未知环境下基于深度序列蒙特卡罗树搜索的信源导航方法

段世红[1,2],何 昊[1,2],徐 诚[1,2],殷 楠[1],王 然[1,2]
(1. 北京科技大学计算机与通信工程学院,北京 100083; 2. 北京科技大学顺德研究生院,广东佛山 528399)

摘 要: 信源导航在应急救援、工业巡检及其他危险作业中具有重要应用意义. 在实际应用中,环境的状态信息往往是难以完全观测的,即部分可观测环境. 如何利用观测到的部分环境信息做出实时决策,并基于历史序列信息对系统未来状态进行有效的预测,成为信源导航相关研究所面临的挑战性问题. 本文提出一种基于深度序列蒙特卡洛树搜索(Deep Sequential Monte-Carlo Tree Search,DS-MCTS)的信源导航算法和系统框架,基于序列动作预测(Sequential Action Prediction,SAP)网络为MCTS决策提供先验知识,构建奖励分配预测(Reward Allocation Prediction,RAP)网络提高奖励分配精度,最终实现系统的最优化决策. 仿真实验表明,DS-MCTS方法提供了一种端到端的信源导航解决方案,可以实现智能体动作的有效预测,实现高效、鲁棒的路径规划.

关键词: 信源导航; 蒙特卡洛树搜索; 序贯决策; 路径规划; 深度强化学习
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2022)07-1744-09
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20211252

DS-MCTS: A Deep Sequential Monte-Carlo Tree Search Method for Source Navigation in Unknown Environments

DUAN Shi-hong[1,2], HE Hao[1,2], XU Cheng[1,2], YIN Nan[1], WANG Ran[1,2]
(1. *School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing* 100083, *China*;
2. *Shunde Graduate School, University of Science and Technology Beijing, Foshan, Guangdong* 528399, *China*)

Abstract: Source navigation has important application significance in emergency rescue, industrial patrol, and other dangerous operations. In practical applications, it is often difficult to fully observe the state information of the environment, that is, a partially observable environment. Making real-time decisions using part of the observed environmental information and effectively predicting the system's future state based on the historical sequence information have become a challenge faced by research institutes related to source navigation. This paper proposes a source navigation algorithm and system framework based on deep sequential Monte-Carlo tree search(DS-MCTS). Prior knowledge is provided to MCTS decision-making based on a sequential action prediction(SAP) network. A reward allocation prediction(RAP) network is built to improve the accuracy of reward distribution and finally realize the system's optimal decision-making. The simulation results show that the DS-MCTS method provides an end-to-end source navigation solution, which can effectively predict agents' actions and achieve efficient and robust path planning.

Key words: source navigation; Monte-Carlo tree search; sequential decision-making; path planning; deep reinforcement learning

1 引言

信源导航在应急救援、工业巡检及其他危险作业中具有重要应用意义. 在帮助寻找矿井中的幸存者、在核电站中寻找辐射源、在海洋里寻找石油泄漏源等应

收稿日期:2021-09-13;修回日期:2021-12-28;责任编辑:崔兴华
基金项目:国家自然科学基金(No.62101029);博士后创新人才支持计划(No.BX20190033);广东省基础与应用基础研究基金联合基金(No.2019A1515110325);中国博士后基金面上项目(No.2020M670135);北京科技大学顺德研究生院博士后科研经费(No.2020BH001);中央高校基本科研业务费(No.06500127)

用中,非常需要小型、灵活的机器人在这些复杂环境中实现完全自主的导航与搜索,快速部署智能体.

梯度方法是寻源问题领域最早研究的算法,也是解决寻源问题的有效的方法之一. 梯度方法利用信号场中信号的梯度变化信息引导智能体移动到信号源所在位置. 路永鑫等人[1]提出一种梯度下降法和改进A*算法相结合的应急机器人路径规划方法. 该方法在运动过程中结合梯度下降法进行局部动态路径规划,解决了传感器探测能力局限性和灾情蔓延产生新危险源等情况下的风险规避困难问题. 但是梯度下降法容易陷入局部最优,且梯度计算复杂度较高,效率较低.

由于梯度下降相关算法存在上述问题,近年来解决路径规划问题的方法大多是启发式群智能算法[2~7]. 即通过模拟一些自然现象或生物行为过程来解决路径规划问题,如粒子群优化算法[3]、蚁群算法[4]、遗传算法[5]和克隆选择算法[6]等. 其中,文献[7]提出了一种改进的移动机器人路径规划优化人工蜂群算法,利用贝塞尔曲线描述路径,将路径优化问题转化为生成贝塞尔曲线点的位置优化问题. 这些生物群体式的群启发式算法能较好地避开局部最优值,但都依赖相关参数的设置,极大地影响了算法解决实际问题的能力. 面对动态环境中的路径规划问题,无法预测计划中可能进一步出现的约束和冲突.

动态环境中的路径规划可以表述为一个序列决策问题. 序列是许多信息系统的重要组成部分,在许多应用和系统上起着重要的作用,例如,蜂窝码分多址系统[8]利用扩频序列来区分来自不同用户的信号;脉冲压缩雷达系统[9]利用相位编码序列调制的探测脉冲来实现远距离物体的高分辨率探测. 此外,有许多关于动态环境中路径规划和运动预测的文献并有大量调查[10~12]. 例如,陈劲峰等[13]提出动态环境下基于改进人工势场法的路径规划算法,并表明改进的人工势场法可解决局部最小值和目标不可达问题,且有良好的动态避障能力. 但是有研究发现当周围环境变得越来越复杂时,机器人失去了寻找路径的能力,并选择停止或者不规则行动[14]. 为了克服上述问题,Helbing等[15]提出了一个社会能力模型(Social Force Model,SFM),将智能体之间的协作和交互描述为高斯过程,预测智能体在导航期间的未来运动.

由以上分析可知,梯度方法效率较低且不易推广,生物群体式的启发式算法容易陷入局部最优且难以完全实现自主决策. 强化学习(Reinforcement Learning,RL)可以实现自主学习和决策,是机器学习的一个重要分支[16],其通过不断学习求出马尔可夫决策过程(Markov Decision Process,MDP)[17,18]的解. 强化学习的一个显著特征是“从互动中学习”,智能体通过一系列离散时间步骤与环境进行交互. 在时间t下,智能体观察到环境处于状态S_t,基于对S_t的观察,智能体采取行动a,这导致智能体接收到奖励$R(S_t,a)$,并且环境变成新的状态S_{t+1}. 蒙特卡洛树搜索(Monte-Carlo Tree Search,MCTS)是一种强化学习方法[19],在面临决策问题的多种选择下选出最优的决策结果. 文献[20]提出了一种基于全扩展的MCTS方法,通过减少模拟的步数来加快树的搜索效率. 受此启发,由于MCTS的性能受其有效搜索深度的约束[21],本文希望能够将历史序列决策信息作为MCTS的先验知识,减少MCTS的搜索深度,以促进MCTS根据历史最优决策信息做出最佳决策. 此外,围棋领域最强大的AlphaGo算法[22],让人类领略到深度强化学习的威力,其主要是将深度学习与强化学习融合,对算法进行优化,使之能够在短时间内做出正确的决策. 许多研究亦可证明,深度强化学习在路径规划领域对提高智能体的导航能力是有效的[23,24].

出于上述考虑,本文充分利用深度强化学习的强大优势,面向部分可观测环境下的信源导航问题提出一个健壮且有效的算法,即基于深度序列蒙特卡洛树搜索的信源导航(Deep Sequential Monte-Carlo Tree Search,DS-MCTS)方法. 进一步根据该方法提出一个结合长短期记忆网络(Long Short-Term Memory,LSTM)和蒙特卡洛树搜索的集成信源导航框架. 对智能体在信源导航过程中的序列轨迹信息和决策信息采样保存,序列动作预测(Sequential Action Prediction,SAP)网络利用历史序列信息给MCTS方法提供先验知识,奖励分配预测(Reward Allocation Prediction,RAP)网络在训练中提高奖励分配精度,促进MCTS方法最优化决策. 本文还将提出的DS-MCTS方法在模拟信号场中进行了相关实验,实验结果表明,该方法能够在部分可观测环境下有效的进行路径规划,并且具有非常稳定的性能. 同时,也能证明深度学习与强化学习融合是机器人应用中一组有前途的算法,快速发展的深度强化学习领域使得应用更加健壮和准确.

本文主要贡献包括以下几个方面.

(1) 提出基于DS-MCTS方法和框架,将该方法和框架应用于智能体信源导航过程中. 研究表明,本文提出的方法和框架能较大程度地利用序列数据优化信源导航过程中的决策以及提高智能体的信源导航成功率,解决了传统MCTS过程缺乏对历史信息的提取利用问题,避免复杂环境下智能体陷入局部最优.

(2) 提出了序列动作预测(SAP)网络,利用LSTM和全连接层的特性,根据智能体的历史轨迹数据信息预测当前时刻下智能体的动作可能性,为蒙特卡洛树搜索决策提供先验知识,促进最优化决策.

（3）提出了端到端的奖励分配预测（RAP）网络，解决之前模拟阶段过于复杂的仿真计算问题，提高MCTS方法中的奖励分配精度以及搜索效率.

2 系统建模

2.1 问题定义

信源导航问题主要是指智能体在部分可观测的信号场中寻找信号源的问题. 未知环境下的信源搜索可以定义为一个部分可观测马尔科夫决策过程（Partially Observable Markov Decision Process，POMDP）. POMDP模型是马尔科夫决策模型的扩展，在强化学习中，MDP是对完全可观测的环境进行描述的，也就是观测到的状态内容完整地决定了决策需要的特征[25]. 但是很多情况下，系统的完整的状态信息难以获取，特别是测量环境信息的传感器信号容易受到噪声的影响. 同时，POMDP假设系统的状态信息不能直接观测得到，是部分可知的，即系统状态仅部分可见情况下的马尔可夫决策过程，这符合本文信源导航问题的实际情况. 所以在本文所提出的方法中，智能体根据自身传感器获得的部分环境信息经由蒙特卡洛树搜索输出角度移动到下一步的目标位置，直至找到信号源，整个过程可以建模为一个POMDP模型，其由八元组$\left(S,A,Z,T,O,R,\gamma,b_0\right)$组成.

S：智能体的连续状态空间，其中状态由位置表示. S_t是智能体在t时刻的位置，$S_t=(x_t,y_t)$，$S=\left\{S_i,S_{i+1},\cdots,S_t\right\}$可以理解成智能体的轨迹信息.

A：动作的离散集合. A_t是智能体在t时刻的运动方向，$A=\left\{A_i,A_{i+1},\cdots,A_t\right\}$代表智能体的历史运动方向信息，其中，$A_t\in\left\{0,\pi/4,\pi/2,3\pi/4,\pi,5\pi/4,3\pi/2,7\pi/4\right\}$.

Z：观测到的环境信息. t时刻下的观测信息$Z_t=I_{S_t}+\omega$，I_{S_t}表示S_t位置下信号强度，ω是观测噪声.

T：$S\times A\to S'$，状态转变函数，可以理解成智能体的运动方程，表示为

$$\begin{cases}x_t=v\times\cos A_t+x_{t-1}\\ y_t=v\times\sin A_t+y_{t-1}\end{cases}\tag{1}$$

其中，v是智能体的前进速度，A_t是动作方向，x_{t-1}和y_{t-1}是智能体上一时刻于信号场的横坐标和纵坐标.

O：$S\times A\to O(Z)$，观测模型，例如$O(Z_{t+1}=Z|S_{t+1}=s,A_t=a)$.

R：$S\times A\to R$，智能体在状态s下采取动作a获得的奖励$R(s,a)$.

γ：折扣因子，$0\leqslant\gamma\leqslant1$.

b_0：智能体初始信念状态.

首先在状态、动作空间上训练SAP网络，再将训练好的网络用于智能体在下一时间步的蒙特卡洛树搜索决策上，并在后续路径规划上递归应用. 图1为集成信源导航框架图，概述了如何使用SAP网络根据t的前m时刻的轨迹信息和历史运动方向信息，预测下一时刻的动作方向概率p_t. 在创建搜索树的过程中，对每个扩展节点进行模拟，参考p_t后通过RAP网络给出预测奖励值，同时不断训练RAP网络.

2.2 系统框架

动态环境中的路径规划可以表述为一个顺序决策问题. 在信源导航过程中倘若把整个过程分为若干个连续的阶段，各个阶段的决策结果前后衔接，这样可通过历史决策序列为下一时刻做出最佳决策提供有效信息. 此外，由于信号源环境部分可观测，且在探索过程中奖励函数是稀疏的，稀疏的奖励计划往往需要长期的信息收集，希望能够通过神经网络来近似真实奖励值. 为此，本文提出集成信源导航框架来解决这一问题，体现DS-MCTS方法的有效性.

框架主要分为3个部分. 一是SAP网络. 核心是长短期记忆神经网络，能够使神经元在管道中保持前后序列记忆. 滑动窗口单步向前移动更新历史数据信息，解决了梯度消失问题，通过对智能体轨迹和历史动作选择输出先验动作概率知识，促进MCTS算法做出最佳策略. 二是MCTS算法. 在先验动作概率信息下，经过树搜索给出唯一最佳动作方向决策. 三是RAP网络. 端到端输出预测奖励值，在树搜索模拟阶段通过不断训练，使得模拟奖励逼近真实奖励，提高奖励分配的精度，提升MCTS算法的决策效率，降低树搜索模拟阶段的复杂度.

3 深度序列蒙特卡洛树搜索方法

3.1 蒙特卡洛树搜索

MCTS方法是一种用于决策问题的启发式搜索算法[26~28]，最著名的是在博弈游戏中使用，如AlphaGo[22]. MCTS方法的核心思想是通过迭代地对动作空间进行随机采样并根据采样结果构建搜索树来找到最优决策. 在搜索树中，每个节点表示决策域的一个状态，指向其子节点的链接表示导致后续状态的动作. 如图2所示，在每次迭代中，MCTS方法执行4个步骤即选择、扩展、模拟和反向传播. 蒙特卡洛树搜索过程根据智能体的轨迹信息，迭代搜索过程以得到动作策略π，蒙特卡洛树搜索从根节点开始，树中每扩展节点都会包含信息$\left\{I(n),s(n),a(n),p(n),R(n),N(n)\right\}$. 其中，$I(n)$表示节点$n$所处位置$s(n)$的信号值，$a(n)$表示节点$n$的父节点到节点$n$的动作方向，$p(n)$表示节点$n$的先验动作选择概率，$R(n)$表示节点$n$的累积奖励，$N(n)$表示节点$n$的被访问次数. MCTS方法主要分为以下几个步骤：

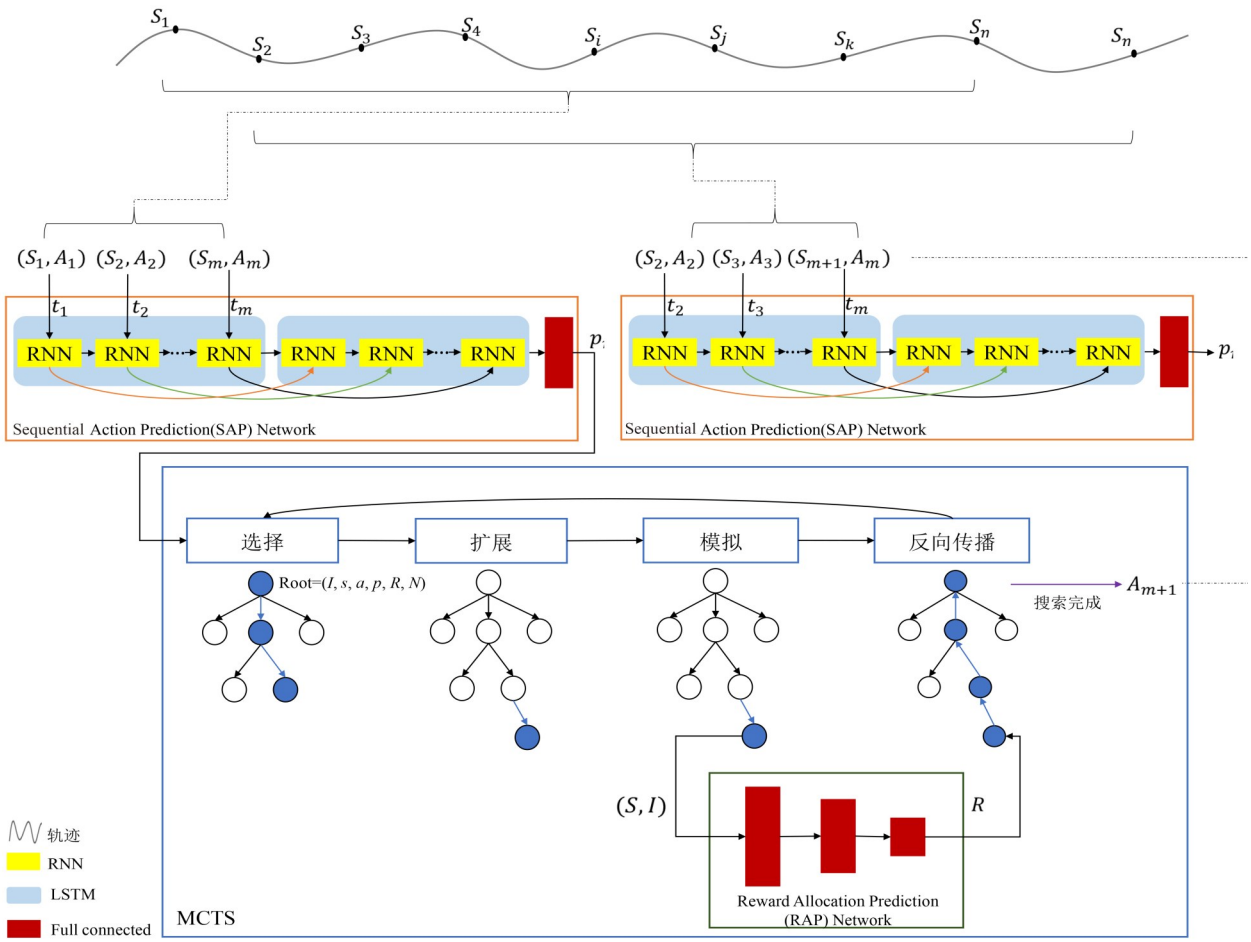


图1　集成信源导航框架图

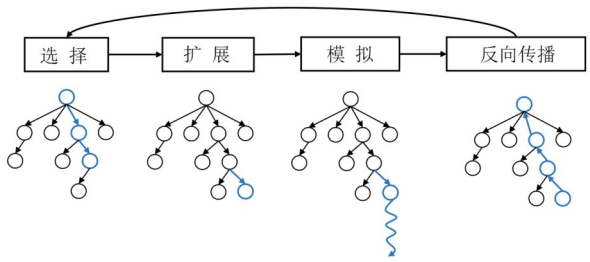


图2　蒙特卡洛树搜索示意图

（1）选择:从根节点开始,应用树的上限置信度公式(Upper Confidence bound apply to Trees,UCT)[29]来选择子节点,UCT平衡了节点的探索和利用.UCT公式为

$$U=\frac{R(n)}{N(n)}+C\sqrt{\frac{\ln N(n_h)}{N(n)}} \tag{2}$$

其中,$N(n_h)$表示n节点的父节点被遍历的次数.

（2）扩展:如果节点n不是终止节点,有节点n未扩展过的可选动作集合,随机选择集合中的动作,根据式(1)生成子节点的位置信息,以此扩展搜索树,并将子节点相关信息初始化.

（3）模拟:扩展子节点后,RAP网络预测更新该节点的奖励值.

（4）反向传播:该步骤中,奖励值和访问次数被传播回根节点,更新每个节点的统计信息:

步骤(1)~(4)反复执行,直到达到最大迭代次数,在根节点下根据式(2)选择其最佳子节点以及对应的动作a,作为该时间步MCTS方法输出的策略.

3.2　SAP网络

本文提出的SAP网络如图3所示,其结构主要由LSTM和全连接网络组成.LSTM是一种特殊的RNN,克服RNN的“梯度消失”问题.在智能体信源导航过程中,有一个长度为m的变长滑动窗口.每个时刻滑动窗口往前移动一步,与此同时将当前时刻前m时间步的轨迹信息和对应的动作方向作为输入,输出智能体下一时刻针对动作选择的概率信息,作为MCTS方法的先验知识.因为包括智能体轨迹在内的历史信息能够反映智能体在这一小段时间的运动趋势,通过训练SAP网络学习智能体的动作选择概率来预测运动趋势,这样能大大提高智能体的信源导航效率并避免局部最优.此外,使用已知的历史信息作为输入已经被证明可以学习智能体的动作和下一个时刻位置之间的关

系[30]. 该网络是从现实世界的相互作用中学习得到的，然后用来模拟连续动作的转换. 网络包括接受输入的非线性嵌入层和LSTM层，非线性嵌入层使用校正线性单元激活，输出通过线性层传递，映射到智能体在下一时刻各个动作方向上的概率，表示为

$$p_t=f_v(S_t,A_t|W_{\mathrm{SAP}}) \tag{3}$$

其中，W_{SAP}表示要配置的SAP网络的参数.

本文使用m个时间变长的编码序列，训练过程是将式(4)的损失降至最低，即

$$\mathrm{Loss}_{\mathrm{SAP}}=\sqrt{\frac{1}{m}\sum_{a\in A}^{m}(p'_a-p_a)^2} \tag{4}$$

其中，A表示运动方向集合，p'_a表示a方向的预测概率，p_a表示a方向的真实概率.

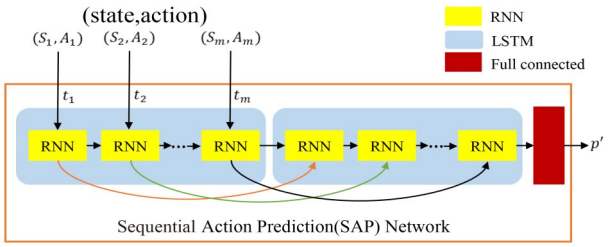


图3 SAP网络结构图

3.3 RAP网络

本文提出的RAP网络如图4所示，使用智能体的轨迹信息和预测的动作方向概率作为输入，预测当前位置应该被分配的奖励值. 如图1所示，RAP嵌入MCTS方法中，应用于MCTS方法的模拟阶段. 传统MCTS的模拟阶段需要一直模拟到达终止状态，加入RAP网络可以并行单步模拟估计节点的奖励值，所有节点在模拟阶段直接通过RAP网络就可直接获得奖励值[31]. 这样能将模拟阶段的步骤简化并大大提高搜索的性能. 蒙特卡洛树搜索过程中的行走轨迹也将会及时更新到智能体的轨迹信息中，蒙特卡洛树当前节点的预测奖励值通过式(5)得出：

$$R_t=f_v(S_t,I_{S\_t}|W_{\mathrm{RAP}}) \tag{5}$$

其中，I_{S_t}表示t时刻智能体在位置S_t的信号强度，W_{RAP}表示要配置的RAP网络的参数.

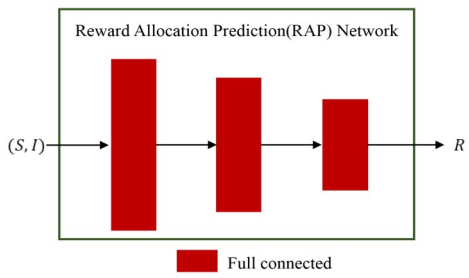


图4 RAP网络结构图

RAP网络是用来预测当前位置的奖励值，通过不断训练RAP网络提高预测精度，训练的损失函数为

$$\mathrm{Loss}_{\mathrm{RAP}}=\left\|R_t-R'_t\right\| \tag{6}$$

$$R=\alpha\left(I_{\mathrm{c\_pos}}-I_{\mathrm{l\_pos}}\right)+(1-\alpha)\left(D_{\mathrm{c\_pos}}-D_{\mathrm{l\_pos}}\right) \tag{7}$$

其中，R'_t为当前位置的真实奖励值，α是权重系数，$I_{\mathrm{c_pos}}$和$I_{\mathrm{l_pos}}$分别为当前位置和上一时刻位置的信号强度，$D_{\mathrm{c_pos}}$和$D_{\mathrm{l_pos}}$分别是当前时刻和上一时刻距离信号源的曼哈顿距离，D的计算方式为

$$D=\left\|P-P_s\right\| \tag{8}$$

其中，P_s表示信号源所处的位置.

4 实验结果和分析

本节对本文提出的DS-MCTS方法进行实验验证，以评估所提出的方法的性能表现，验证其有效性.

4.1 实验设置

实验模拟一个信号源在大小20 m×20 m的信号场中，本文将信源强度建模为离信源距离的函数，智能体在信号场各位置观测到的信号强度由式(9)[32]给出：

$$I_P=100-20\log_{10}2400-20\log_{10}\left\|P-P_s\right\|-R \tag{9}$$

其中，P代表信号点的位置，P_s代表信号源位置. R为模拟智能体对于真实信号的实际接收情况而加入的噪声，$R=\sqrt{X^2+Y^2}$，X和Y分别服从正态分布，即

$$X\sim N\left(v\cos\theta,\sigma^2\right),Y\sim N(v\sin\theta,\sigma^2).$$

本文模拟这样一个信号场去训练智能体寻找信号源，将信号场区域划分成方形网格，简化搜索区域. 这一方法把信号场区域简化为一个二维数组，数组的每一个元素是信号场的一个方块. 初始化智能体在信号场中的起始位置，信号源位置固定，智能体采样信源导航过程中的轨迹信息和信号强度信息，通过训练SAP和RAP网络，能得到一个策略，可以使智能体决策朝着离信号源梯度上升最快方向行动，并且该策略适用于不同的信源环境.

对于SAP网络，网络结构如表1所示，采用两层LSTM叠加，再通过具有8个神经元的全连接层预测动作概率信息. 间隔100个采样样本训练一次，batch size设置为10，学习速率设置为0.000 1. 输入为智能体当前时刻的前m时间步的位置和动作信息，每一时间间隔类似于单步滑动窗口，输出为每个动作方向的概率，为MCTS方法提供先验知识.

表1 SAP网络结构

网络层数	神经元个数	激活函数
LSTM	72	—
LSTM	100	—
Full connected	8	SoftMax

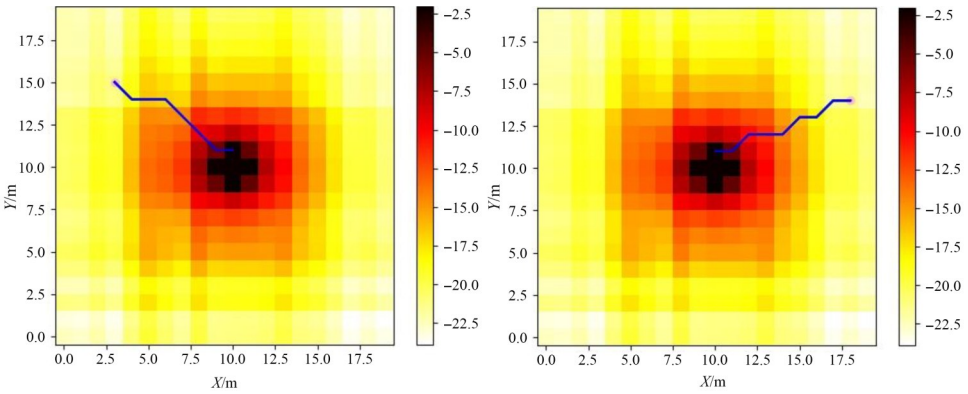
对于RAP网络,网络结构如表2所示,采用三层全连接神经网络叠加. 间隔2 000个采样样本训练一次,batch size设置为50,学习速率设置为0.000 1,训练中MCTS方法迭代次数设置为200,深度设置为4. 输入为智能体当前节点的前m时间步的位置和信号强度信息以及先验动作概率信息,输出为当前节点模拟的预测奖励值,通过训练不断提高奖励分配精度,提高MCTS方法的决策效率.

表2　RAP网络结构

网络层数	神经元个数	激活函数
Full connected	18	tanh
Full connected	10	tanh
Full connected	1	—

4.2　性能分析

本文进行了如图5所示的仿真实验,可选动作方向角度设置为{0, π/4, π/2, 3π/4, π, 5π/4, 3π/2, 7π/4}. 本实验智能体获取到的信号源分布噪声R的均值设置为1、方差设置为2. 图5是智能体在同一信号源场景下将SAP和RAP网络训练后的两次迭代信源导航,信号源位置不变并设置为[10,10],每一次智能体的初始位置不同,分别是[3,15],[18,14]. 本文提出的方法降低了信号场中噪声对梯度计算的干扰,从图中可看出,智能体不论以哪个方向作为起点,均能与信号场中梯度上升最快的方向吻合. 这说明了DS-MCTS方法具有较好的稳定性和抗噪声干扰的特性,进一步说明了SAP网络预测的动作概率先验知识是比较准确的. 此外RAP网络在训练后能够给予智能体在MCTS方法模拟阶段精准的奖励,从而促进智能体选择一条更优的信源导航路径,提高信源导航效率.



(a)　初始位置[3,15]　　(b)　初始位置[18,14]

图5　仿真实验信源导航路径图

图6和图7分别表示RAP网络和SAP网络的训练损失曲线图. RAP网络训练损失收敛于大概300个epoch时,SAP训练损失收敛于100个epoch左右,且之后都保持在一个较低的损失. 特别注意的是,虽然SAP损失在图中表示的波动较大,实际上收敛于(0.5,0.7)区间,明显较之前损失低且稳定,同时也说明两个神经网络的网络结构和参数设计合理,能够快速的训练网络,促进DS-MCTS方法对信源导航的高效决策.

图8表示智能体在训练期间每次迭代信源导航的步数图线,该实验信号源位置保持不变设置为[10,10],智能体的位置每次迭代均随机生成(距离信号源不低于10),信号场分布均值为1,方差为1.7,智能体可选8个动作方向. 由图8可知,智能体在开始25次迭代次数中,步数显著较高,说明智能体并未找到一条最优信源导航路径或者未能在迭代终止条件前成功寻找到信号源. 同时,通过信源导航期间中不断的训练SAP和RAP网络,在迭代60次之后,步数稳定在20步左右,说明智能体能以较快速度收敛到信号源,以及

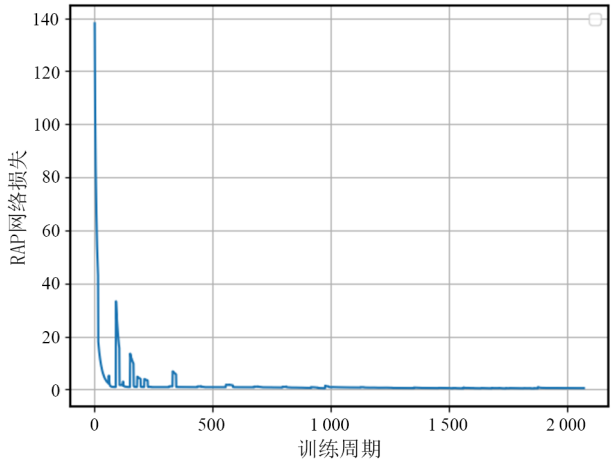


图6　RAP网络训练损失图

本文提出的DS-MCTS方法能够在学习中不断优化. 实验中智能体通过多次迭代学习到一个对于当前信源环境的最佳路径规划策略,并在之后的迭代步数中应用该策略寻找信号源,说明本方法具有非常稳定的性

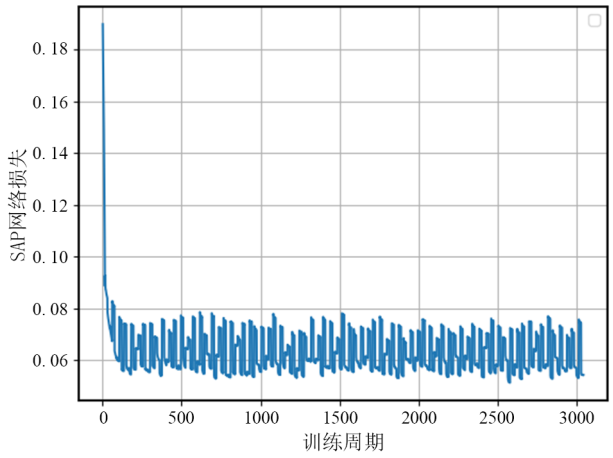


图7 SAP网络训练损失图

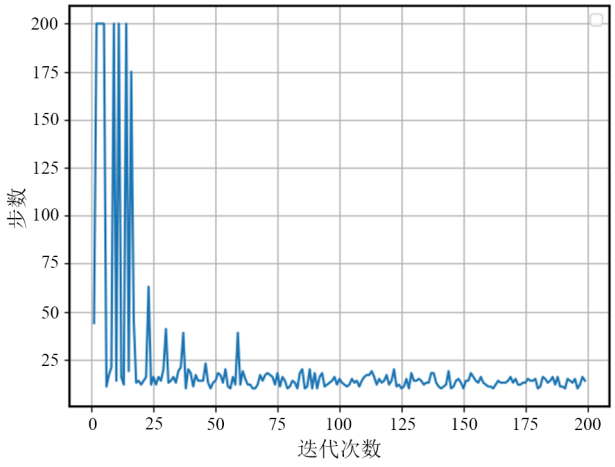


图8 迭代步数图

能表现.

本文在对所提出的DS-MCTS方法进行验证的同时,还使用梯度下降法(Gradient Descent,GD)[33]、蒙特卡洛树与高斯过程(Monte Carlo Tree Search with Gaussian Process,MCTS-GP)结合方法[34]作对比. 在相同的信源环境下,随机进行100次仿真实验,每次实验的初始位置随机设定. 分别对3种方法进行信源导航实验,并将每次实验迭代的步数从小到大排序,结果如图9和表3所示.

可以看出,单纯梯度下降法虽然决策效率较高,但是相对于另外两种方法,成功率明显降低,难以满足应用需求;DS-MCTS和MCTS-GP的寻源成功率相近,但是本文提出DS-MCTS方法平均步数更低,执行时间更少,效率更高;MCTS-GP方法由于引入高斯过程预测模拟阶段的奖励,从而导致计算开销大,决策时间显著增加,在实际场景中难以满足应用实时性的需求,不利于推广. 由此可见,本文所提出的DS-MCTS方法在点源环境下寻源具有良好的鲁棒性和较高的效率.

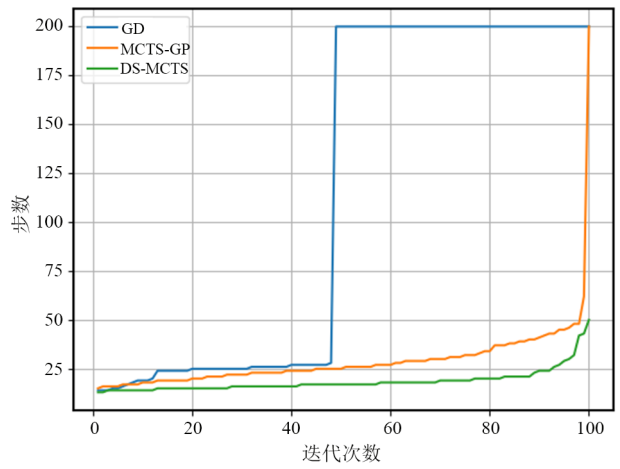


图9 不同方法对比实验步数

表3 不同方法信源导航结果比较

方法	寻源成功率/%	平均步数	时间/s
GD[32]	48	115	1.2
MCTS-GP[33]	99	29	9.5
DS-MCTS	100	18	0.9

5 总结

本文探讨了寻源的研究前景和研究意义,提出了DS-MCTS方法和框架,并通过实验验证本文提出方法框架能大大提高智能体的信源导航成功率并降低信源搜索过程的导航时间. 此外,本文提出的网络还展示智能体在信源导航过程中能够利用历史数据信息准确预测智能体的动作趋势,为MCTS算法决策提供先验知识,提升决策效果. 同时,在MCTS模拟阶段中加入端到端的RAP网络,提高了搜索效率以及奖励分配精度,促进蒙特卡洛树最优化决策. 在后续研究中,如何在保持信源搜索效率的同时提高定位精度,将会是一个主要工作方向.

参考文献

[1] 路永鑫, 魏云冰, 赵启承, 等. 基于层次分析法和改进A*算法的电力应急机器人路径规划[J]. 电力系统保护与控制, 2021, 49(9): 82-89.
LU Y X, WEI Y B, ZHAO Q C, et al. Path planning of a power emergency robot based on an analytic hierarchy process and improved A* algorithm[J]. Power System Protection and Control, 2021, 49(9): 82-89. (in Chinese)

[2] MANDAL S, SAHA D, MAHANTI A. A heuristic search for generalized cellular network planning[C]//2002 IEEE International Conference on Personal Wireless Communications. New Delhi, India: IEEE, 2002: 105-109.

[3] LIN Q Z, LIU S B, ZHU Q L, et al. Particle swarm optimization with a balanceable fitness estimation for many-ob-

jective optimization problems[J]. IEEE Transactions on Evolutionary Computation, 2018, 22(1): 32-46.

[4] ROJAS SANTIAGO M, MUTHUSWAMY S, HULETT M. An ACO algorithm for scheduling a flow shop with set-up times[J]. International Journal of Industrial and Systems Engineering, 2020, 36(1): 98-109.

[5] LI J, SHU Z. Research on path planning based on improved D* lite genetic algorithm[J]. Machine Tool & Hydraulics, 2019, 47(11): 39-42.

[6] YAN X S, LI P P, TANG K, et al. Clonal selection based intelligent parameter inversion algorithm for prestack seismic data[J]. Information Sciences, 2020, 517: 86-99.

[7] WANG H Q, HU Y Y, LIAO W D, et al. Path planning algorithm based on improved artificial bee colony algorithm [J]. Control Engineering, 2016, 23(95): 1407-1411.

[8] VITERBI A J. CDMA: Principles of Spread Spectrum Communication[M]. Wokingham: Addison-Wesley, 1995.

[9] SKOLNIK M I. Radar Handbook[M]. 3rd ed. New York: McGraw-Hill, 2008.

[10] XIE G, SHANGGUAN A Q, FEI R, et al. Motion trajectory prediction based on a CNN-LSTM sequential model[J]. Science China Information Sciences, 2020, 63(11): 1-21.

[11] 许凯波, 鲁海燕, 黄洋, 等. 基于双层蚁群算法和动态环境的机器人路径规划方法[J]. 电子学报, 2019, 47(10): 2166-2176.
XU K B, LU H Y, HUANG Y, et al. Robot path planning based on double-layer ant colony optimization algorithm and dynamic environment[J]. Acta Electronica Sinica, 2019, 47(10): 2166-2176. (in Chinese)

[12] WEN T, YANG D C, LIU W F, et al. A novel integrated path planning algorithm for warehouse AGVs[J]. Chinese Journal of Electronics, 2021, 30(2): 331-338.

[13] 陈劲峰, 黄卫华, 章政, 等. 动态环境下基于改进人工势场法的路径规划算法[J]. 组合机床与自动化加工技术, 2020(12): 6-9, 14.
CHEN J F, HUANG W H, ZHANG Z, et al. Path planning algorithm based on improved artificial potential field method in dynamic environment[J]. Modular Machine Tool & Automatic Manufacturing Technique, 2020(12): 6-9, 14. (in Chinese)

[14] TRAUTMAN P, KRAUSE A. Unfreezing the robot: Navigation in dense, interacting crowds[C]//2010 IEEE/RSJ International Conference on Intelligent Robots and Systems. Taipei, China: IEEE, 2010: 797-803.

[15] HELBING D, MOLNÁR P. Social force model for pedestrian dynamics[J]. Physical Review. E, Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics, 1995, 51(5): 4282-4286.

[16] SUTTON R S, Barto A G. Reinforcement learning: An introduction[M]. MIT press, 2018.

[17] PUTERMAN M L. Markov Decision Processes: Discrete Stochastic Dynamic Programming[M]. Hoboken: John Wiley & Sons, 1994.

[18] PENG P, ZHU F, LIU Q, et al. Achieving safe deep reinforcement learning via environment comprehension mechanism[J]. Chinese Journal of Electronics, 2021, 30(6): 1049-1058.

[19] VODOPIVEC T, SAMOTHRAKIS S, STER B. On Monte Carlo tree search and reinforcement learning[J]. Journal of Artificial Intelligence Research, 2017, 60: 881-936.

[20] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[21] KEARNS M, MANSOUR Y, NG A Y. A sparse sampling algorithm for near-optimal planning in large Markov decision processes[J].Machine Learning, 2002, 49(2/3): 193-208.

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

[23] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[J]. arXiv preprint, arXiv:1509.02971, 2015.

[24] 黄志清, 曲志伟, 张吉, 等. 基于深度强化学习的端到端无人驾驶决策[J]. 电子学报, 2020, 48(9): 1711-1719.
HUANG Z Q, QU Z W, ZHANG J, et al. End-to-end autonomous driving decision based on deep reinforcement learning[J]. Acta Electronica Sinica, 2020, 48(9): 1711-1719. (in Chinese)

[25] CHEN X, LI Z, WANG K, et al. MDP-based network selection with reward optimization in HetNets[J]. Chinese Journal of Electronics, 2018, 27(1): 183-190.

[26] COULOM R. Efficient selectivity and backup operators in Monte-Carlo tree search[C]//International Conference on Computers and Games. Turin, Italy: Springer, 2007: 72-83.

[27] CHASLOT G, BAKKES S, SZITA I, et al. Monte-Carlo tree search: A new framework for game AI[C]//Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment. Palo Alto, California, USA: The AAAI Press. 2008, 4(1): 216-217.

[28] BROWNE C B, POWLEY E, WHITEHOUSE D, et al. A

survey of monte carlo tree search methods[J]. IEEE Transactions on Computational Intelligence and AI in games, 2012, 4(1): 1-43.

[29] KRONBERGER G, BRAUNE R. Bandit-based Monte-Carlo planning for the single-machine total weighted tardiness scheduling problem[C]//International Conference on Computer Aided Systems Theory. Las Palmas de Gran Canaria, Spain: Springer, 2007: 837-844.

[30] ZHANG J J, LIU H Y, CHANG Q, et al. Recurrent neural network for motion trajectory prediction in human-robot collaborative assembly[J]. CIRP Annals, 2020, 69(1): 9-12.

[31] 徐诚, 何昊, 段世红, 等. 一种基于深度蒙特卡洛树搜索的信源导航方法及装置: 202110316103.9[P].2021.

[32] DENKER A, İŞERI M C. Design and implementation of a semi-autonomous mobile search and rescue robot: SALVOR[C]//2017 International Artificial Intelligence and Data Processing Symposium (IDAP). Malatya, Turkey: IEEE, 2017: 1-6.

[33] 方朋朋, 杨家富, 施杨洋, 等. 基于梯度下降法和改进人工势场法的无人车避障方法[J]. 制造业自动化, 2018, 40(11):81-84.
FANG P P, YANG J F, SHI Y Y, et al. Gradient descent method and improved artificial potential field method for obstacle avoidance of unmanned vehicle[J]. Manufacturing Automation, 2018, 40(11): 81-84. (in Chinese)

[34] VISERAS A, SHUTIN D, MERINO L. Robotic active information gathering for spatial field reconstruction with rapidly-exploring random trees and online learning of Gaussian processes[J]. Sensors(Basel, Switzerland), 2019, 19(5): 1016.

作者简介

段世红 女,1973年生,山西太原人. 北京科技大学副教授. 研究方向为多智能体系统、嵌入式计算与无线定位、数值优化与分布式计算.

何 昊 男,1997年生,湖南永州人. 研究方向为多智能体系统、无线定位.

徐 诚(通讯作者) 男,1988年生,辽宁开原人. 北京科技大学副教授. 研究方向为群体智能与协同计算、多智能体系统与分布式安全.
E-mail: xucheng@ustb.edu.cn

殷 楠 女,1996年生,山西忻州人. 研究方向为多智能体系统、无线定位.

王 然 女,1991年生,北京人. 主要研究方向为群体智能与协同计算、多智能体系统与分布式安全.