text{int}}\in R^{\text{int}}\right)$，并进入一个新的状态$s_{t+1}\in S$。

2.1　分步设置里程碑

使用次优策略得到一条从起点到终点的先验次优路径l_{init}，之后分两步设置里程碑。

在智能体学习开始之前，除了先验次优路径的位置信息之外没有任何其他附加信息，因此在l_{init}上等长地设置一定数量的里程碑。即从起点开始，每隔d步设置一个里程碑，重复该操作直至达到l_{init}的终点，得到里程碑序列$M=[m_1,m_2,\cdots,m_n]$，其中里程碑标签单调递增。

当智能体第一次到达目的地时，将其轨迹标记为l_{first}。SPQRP 算法会鼓励智能体探索而不是简单地克隆[16]，其内在奖励会根据当前状态和初始状态之间智能体移动的总步数设置折扣，这意味着它会引导智能体探索出一条优于先验次优路径的新路径，即新路径l_{first}优于初始次优路径l_{init}。接下来找到l_{init}与l_{first}重合的路段，重合路段上的里程碑对智能体学习的贡献度较大，保留这部分里程碑，丢弃只在l_{init}而不在l_{first}上的里程碑。除此之外，额外再增加一些里程碑以引导智能体进行后续的探索与学习。

在学习过程中重新设置一次里程碑意味着在训练过程中修改了之前定义的马尔可夫决策过程模型$G=<S,A,R^{\text{ext}},R^{\text{int}},t>$，这样做可能会造成一定的模型偏移。如果不对里程碑的新增方式以及数量进行严格限制，那么修改后的里程碑可能会降低智能体的学习效率，甚至导致学习失败。考虑到里程碑对应内在奖励r^{int}的大小与标签值有很大关系，为了缓解模型偏移问题，规定增设里程碑时必须保证留下的里程碑标签不变。沿着l_{first}在标签为p、q的里程碑之间等长地增加$\lfloor\frac{q-p-1}{2}\rfloor$个里程碑，每个里程碑的标签为$p+2k$，其中$k=1,2,\cdots,\lfloor\frac{q-p-1}{2}\rfloor$。

如初始里程碑序列为$[m_1,m_2,\cdots,m_n]$，其中$n=7$。在智能体探索到外部奖励后，保留重合路段上的里程碑，得到保留里程碑序列（假设该序列是$[m_1, m_3, m_6]$）。之后，按照上述规则增设里程碑，得到一个新的里程碑序列$[m_1, m_3, m_5, m_6]$，该序列能够更有效地引导智能体学习。

2.2　内在奖励

内在奖励的设置旨在引导智能体放弃不完美的次优路径，不断学习与强化探索到的更优路径。由于给每个里程碑设置了标签，并且同一条次优路径上里程碑的标签是单调递增的，因此次优路径上的里程碑标签值越大，里程碑就越靠近最终的学习目标，应该给予更多的内在奖励用于引导。里程碑的基本奖励为r^{basic}，那么在先验路径上的第j个里程碑的基本奖励为：

$$r_j^{\text{basic}}=\lambda j \tag{1}$$

式中，r_j^{basic}表示先验路径上第j个里程碑的基本奖励；λ是基本奖励r_j^{basic}的权重系数。

当智能体跳过一些不完美的里程碑路段时，里程碑标签会实现一个飞跃，为了鼓励智能体的这种行为，需要给它比基础奖励更多的内在奖励，本文称这种内在奖励为飞跃奖励。同时，为了防止智能体陷入局部最优，规定智能体不能重复获取里程碑的奖励。当智能体到达标签为j的里程碑时，它得到的内在奖励为：

$$r^{\text{init}}=\sum_{i=j-h}^{j}\gamma^k r_i^{\text{basic}}\qquad k=h,h-1,\cdots,0 \tag{2}$$

式中，i是当前可达里程碑的最小标签值；h是智能体跃过的里程碑个数；γ是折扣因子。为了防止智能体“返回”去获取之前未获取的内在奖励，本文规定，一旦智能体获得某个里程碑的内在奖励，该里程碑之前的内在奖励将置 0。

2.3　DSQSP 算法

本文提出的基于次优策略的动态分步强化学习路径规划算法的执行过程如下。

初始化 Q 网络以及里程碑序列M，计算出每个里程碑对应的内在奖励值r^{int}；

for 训练轮数episode $=0,1,\cdots,n$ do

　初始化状态s_0；

　for 时间步$t=0,1,\cdots,m$ do

　　在当前状态s_t下，根据$\varepsilon-$greedy策略获取动作a_t；

　　执行动作a_t，获取奖励r_t和s_{t+1}；

　　将(s_t,a_t,r_t,s_{t+1})存入经验池中；

从经验池中随机取出一批样本(s_i, a_i, r_i, s_{i+1})更新 Q 网络；

end for

if 智能体第一次获得外部奖励r^{ext} do

更新里程碑序列M；

更新里程碑对应的奖励值r^{int}；

end if

end for

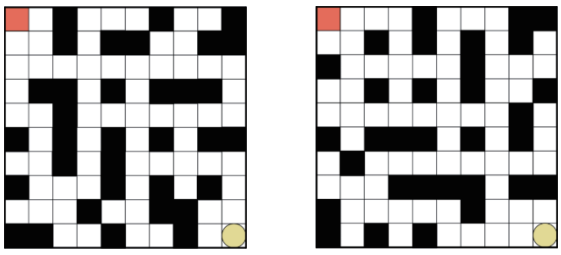
该算法的时间复杂度主要受训练轮数n和每轮的时间步数m的影响。具体来说，在每个训练轮次中，内层循环的时间复杂度为$O(mB)$，其中B是批大小，外层循环的时间复杂度为$O(n)$，因此总的时间复杂度为$O(mn)$，B是常数可以忽略不计。

3 实验结果及分析

本节通过设计对比实验来验证改进后的 DSQSP 算法可有效应用在不同的迷宫环境中，能够有效破除次优策略带来的天花板效应，并且与 SPQRP 算法相比，该算法能提高次优策略的利用效率，从而提高智能体的学习效率。

3.1 实验设计

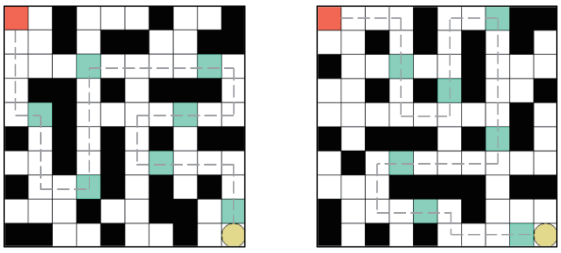
使用 tkinter 库构建迷宫环境，设计图 2 所示的迷宫 A 与迷宫 B 两个场景。在迷宫环境中，红色方块代表智能体，它需要通过上、下、左、右 4 个方向的移动成功到达黄色实心圆所表示的终点或者达到最大探索次数。智能体到达终点会获取 1 000 的外部奖励，该外部奖励的大小会随着智能体的运动线性递减，这一设置的目的是让智能体能够找到从起点到终点的最短可行路径。实验开始时，外部奖励为 1 000，该奖励的大小会随着时间步 (timesteps) 线性递减 ($r^{\text{ext}} = \max(1000 - 10\times \text{timesteps}, 0)$)，这一设置的目的是让智能体能够找到从起点到终点的最短可行路径。黑色方块表示的障碍物不可通行，如果智能体碰到这些方块，一轮学习将会结束，智能体回到初始位置开始新一轮的学习。



a. 迷宫A b. 迷宫B

图 2 迷宫环境

分别在两个迷宫中随机选取次优路径，得到图 3 所示的示意图，其中灰色虚线就是次优路径，在两个环境中对比 SPQRP 算法、DSQSP 算法与 DQfD 算法。由于 SPQRP 算法在里程碑设置间隔 $d = 4$时实验效果较好，因此在本实验的初始化过程中，每隔 4 步就在次优路径上选取一个里程碑。图中的绿色方块是里程碑，智能体到达这些状态会获取内在奖励。内在奖励旨在有效引导智能体探索，其与外部奖励的比例不宜过小，过小会导致引导效果不明显；也不宜超过 1，超过 1 会使智能体偏离最终目标。实验发现，当内在奖励的最大值与外在奖励的比例为 1∶10 时，效果较为理想。本文设置 $\lambda = 3$，$\gamma = 0.9$，以保证内在奖励的最大值与外在奖励的比例约为 1∶10。



a. 迷宫A b. 迷宫B

图 3 加入次优路径的迷宫环境

3.2 实验结果及分析

迷宫 A 与迷宫 B 的实验结果分别如图 4 和图 5 所示。

可以看出，在同一个迷宫中使用 SPQRP 算法与使用 DSQSP 算法的智能体最开始获得 reward 的轮次基本一致，这是因为两种算法在智能体学习开始之前设置里程碑的方式一样，也就是说在智能体第一次到达目的地之前的探索过程中，两种算法对智能体的引导是完全相同的。但在之后的学习中，绿色曲线代表的 DSQSP 算法的学习效率明显高于橙色曲线代表的 SPQRP 算法，更快地达到了收敛状态。这是因为在智能体第一次到达目的地后，算法留下了l_{init}与l_{first}重合路段上的里程碑，抛弃了其他不重要的里程碑，并在不打乱标签单调递增规则的情况下增加了一定数量的里程碑。此外，分析图 4 和图 5 可知，DQfD 算法只能围绕先验路径进行学习，而本文提出的算法鼓励智能体探索优于先验路径的新路径，打破了天花板效应。同时，根据实验结果图得知，本文提出的 DSQSP 算法在不同的迷宫场景中均有效。

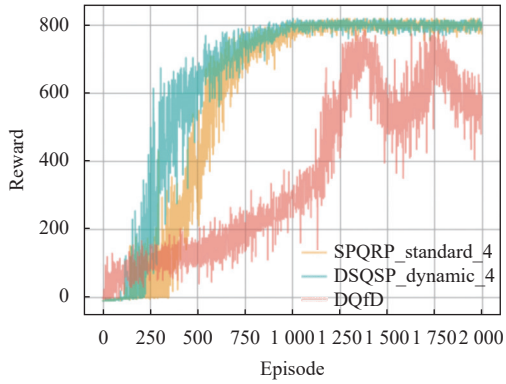


图 4　DSQSP 算法与 SPQRP 算法、DQfD 算法在迷宫 *A* 中的对比结果

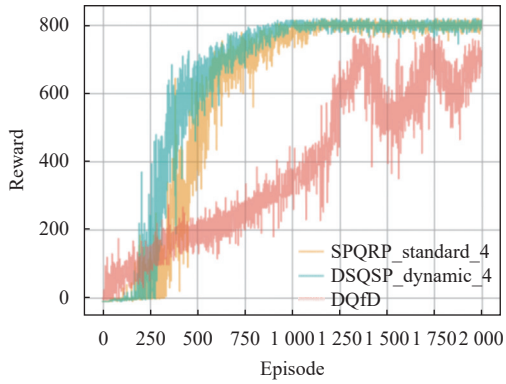


图 5　DSQSP 算法与 SPQRP 算法、DQfD 算法在迷宫 *B* 中的对比结果

4　结束语

本文提出了一种基于次优策略的动态分步强化学习路径规划算法，通过使用次优策略来解决复杂路径规划任务存在的稀疏奖励问题。该算法将次优策略转化为可以提供内在奖励的里程碑，并在训练过程中逐步丢弃较差的里程碑。通过这种方式，智能体可以学习到优于次优策略的新策略。在相同的场景下，与 DQfD 算法和 SPQRP 算法相比，基于次优策略的动态分步强化学习路径规划算法有着更好的表现。

参 考 文 献

[1] 崔炜, 朱发证. 机器人导航的路径规划算法研究综述[J]. 计算机工程与应用, 2023, 59(19): 10-20.
CUI W, ZHU F Z. Review of path planning algorithms for robot navigation[J]. Computer Engineering and Applications, 2023, 59(19): 10-20.

[2] SAADI A A, SOUKANE A, MERAIHI Y, et al. UAV path planning using optimization approaches: A survey[J]. Archives of Computational Methods in Engineering, 2022, 29(6): 4233-4284.

[3] ZHOU C M, HUANG B D, FRÄNTI P. A review of motion planning algorithms for intelligent robots[J]. Journal of Intelligent Manufacturing, 2022, 33(2): 387-424.

[4] YANG Y, LI Z, SHANG Y, et al. Sparse reward for reinforcement learning-based continuous integration testing[J]. Journal of Software: Evolution and Process, 2023, 35(6): e2409.

[5] 杨瑞, 严江鹏, 李秀. 强化学习稀疏奖励算法研究: 理论与实验[J]. 智能系统学报, 2020, 15(5): 888-899.
YANG R, YAN J P, LI X. Survey of sparse reward algorithms in reinforcement learning: Theory and experiment[J]. CAAI Transactions on Intelligent Systems, 2020, 15(5): 888-899.

[6] ROSS S, GORDON G J, BAGNELL J A. A reduction of imitation learning and structured prediction to No-regret online learning[EB/OL]. (2010-11-02)[2024-02-12]. https://arxiv.org/abs/1011.0686.

[7] NAIR A, MCGREW B, ANDRYCHOWICZ M, et al. Overcoming exploration in reinforcement learning with demonstrations[C]//Proceedings of the IEEE International Conference on Robotics and Automation. New York: IEEE, 2018: 6292-6299.

[8] SILVER D, SCHRITTWIESER J, SIMONYAN K, et al. Mastering the game of go without human knowledge[J]. Nature, 2017, 550(7676): 354-359.

[9] 闫皎洁, 张锲石, 胡希平. 基于强化学习的路径规划技术综述[J]. 计算机工程, 2021, 47(10): 16-25.
YAN J J, ZHANG Q S, HU X P. Review of path planning techniques based on reinforcement learning[J]. Computer Engineering, 2021, 47(10): 16-25.

[10] JIANG L, HUANG H Y, DING Z H. Path planning for intelligent robots based on deep Q-learning with experience replay and heuristic knowledge[J]. IEEE/CAA Journal of Automatica Sinica, 2020, 7(4): 1179-1189.

[11] CHEN P Z, PEI J A, LU W Q, et al. A deep reinforcement learning based method for real-time path planning and dynamic obstacle avoidance[J]. Neurocomputing, 2022, 497: 64-75.

[12] MOUSAVI S S, SCHUKAT M, HOWLEY E. Deep reinforcement learning: An overview[C]//Proceedings of SAI Intelligent Systems Conference. Cham: Springer International Publishing, 2017: 426-440.

[13] HESTER T, VECERIK M, PIETQUIN O, et al. Deep Q-learning from demonstrations[EB/OL]. (2017-04-12)[2024-02-12]. https://arxiv.org/abs/1704.03732.

[14] BROWN D S, GOO W, NAGARAJAN P, et al. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations[EB/OL]. [2024-02-12]. https://arxiv.org/abs/1904.06387.

[15] UCHENDU I, XIAO T, LU Y, et al. Jump-start reinforcement learning[EB/OL]. (2022-04-05)[2024-02-12]. https://arxiv.org/abs/2204.02372.

[16] 胡鑫源. 多智能体强化学习中探索策略的研究与实现[D]. 成都: 电子科技大学, 2022.
HU X Y. Research and implementation of exploration strategy in multi-agent reinforcement learning[D]. Chengdu: University of Electronic Science and Technology of China, 2022.

编　辑　张　莉