(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113139644 B
(45)授权公告日 2024.02.09

(21)申请号 202110316103.9

(22)申请日 2021.03.24

(65)同一申请的已公布的文献号
申请公布号 CN 113139644 A

(43)申请公布日 2021.07.20

(73)专利权人 北京科技大学顺德研究生院
地址 528300 广东省佛山市顺德区大良致慧路2号

(72)发明人 徐诚 何昊 段世红 殷楠

(74)专利代理机构 北京市广友专利事务所有限责任公司 11237
专利代理师 张仲波 付忠林

(51)Int.Cl.
G01C 21/20(2006.01)
G06N 3/0442(2023.01)
G06N 3/045(2023.01)
G06N 3/006(2023.01)
G06N 3/092(2023.01)
G06N 3/084(2023.01)
G06N 3/0464(2023.01)
G06N 5/01(2023.01)

(56)对比文件
CN 109063823 A,2018.12.21
CN 109682392 A,2019.04.26
CN 110427261 A,2019.11.08
CN 110989352 A,2020.04.10
CN 111242246 A,2020.06.05
CN 111506980 A,2020.08.07
CN 111780777 A,2020.10.16
CN 111832501 A,2020.10.27
KR 19990038244 A,1999.06.05
US 2020127457 A1,2020.04.23
仉新;张禹;苏晓明.移动机器人自主定位和导航系统设计与实现.机床与液压.2020,(第10期),全文.
张健;潘耀宗;杨海涛;孙舒;赵洪利.基于蒙特卡洛Q值函数的多智能体决策方法.控制与决策.2020,(第03期),全文.
黄东晋;蒋晨凤;韩凯丽.基于深度强化学习的三维路径规划算法.计算机工程与应用.2020,(第15期),全文.

审查员 吴圆霞

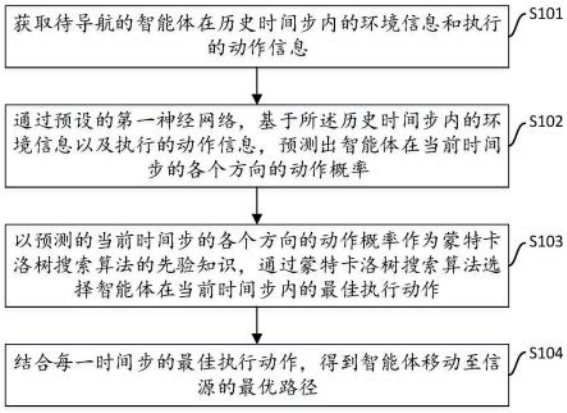
权利要求书1页 说明书7页 附图2页

(54)发明名称

一种基于深度蒙特卡洛树搜索的信源导航方法及装置

(57)摘要

本发明公开了一种基于深度蒙特卡洛树搜索的信源导航方法及装置,该方法包括:获取待导航智能体在历史时间步内的环境信息和执行的动作信息;通过预设的第一神经网络,基于历史时间步内的环境信息和动作信息,预测出智能体在当前时间步的各个方向的动作概率;以预测的动作概率作为蒙特卡洛树搜索算法的先验知识,选择智能体在当前时间步内的最佳执行动作;结合每一时间步的最佳执行动作,得到智能体移动至信源的最优路径。本发明提出在蒙特卡洛树中使用循环神经网络的集成规划路径框架,帮助提高导航控制的稳定性和性能,通过对时间动作序列数据的处理,解决连续空间中的路径规划问题。



CN 113139644 B

1.一种基于深度蒙特卡洛树搜索的信源导航方法,其特征在于,包括:

获取待导航的智能体在历史时间步内的环境信息和执行的动作信息;

通过预设的第一神经网络,基于所述历史时间步内的环境信息以及执行的动作信息,预测出智能体在当前时间步的各个方向的动作概率;

以预测的当前时间步的各个方向的动作概率作为蒙特卡洛树搜索算法的先验知识,通过蒙特卡洛树搜索算法选择智能体在当前时间步内的最佳执行动作;

结合每一时间步的最佳执行动作,得到智能体移动至信源的最优路径;

所述预设的第一神经网络为长短期记忆人工记忆神经网络;

在通过蒙特卡洛树搜索算法选择智能体在当前时间步内的最佳执行动作时,在蒙特卡洛树搜索算法的模拟阶段,所述方法还包括:

将预测的动作概率和当前节点的状态信息输入预设的第二神经网络,通过预设的第二神经网络为当前节点分配奖励值,再将奖励值反向传播至根节点;

所述预设的第二神经网络为卷积神经网络;

在通过预设的第二神经网络为当前节点分配奖励值,再将奖励值反向传播至根节点之后,所述方法还包括:

用获取的奖励值继续训练所述预设的第二神经网络,以提高预测能力。

2.一种基于深度蒙特卡洛树搜索的信源导航装置,其特征在于,包括:

历史环境信息及动作信息获取模块,用于获取待导航的智能体在历史时间步内的环境信息和执行的动作信息;

动作概率预测模块,用于通过预设的第一神经网络,基于所述历史环境信息及动作信息获取模块所获取的历史时间步内的环境信息以及执行的动作信息,预测出智能体在当前时间步的各个方向的动作概率;

最佳执行动作决策模块,用于以所述动作概率预测模块所预测的当前时间步的各个方向的动作概率作为蒙特卡洛树搜索算法的先验知识,通过蒙特卡洛树搜索算法选择智能体在当前时间步内的最佳执行动作;

最优路径获取模块,用于结合所述最佳执行动作决策模块输出的每一时间步的最佳执行动作,得到智能体移动至信源的最优路径;

所述预设的第一神经网络为长短期记忆人工记忆神经网络;

在通过蒙特卡洛树搜索算法选择智能体在当前时间步内的最佳执行动作时,在蒙特卡洛树搜索算法的模拟阶段,所述最佳执行动作决策模块还用于:

将预测的动作概率和当前节点的状态信息输入预设的第二神经网络,通过预设的第二神经网络为当前节点分配奖励值,再将奖励值反向传播至根节点;

所述预设的第二神经网络为卷积神经网络;

在通过预设的第二神经网络为当前节点分配奖励值,再将奖励值反向传播至根节点之后,所述最佳执行动作决策模块还用于:

用获取的奖励值继续训练所述预设的第二神经网络,以提高预测能力。

一种基于深度蒙特卡洛树搜索的信源导航方法及装置

技术领域

[0001] 本发明涉及计算机科学技术领域,特别涉及一种基于深度蒙特卡洛树搜索的信源导航方法及装置。

背景技术

[0002] 针对信息残缺的环境的决策和求解过程,假设系统的状态信息不能直接观测得到,是部分可知的,因而对只有不完全状态信息的系统建模,依据当前的不完全状态信息做出决策。例如,在许多环境和地球科学应用中,专家希望收集具有最大科学价值的样本(例如溢油源),通过收集的样本做下一步决策,但现象的分布最初是未知的。典型地,样本是由技术人员或移动平台按照预定的覆盖轨迹在预定位置收集的。这些非自适应策略最大程度地导致了样本稀疏性,并且当环境的几何结构未知(例如,巨石场)或变化(例如,潮汐带)时可能不可行,最大程度地增加有价值的样本数量需要自适应的定位和导航。

[0003] 部分可观测环境下进行定位导航,由于状态的不可观测性,不能通过状态直接做出决策,对于决策过程,蒙特卡洛树搜索是一种启发式的最优搜索算法,自提出以来,对许多游戏来说都是一个重大突破。因为它可以平衡探索和开发的同时进行搜索。在搜索过程中,首先需要对状态进行预测,如果是离散状态空间,有研究发现可通过高斯过程回归预测状态。但面对连续状态空间,如何有效准确的预测状态是亟需解决的一个问题。

[0004] 此外,在复杂环境情况下智能体进行定位导航只能依靠视觉信息来探索控制,开发基于视觉的定位和导航是在模仿人类在感知环境的思维过程,在探索过程中,奖励函数往往是稀疏的,稀疏的奖励计划需要长期的信息收集,这是现在智能体技术中一个挑战性问题。此外,随着计算机视觉的发展,传统导航的方法对图像状态的预测有一些缺陷,循环学习有待能成为基于视觉的导航设计的一种有效的解决方法。

[0005] 综上,针对部分可观测环境下的智能体路径规划问题,现有技术的稳定性和性能还不够理想,因此,亟需研发一种新的信源导航方法,实现在给定含信号源的信号场后,能够高效、稳定地寻求智能体的最优路径规划。

发明内容

[0006] 本发明提供了一种基于深度蒙特卡洛树搜索的信源导航方法及装置,以解决针对部分可观测环境下的智能体路径规划问题,现有技术的稳定性和性能还不够理想的技术问题。

[0007] 为解决上述技术问题,本发明提供了如下技术方案:

[0008] 一方面,本发明提供了一种基于深度蒙特卡洛树搜索的信源导航方法,该基于深度蒙特卡洛树搜索的信源导航方法包括:

[0009] 获取待导航的智能体在历史时间步内的环境信息和执行的动作信息;

[0010] 通过预设的第一神经网络,基于所述历史时间步内的环境信息以及执行的动作信息,预测出智能体在当前时间步的各个方向的动作概率;

[0011] 以预测的当前时间步的各个方向的动作概率作为蒙特卡洛树搜索算法的先验知识,通过蒙特卡洛树搜索算法选择智能体在当前时间步内的最佳执行动作;

[0012] 结合每一时间步的最佳执行动作,得到智能体移动至信源的最优路径。

[0013] 可选地,所述预设的第一神经网络为长短期记忆人工记忆神经网络。

[0014] 进一步地,在通过蒙特卡洛树搜索算法选择智能体在当前时间步内的最佳执行动作时,在蒙特卡洛树搜索算法的模拟阶段,所述方法还包括:

[0015] 将预测的动作概率和当前节点的状态信息输入预设的第二神经网络,通过预设的第二神经网络为当前节点分配奖励值,再将奖励值反向传播至根节点。

[0016] 可选地,所述预设的第二神经网络为卷积神经网络。

[0017] 进一步地,在通过预设的第二神经网络为当前节点分配奖励值,再将奖励值反向传播至根节点之后,所述方法还包括:

[0018] 用获取的奖励值继续训练所述预设的第二神经网络,以提高预测能力。

[0019] 另一方面,本发明还提供了一种基于深度蒙特卡洛树搜索的信源导航装置,该基于深度蒙特卡洛树搜索的信源导航装置包括:

[0020] 历史环境信息及动作信息获取模块,用于获取待导航的智能体在历史时间步内的环境信息和执行的动作信息;

[0021] 动作概率预测模块,用于通过预设的第一神经网络,基于所述历史环境信息及动作信息获取模块所获取的历史时间步内的环境信息以及执行的动作信息,预测出智能体在当前时间步的各个方向的动作概率;

[0022] 最佳执行动作决策模块,用于以所述动作概率预测模块所预测的当前时间步的各个方向的动作概率作为蒙特卡洛树搜索算法的先验知识,通过蒙特卡洛树搜索算法选择智能体在当前时间步内的最佳执行动作;

[0023] 最优路径获取模块,用于结合所述最佳执行动作决策模块输出的每一时间步的最佳执行动作,得到智能体移动至信源的最优路径。

[0024] 可选地,所述预设的第一神经网络为长短期记忆人工记忆神经网络。

[0025] 进一步地,在通过蒙特卡洛树搜索算法选择智能体在当前时间步内的最佳执行动作时,在蒙特卡洛树搜索算法的模拟阶段,所述最佳执行动作决策模块还用于:

[0026] 将预测的动作概率和当前节点的状态信息输入预设的第二神经网络,通过预设的第二神经网络为当前节点分配奖励值,再将奖励值反向传播至根节点。

[0027] 可选地,所述预设的第二神经网络为卷积神经网络。

[0028] 进一步地,在通过预设的第二神经网络为当前节点分配奖励值,再将奖励值反向传播至根节点之后,所述最佳执行动作决策模块还用于:

[0029] 用获取的奖励值继续训练所述预设的第二神经网络,以提高预测能力。

[0030] 再一方面,本发明还提供了一种电子设备,其包括处理器和存储器;其中,存储器中存储有至少一条指令,所述指令由处理器加载并执行以实现上述方法。

[0031] 又一方面,本发明还提供了一种计算机可读存储介质,所述存储介质中存储有至少一条指令,所述指令由处理器加载并执行以实现上述方法。

[0032] 本发明提供的技术方案带来的有益效果至少包括:

[0033] 针对部分可观测环境下的智能体路径规划问题,本发明使用了学习型的社会反应

模型来预测整个行动空间规划过程中的智能体动态。本发明应用在智能体系统中,智能体在移动过程中观测环境信息的同时,不断的训练循环神经网络的参数,以提高移动过程中状态的预测能力,使得奖励分配更加合理化,再结合蒙特卡洛树搜索对时间动作序列数据处理,以提高智能体到达信源位置的路径规划能力。从而解决了在部分可观测环境下,智能体高效路径规划的问题。

附图说明

[0034] 为了更清楚地说明本发明实施例中的技术方案,下面将对实施例描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本发明的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他的附图。

[0035] 图1为本发明实施例提供的基于深度蒙特卡洛树搜索的信源导航方法的流程示意图;

[0036] 图2为本发明实施例提供的基于蒙特卡洛树搜索和神经网络的算法框架图;

[0037] 图3为蒙特卡洛树搜索算法的执行流程图。

具体实施方式

[0038] 为使本发明的目的、技术方案和优点更加清楚,下面将结合附图对本发明实施方式作进一步地详细描述。

[0039] 第一实施例

[0040] 在部分可观测环境下,智能体配备了传感器,可以感知有限范围内的环境信息,智能体需要根据有限的环境信息来进行路径规划,其中就涉及到如何在每一个时间步做出动作决策以及如何像传统强化学习一样有明确的值函数来定义每一时间步的奖励。根据这些问题,本实施例提供了一种基于深度蒙特卡洛树搜索的信源导航方法,应用于在智能体在部分复杂可观测环境,针对获得有限状态信息下出现的高效路径规划问题。便于在给定条件下确定未来一段时间内的智能体动作序列,在整个环境下所有动作组成一条最优路径。

[0041] 本实施例的信源导航方法可以由电子设备实现,该电子设备可以是终端或者服务器。该方法的执行流程如图1所示,包括以下步骤:

[0042] S101,获取待导航的智能体在历史时间步内的环境信息和执行的动作信息;

[0043] S102,通过预设的第一神经网络,基于所述历史时间步内的环境信息以及执行的动作信息,预测出智能体在当前时间步的各个方向的动作概率;

[0044] S103,以预测的当前时间步的各个方向的动作概率作为蒙特卡洛树搜索算法的先验知识,通过蒙特卡洛树搜索选择智能体在当前时间步内的最佳执行动作;

[0045] S104,结合每一时间步的最佳执行动作,得到智能体移动至信源的最优路径。

[0046] 进一步地,在通过蒙特卡洛树搜索算法选择智能体在当前时间步内的最佳执行动作时,在蒙特卡洛树搜索算法的模拟阶段,所述方法还包括:

[0047] 将预测的动作概率和当前节点的状态信息输入预设的第二神经网络,通过预设的第二神经网络为当前节点分配奖励值,再将奖励值反向传播至根节点。

[0048] 通过采用上述技术方案,在给定信号场下,经过训练,智能体能在移动过程中预测

下一时刻状态,并且能低成本高效的从初始位置向信源位置移动。

[0049] 本实施例的信源导航方法将循环神经网络融入蒙特卡洛树搜索中,蒙特卡洛树搜索使用循环神经网络的预测动作概率来估算一个搜索树中每一个状态的值。随着进行了越来越多的模拟,搜索树会变得越来越庞大,而相关的值也会变得越来越精确。通过选取值更高的子树,用于选择行动的策略概率在搜索的过程中会一直随着时间而有所改进。通过将监督和强化学习两种方法结合起来从而训练循环神经网络,引入了一个新搜索算法,这一算法成功的整合了神经网络评估和蒙特卡洛树模拟算法,从而可实现低成本且高效的信源导航。

[0050] 而且,需要说明的是,在强化学习系统变得越来越普通的时候,引发期望行为的奖励机制的设计变得更加重要和困难。另外,在奖励分配中,我们发现奖励形成对强化学习速度至关重要,奖励范围是一个重要的参数,它涉及整形的有效性,并证明它对于一个简单的强化学习算法在运行时间上具有最强的影响力。因此,为了能够给予合理的奖励分配,本实施例利用神经网络来学习奖励的分配,奖励分配是由卷积神经网络对获取的状态信息和动作信息做出预测,通过不断训练,可以使神经网络的输出逼近真实奖励。

[0051] 此外,动态环境中的路径规划可以表述为顺序决策问题。序列在许多应用程序和系统中起着重要的作用,在本实施例中,时间动作序列是由循环神经网络对K个历史时间步内的状态信息和动作信息做出下一时间步动作概率分配,为蒙特卡洛树搜索提供先验知识。然后,采用蒙特卡洛树搜索根据循环神经网络预测的动作概率和当前状态,选择当前时间步内的最佳动作。

[0052] 而且,需要说明的是,与传统蒙特卡洛树搜索的模拟操作不同,传统的模拟操作模拟的是最终状态。从添加节点的状态开始,模拟运行,运行是随机执行的或者是根据启发式策略执行的,直到达到最终状态。本方法的模拟操作模拟的是奖励值的分配,通过获取当前节点的状态作为循环神经网络的输入,输出在当前状态下应该分配的奖励值,并且将该奖励值反向传播到根节点上。

[0053] 以蒙特卡洛树搜索算法对动作的输出为决策主线,循环神经网络对时间动作序列的处理为时间主线,两者结合可促进智能体在线训练和学习。

[0054] 具体地,在本实施例中,所述预设的第一神经网络为长短期记忆人工记忆神经网络LSTM。所述预设的第二神经网络为卷积神经网络。

[0055] 本实施例的基于深度蒙特卡洛树搜索的信源导航方法的框架如图2所示,蒙特卡洛树搜索在博弈游戏上有着不错的效果,于是本实施例将蒙特卡洛树搜索引入到导航问题上,不过与博弈游戏不同的是,导航处理的是连续状态空间上的动作决策问题,针对这个问题,本实施例提出将神经网络融入蒙特卡洛树搜索中,用神经网络去处理庞大复杂的状态空间数据。具体步骤如下:

[0056] ①智能体执行动作,并观测每一步动作的环境信息,

[0057] ②取智能体前K个时间步(不包括当前时间步)所观测到的环境信息以及执行的动作输入到LSTM网络中,输出当前时间步的各个方向的动作概率向量,

[0058] ③将预测的动作概率信息和当前观测到的状态信息输入蒙特卡洛树搜索,并作为根节点信息,

[0059] ④对根节点进行选择、扩展、模拟、反向传播操作

[0060] ⑤模拟过程中将动作概率信息和当前节点的状态信息传入卷积神经网络,为当前节点分配奖励值,再将奖励值反向传播,

[0061] ⑥重复④、⑤直至满足蒙特卡洛树搜索次数,并输出当前时间步最佳下一步动作,

[0062] ⑦循环①-⑥,直至满足程序迭代终止条件。

[0063] 下面介绍蒙特卡洛树搜索以及说明算法中时间动作序列数据处理和奖励分配的过程。

[0064] 蒙特卡洛树搜索:

[0065] 蒙特卡洛树搜索是基于特定域状态空间的蒙特卡洛模拟的随机采样的最佳优先搜索方法,意味着根据随机模拟的结果做出决策。MCTS的执行流程如图3所示,由四个步骤组成,这些步骤将反复执行,直到达到计算阈值为止,即设置的迭代次数,内存使用上限或时间限制。每次迭代的四个步骤为:

[0066] • 选择:从根节点开始,根据选择策略以递归方式选择子节点。当到达不代表终端状态的叶节点时,将其选择进行扩展。

[0067] 在该步骤中,需要一项策略来探索树以做出有意义的决定,并最终收敛到最有价值的树。应用于树的上限置信区间(UCT),该函数用于最大化多臂机的报酬,UCT平衡了对奖励节点的利用,同时允许探索访问较少的节点。确定给定当前节点选择哪个子代的策略是使以下等式最大化的策略:

[0068] $$V_i + C\sqrt{\frac{lnn_p}{n_i}}$$

[0069] V_i是基于定义的主动策略的当前子代的得分。在第二项中,n_p是节点的访问次数和当前子节点的访问次数。C是通过实验确定的勘探常数。当子节点的访问计数高于阈值T时,将应用UCT。当节点的访问计数低于此阈值时,将随机选择一个子节点扩展。

[0070] • 扩展:给定可用的动作序列,所有子节点都将添加到所选叶节点。

[0071] • 模拟:从添加节点的状态开始,模拟运行。运行是随机执行的,或者是根据启发式策略执行的,直到达到最终状态。

[0072] • 反向传播:模拟的结果立即从所选节点传播到根节点。对于在选择阶段选择的每个节点,沿树更新统计信息,并增加访问次数。

[0073] 时间动作序列数据处理:

[0074] 在复杂动态环境中,许多进行的任务和运动计划需要对可能的未来环境进行有效的探索,在现实世界中的顺序决策问题(例如机器人技术)中,采样的收集顺序至关重要,尤其是在机器人需要优化时间上非平稳的目标函数时。无模型强化学习已在许多挑战性任务中被证明是成功的,但在需要长期计划的任务上却表现不佳,在本发明中,我们将蒙特卡洛树搜索与长短期记忆人工记忆神经网络(LSTM)融合,使得强化学习和深度学习相辅相成。使用LSTM对时间动作序列的处理来预测动作概率的流程如下:

[0075] ①智能体在行进过程中将每一时刻观测到的状态信息T_k和蒙特卡洛树搜索输出的动作决策信息保存,

[0076] ②在当前时间T_t下,读取前6个时刻的相关信息,作为LSTM的输入,输出当前时间下各个动作的概率预测向量,

[0077] ③蒙特卡洛树搜索以该动作预测概率为先验知识,做出当前时间的最佳动作决策。

[0078] 奖励分配:

[0079] 根据以上蒙特卡洛树搜索过程可以注意到,模拟阶段后的反馈对整个蒙特卡洛树搜索至关重要。在模拟阶段,本实施例引入卷积神经网络模拟智能体在行进过程中获得的奖励值,使卷积神经网络替代强化学习值函数,逼近真实奖励,奖励分配的流程如下:

[0080] ①蒙特卡洛树搜索扩展至当前节点,对当前节点进行模拟,

[0081] ②获取当前节点观测到的状态信息S_i以及该时间步的动作预测概率$P(p_i^0,p_i^1,p_i^2,p_i^3)$,

[0082] ③将以上信息输入卷积神经网络,输出当前节点应分配得到的奖励值,同时用该奖励值继续训练卷积神经网络,提高预测能力。

[0083] 综上,针对部分可观测环境下的智能体路径规划问题,本实施例使用了学习型的社会反应模型来预测整个行动空间规划过程中的智能体动态。该方法应用在智能体系统中,智能体在移动过程中观测环境信息的同时,不断训练循环神经网络的参数,以提高移动过程中状态预测能力,使得奖励分配更合理,再结合蒙特卡洛树搜索对时间动作序列数据处理,以提高智能体到达信源位置的路径规划能力。从而解决了在部分可观测环境下,智能体高效路径规划的问题。

[0084] 第二实施例

[0085] 本实施例提供了一种基于深度蒙特卡洛树搜索的信源导航装置,包括:

[0086] 历史环境信息及动作信息获取模块,用于获取待导航的智能体在历史时间步内的环境信息和执行的动作信息;

[0087] 动作概率预测模块,用于通过预设的第一神经网络,基于所述历史环境信息及动作信息获取模块所获取的历史时间步内的环境信息以及执行的动作信息,预测出智能体在当前时间步的各个方向的动作概率;

[0088] 最佳执行动作决策模块,用于以所述动作概率预测模块所预测的当前时间步的各个方向的动作概率作为蒙特卡洛树搜索算法的先验知识,通过蒙特卡洛树搜索算法选择智能体在当前时间步内的最佳执行动作;

[0089] 最优路径获取模块,用于结合所述最佳执行动作决策模块输出的每一时间步的最佳执行动作,得到智能体移动至信源的最优路径。

[0090] 本实施例的基于深度蒙特卡洛树搜索的信源导航装置与上述第一实施例的基于深度蒙特卡洛树搜索的信源导航方法相对应;其中,该基于深度蒙特卡洛树搜索的信源导航装置中的各功能模块所实现的功能与上述的基于深度蒙特卡洛树搜索的信源导航方法中的各流程步骤一一对应;故,在此不再赘述。

[0091] 第三实施例

[0092] 本实施例提供一种电子设备,其包括处理器和存储器;其中,存储器中存储有至少一条指令,所述指令由处理器加载并执行,以实现第一实施例的方法。

[0093] 该电子设备可因配置或性能不同而产生比较大的差异,可以包括一个或一个以上处理器(central processing units,CPU)和一个或一个以上的存储器,其中,存储器中存储有至少一条指令,所述指令由处理器加载并执行上述方法。

[0094] 第四实施例

[0095] 本实施例提供一种计算机可读存储介质,该存储介质中存储有至少一条指令,所述指令由处理器加载并执行,以实现上述第一实施例的方法。其中,该计算机可读存储介质可以是ROM、随机存取存储器、CD-ROM、磁带、软盘和光数据存储设备等。其内存储的指令可由终端中的处理器加载并执行上述方法。

[0096] 此外,需要说明的是,本发明可提供为方法、装置或计算机程序产品。因此,本发明实施例可采用完全硬件实施例、完全软件实施例或结合软件和硬件方面的实施例的形式。而且,本发明实施例可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质上实施的计算机程序产品的形式。

[0097] 本发明实施例是参照根据本发明实施例的方法、终端设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、嵌入式处理机或其他可编程数据处理终端设备的处理器以产生一个机器,使得通过计算机或其他可编程数据处理终端设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0098] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理终端设备以特定方式工作的计算机可读存储器中,使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品,该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。这些计算机程序指令也可装载到计算机或其他可编程数据处理终端设备上,使得在计算机或其他可编程终端设备上执行一系列操作步骤以产生计算机实现的处理,从而在计算机或其他可编程终端设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0099] 还需要说明的是,在本文中,诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来,而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、物品或者终端设备不仅包括那些要素,而且还包括没有明确列出的其他要素,或者是还包括为这种过程、方法、物品或者终端设备所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括所述要素的过程、方法、物品或者终端设备中还存在另外的相同要素。

[0100] 最后需要说明的是,以上所述是本发明优选实施方式,应当指出,尽管已描述了本发明优选实施例,但对于本技术领域的技术人员来说,一旦得知了本发明的基本创造性概念,在不脱离本发明所述原理的前提下,还可以做出若干改进和润饰,这些改进和润饰也应视为本发明的保护范围。所以,所附权利要求意欲解释为包括优选实施例以及落入本发明实施例范围的所有变更和修改。

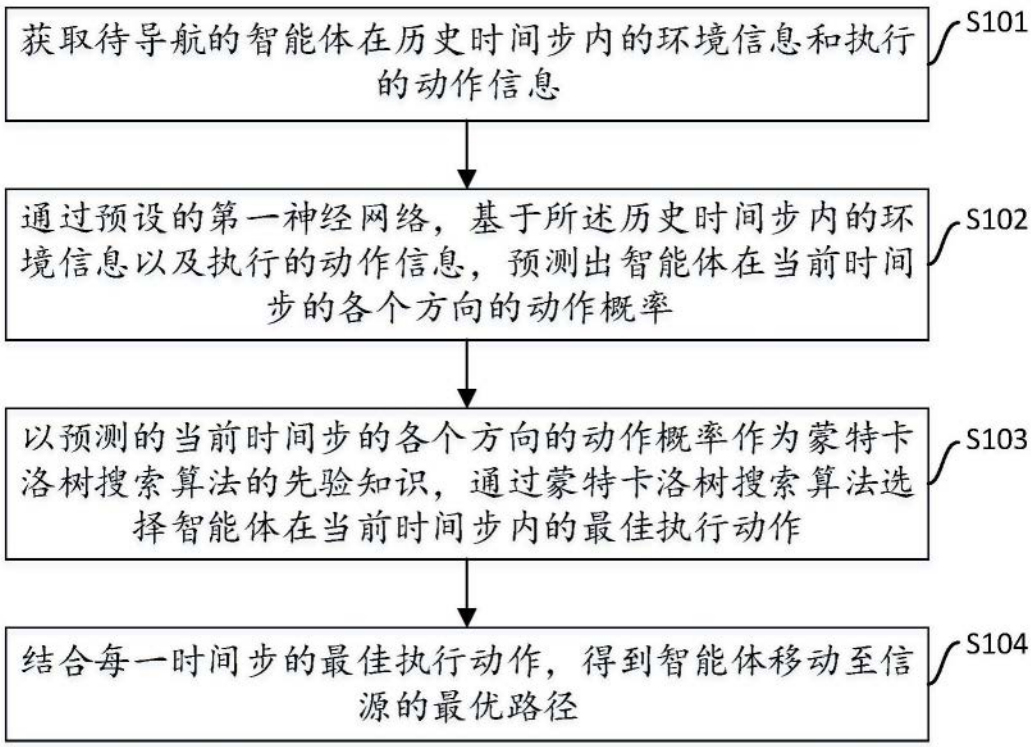

获取待导航的智能体在历史时间步内的环境信息和执行的动作信息
S101
通过预设的第一神经网络，基于所述历史时间步内的环境信息以及执行的动作信息，预测出智能体在当前时间步的各个方向的动作概率
S102
以预测的当前时间步的各个方向的动作概率作为蒙特卡洛树搜索算法的先验知识，通过蒙特卡洛树搜索算法选择智能体在当前时间步内的最佳执行动作
S103
结合每一时间步的最佳执行动作，得到智能体移动至信源的最优路径
S104

图1

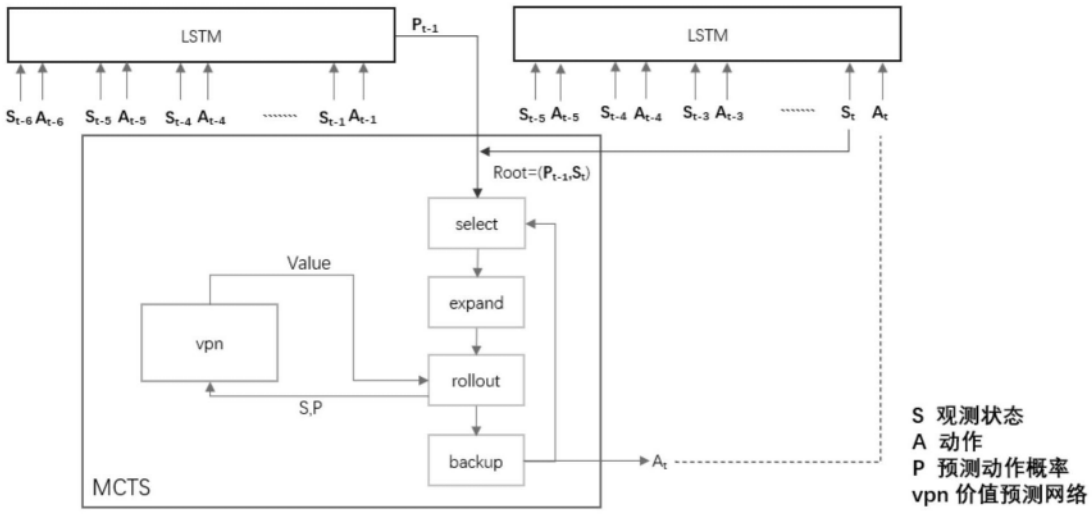

LSTM
LSTM
Root=(Pt-1,St)
select
expand
rollout
backup
Value
vpn
S,P
MCTS
S 观测状态
A 动作
P 预测动作概率
vpn 价值预测网络

图2

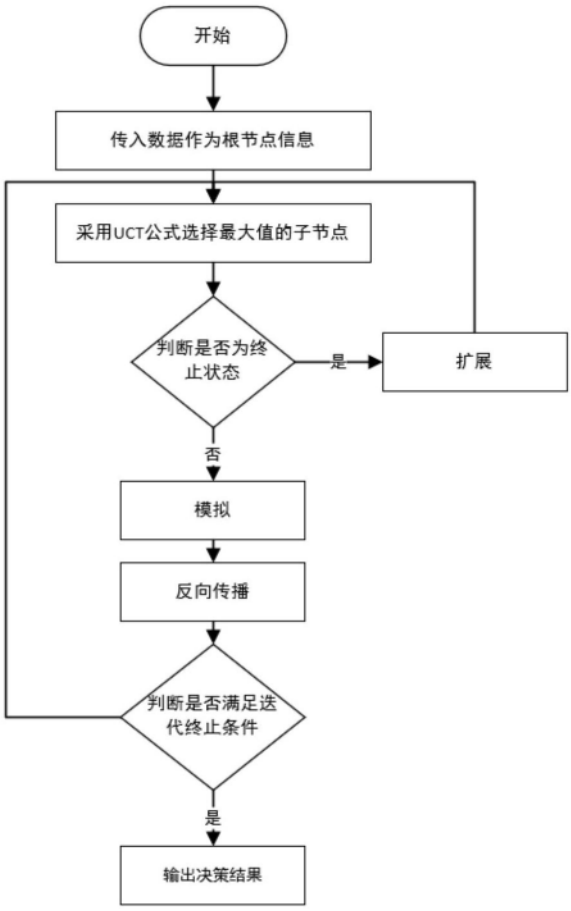

开始
传入数据作为根节点信息
采用UCT公式选择最大值的子节点
判断是否为终止状态
是
扩展
否
模拟
反向传播
判断是否满足迭代终止条件
是
输出决策结果

图3