(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115857556 A
(43)申请公布日 2023.03.28

(21)申请号 202310046016.5

(22)申请日 2023.01.30

(71)申请人 中国人民解放军96901部队
地址 100094 北京市海淀区北清路109号

(72)发明人 武应华 赵国宏 宋天莉 张帅

(74)专利代理机构 北京众元弘策知识产权代理事务所(普通合伙) 11462
专利代理师 白元群

(51)Int.Cl.
G05D 1/10(2006.01)
G06F 18/214(2023.01)

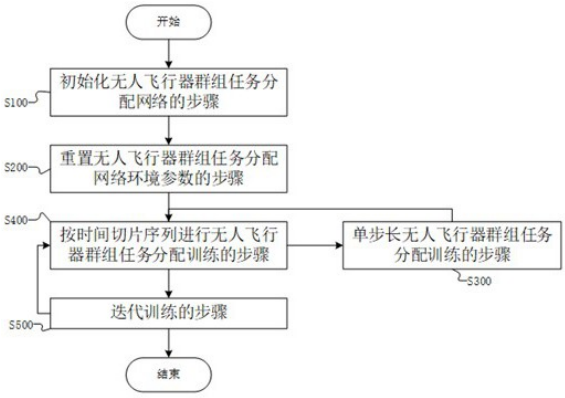
权利要求书4页 说明书11页 附图2页

(54)发明名称

一种基于强化学习的无人飞行器协同探测规划方法

(57)摘要

本发明公开了一种基于强化学习的无人飞行器协同探测规划方法。本发明的方法包括初始化无人飞行器群组任务分配网络的步骤,重置无人飞行器群组任务分配网络环境参数的步骤,单步长无人飞行器群组任务分配训练的步骤,按时间切片序列进行无人飞行器群组任务分配训练的步骤,迭代训练的步骤。本发明基于Q-learning强化学习技术,构造了强化学习无人飞行器群组任务分配网络,提出了将强化学习方法用于工程实现的流程框架,实现了一种智能化强化学习无人飞行器协同探测规划方法。本发明能够节约计算资源,适于工程应用。



CN 115857556 A

1.一种基于强化学习的无人飞行器协同探测规划方法，其特征在于包括：初始化无人飞行器群组任务分配网络的步骤，重置无人飞行器群组任务分配网络环境参数的步骤，单步长无人飞行器群组任务分配训练的步骤，按时间切片序列进行无人飞行器群组任务分配训练的步骤，迭代训练的步骤；其中，

所述初始化无人飞行器群组任务分配网络的步骤中，基于Q-learning强化学习方法构造部署在单个无人飞行器的强化学习任务分配网络，包括状态集合S、动作集合A、动作选择策略π、奖励函数R、奖励衰减因子γ、探索率ε和学习速率α；

所述状态集合S是由状态s_t构成的集合空间，所述状态s_t包括在时间切片t时探测任务分配场景中的m个探测区域的状态和n架无人飞行器所处的环境状态；

所述动作集合A是由的动作a_t构成的集合空间，所述动作a_t包括时间切片t时每个无人飞行器根据状态s_t选择的动作和动作执行后的结果；

所述动作选择策略π表示选择动作的策略，所述动作选择策略π是在状态为s时无人飞行器采取动作a的策略，记为$\pi(a|s), a \in A, s \in S$；

所述奖励函数R是探测任务分配的奖励函数，所述奖励函数R是在时间切片t时，在状态s_t下采取动作a_t变化到状态s_{t+1}对于下一时间切片$t+1$时的奖励R_{t+1}；

所述奖励衰减因子γ表示下一状态和奖励的相关性，$\gamma \in [0,1]$；

所述探索率ε是用于避免陷入局部最优的参数，$\varepsilon \in (0,1)$；

所述重置无人飞行器群组任务分配网络环境参数的步骤中，设置迭代回合数的初始值e为1，设置初始化观察状态为s_1，设置时间切片初始值t为1，设置Q表中的累积奖励的初值为0，所述Q表是一个二维矩阵，Q表以(s_t, a_t)为索引记录相应的累积奖励；

所述单步长无人飞行器群组任务分配训练的步骤中，计算时间切片t对应的奖励以及累积奖励，包括以下子步骤：

步骤step1，若当前状态s_t为s_1，即若当前时间切片$t=1$，从所有动作中随机选择一个动作a_t作为无人飞行器群的执行动作，若当前状态其他状态时，进行步骤step2；

步骤step2，根据当前状态s_t执行动作a_t后得到下一时间切片的状态s_{t+1}；根据状态s_t、动作a_t和状态s_{t+1}采用奖励函数R计算奖励R_{t+1}；以探索率ε采用动作选择策略$\pi(a|s_{t+1})$从动作集合A中选择下一时间切片的动作a_{t+1}，即下一时间切片的动作a_{t+1}采用以下公式选择，

$$a_{t+1}=\begin{cases}\pi(a|s_{t+1}),\ a \in A & ,\ p=\varepsilon \\ A\text{不等于}\ \pi(a|s_{t+1})\text{其他任意动作} & ,\ p=1-\varepsilon\end{cases}$$

2.其中，p表示无人飞行器群组选取下一时间切片动作时，为避免局部最优，按照均匀分布以概率ε采用策略$\pi(a|s_{t+1})$从动作空间A中选择的动作，按照均匀分布以概率1-ε选择动作空间A中不等于$\pi(a|s_{t+1})$其它动作；

步骤step3，更新Q表中对应当前索引(s_t, a_t)的Q值$Q(s_t, a_t)$，Q值$Q(s_t, a_t)$的更新方式为

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[R_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)],$$

所述按时间切片序列进行无人飞行器群组任务分配训练的步骤，该步骤中，根据当前状态s_t和动作a_t按时间切片序列依次进行单步长无人飞行器群组任务分配训练的步骤，直到训练时间切片满足最大时长或满足终止条件时，停止当前回合训练；

迭代训练的步骤，该步骤中，完成一回合所有时间切片的无人飞行器群组任务分配训练后，回合数e计数增加1，时间切片进入下一回合的训练，直到回合数达到预设的训练样本数为止。

3.如权利要求1所述的基于强化学习的无人飞行器协同探测规划方法，其特征在于，所述状态s_t采用如下表示方法：

$$s_t = (E_{plan}(t), A_{plan}(t), \Psi(t))$$

4.其中，向量$E_{plan}(t) = [p_1, p_2, \dots, p_n]$表示无人飞行器任务分配情况，$p_i$表示无人飞行器前往探测区域的编号，$p_i \in \{0,1,2, \dots, m\}$，$i = 1,2,3, \dots, n$，$p_i = 0$表示该无人飞行器节点尚未分配任务；向量$A_{plan}(t) = [c_1, c_2, \dots, c_m]$表示探测区域的侦察任务承载情况，$c_j \in \{0,1,2,3\}$，$j = 1,2,3, \dots, m$，其中$c_j = 0$表示探测区域尚未分配侦察任务，$c_j = 1$表示该探测区域为部分覆盖侦察，$c_j = 2$表示该探测区域为完全覆盖侦察，$c_j = 3$表示该探测区域为冗余覆盖侦察；向量$\Psi(t) = [(x_1, y_1, z_1), \dots, (x_n, y_n, z_n)]$是无人飞行器规划时刻的三维坐标数据。

5.如权利要求2所述的基于强化学习的无人飞行器协同探测规划方法，其特征在于，所述无人飞行器规划时刻的三维坐标数据是原始三维坐标数据或者对原始三维坐标数据归一化处理后得到的数据。

6.如权利要求3所述的基于强化学习的无人飞行器协同探测规划方法，其特征在于，所有规划时刻的向量$\Psi(t)$相等，即$\Psi(t+1) \equiv \Psi(t)$。

7.如权利要求2所述的基于强化学习的无人飞行器协同探测规划方法，其特征在于，所述奖励R_{t+1}采用如下表示方法：

$$R_{t+1} = \frac{\sum_{j=1}^{m} (c_j^t - c_j^{t+1})}{\max\limits_{i=1,2\dots n} t_{endi} - \min\limits_{i=1,2\dots n} t_{starti}}$$

8.其中，c_j^t、c_j^{t+1}分别是状态s_t和状态s_{t+1}中的第j个探测区域任务承载情况，t_{endi}是无人飞行器节点i前往动作a_t的向量$C_{plan}(t)$中p_i所指定探测区域执行探测任务的结束时刻，$\max\limits_{i=1,2\dots n} t_{endi}$是其中的最大值；$t_{starti}$是无人飞行器节点$i$前往动作$a_t$的向量$C_{plan}(t)$中$p_i$所指定探测区域执行探测任务的开始时刻，$\min\limits_{i=1,2\dots n} t_{starti}$是其中的最小值。

9.如权利要求1所述的基于强化学习的无人飞行器协同探测规划方法，其特征在于，所述动作a_t采用如下表示方法：

$$a_t = (A_{stable}(t), C_{plan}(t))$$

10.其中，向量$A_{stable}(t) = [a_{s1}, a_{s2}, \dots, a_{sn}]$表示无人飞行器的行动策略，$a_{si} \in \{0,1,2\}$

表示无人飞行器i的行动策略，其中，$a_{si}=0$表示无人飞行器节点维持原状，$a_{si}=1$表示无人飞行器节点放弃当前任务，$a_{si}=2$表示无人飞行器节点重新选择任务，$C_{plan}(t)=[p_1,p_2,\dots,p_n]$表示执行行动策略$A_{stable}(t)$后无人飞行器的任务重新分配情况，$p_i\in\{0,1,2,\dots,m\}$。

11.如权利要求1所述的基于强化学习的无人飞行器协同探测规划方法，其特征在于，所述动作选择策略$\pi(a|s_{t+1})$为$\underset{a\in A}{argmax}\,Q(s_{t+1},a)$，表示从动作空间$A$中选取使$Q$值$Q(s_{t+1},a)$最大的动作。

12.一种基于强化学习的无人飞行器协同探测规划装置，其特征在于，包括无人飞行器群组任务分配网络初始化模块、无人飞行器群组任务分配网络环境参数重置模块、单步长无人飞行器群组任务分配训练模块、无人飞行器群组任务分配时间切片序列训练控制模块、迭代训练控制模块；其中，

无人飞行器群组任务分配网络初始化模块，该模块用于基于Q-learning强化学习方法构造部署在单个无人飞行器的强化学习任务分配网络，包括状态集合S、动作集合A、动作选择策略π、奖励函数R、奖励衰减因子γ、探索率ε和学习速率α；

所述状态集合S是由状态s_t构成的集合空间，所述状态s_t包括在时间切片t时探测任务分配场景中的m个探测区域的状态和n架无人飞行器所处的环境状态；

所述动作集合A是由的动作a_t构成的集合空间，所述动作a_t包括时间切片t时每个无人飞行器根据状态s_t选择的动作和动作执行后的结果；

所述动作选择策略π表示选择动作的策略，所述动作选择策略π是在状态为s时无人飞行器采取动作a的策略，记为$\pi(a|s)$，$a\in A, s\in S$；

所述奖励函数R是探测任务分配的奖励函数，所述奖励函数R是在时间切片t时，在状态s_t下采取动作a_t变化到状态s_{t+1}对于下一时间切片$t+1$时的奖励R_{t+1}；

所述奖励衰减因子γ表示下一状态和奖励的相关性，$\gamma\in[0,1]$；

所述探索率ε是用于避免陷入局部最优的参数，$\varepsilon\in(0,1)$；

无人飞行器群组任务分配网络环境参数重置模块，该模块用于设置迭代回合数的初始值e为1，设置初始化观察状态为s_1，设置时间切片初始值t为1，设置Q表中的累积奖励的初值为0，所述Q表是一个二维矩阵，Q表以(s_t,a_t)为索引记录相应的累积奖励；

单步长无人飞行器群组任务分配训练模块，该模块用于计算时间切片t对应的奖励以及累积奖励，包括子模块：

子模块step1，若当前状态s_t为s_1，即若当前时间切片$t=1$，从所有动作中随机选择一个动作a_t作为无人飞行器群的执行动作，若当前状态其他状态时，进入子模块step2；

子模块step2，根据当前状态s_t执行动作a_t后得到下一时间切片的状态s_{t+1}；根据状态s_t、动作a_t和状态s_{t+1}采用奖励函数R计算奖励R_{t+1}；以探索率ε采用动作选择策略$\pi(a|s_{t+1})$从动作集合A中选择下一时间切片的动作a_{t+1}，即下一时间切片的动作a_{t+1}采用以下公式选择，

$$a_{t+1}=\begin{cases}\pi(a|s_{t+1}),\ a\in A & ,\ p=\varepsilon\\ A\text{不等于}\ \pi(a|s_{t+1})\text{其他任意动作} & ,\ p=1-\varepsilon\end{cases}$$

13.其中，p表示无人飞行器群组选取下一时间切片动作时，为避免局部最优，按照均匀分布以概率ε采用策略$\pi(a|s_{t+1})$从动作空间A中选择的动作，按照均匀分布以概率1-ε选择动作空间A中不等于$\pi(a|s_{t+1})$其它动作；

子模块step3，用于更新Q表中对应当前索引(s_t,a_t)的Q值$Q(s_t,a_t)$，Q值$Q(s_t,a_t)$的更新方式为

$$Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha[R_{t+1}+\gamma Q(s_{t+1},a_{t+1})-Q(s_t,a_t)],$$

无人飞行器群组任务分配时间切片序列训练控制模块，该模块用于根据当前状态s_t和动作a_t按时间切片序列依次调用单步长无人飞行器群组任务分配训练模块进行单步长无人飞行器群组任务分配训练，直到训练时间切片满足最大时长或满足终止条件时，停止当前回合训练；

迭代训练控制模块，该模块用于在完成一回合所有时间切片的无人飞行器群组任务分配训练后，控制无人飞行器群组任务分配时间切片序列训练控制模块进入下一回合的训练，并且回合数e计数增加1，直到回合数达到预设的训练样本数为止。

14.一种基于强化学习的无人飞行器协同探测规划平台，其特征在于，包括：

至少一个处理器；以及，

与所述至少一个处理器通信连接的存储器；其中，

所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行如权利要求1至7中任一项所述的基于强化学习的无人飞行器协同探测规划方法。

15.一种计算机可读存储介质，存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现权利要求1至7中任一项所述的基于强化学习的无人飞行器协同探测规划方法。

一种基于强化学习的无人飞行器协同探测规划方法

技术领域

[0001] 本发明属于无人飞行器技术领域，具体涉及一种基于强化学习的无人飞行器协同探测规划方法、装置、平台和计算机存储介质。

背景技术

[0002] 强化学习是人工智能一个重要的领域，其在围棋领域已经取得了突破性的成果。强化学习较强的适应性和自学习特征使得其天然契合复杂任务场景下的制导规划，将是未来主要的发展方向，其用于飞行器任务规划和制导控制领域的研究起步未久，一方面缺少满足应用场景实际需要的成熟计算网络模型，另一方面前受制于计算资源和算法研究进展尚未得到成熟应用。

发明内容

[0003] 本发明的目的在于提供一种满足无人飞行器应用需求、节约计算资源、适于实用的一种基于强化学习的无人飞行器协同探测规划方法，以解决上述背景技术中提出的问题。

[0004] 为实现上述目的，本发明提供了一种基于强化学习的无人飞行器协同探测规划方法，其特征在于包括：初始化无人飞行器群组任务分配网络的步骤，重置无人飞行器群组任务分配网络环境参数的步骤，单步长无人飞行器群组任务分配训练的步骤，按时间切片序列进行无人飞行器群组任务分配训练的步骤，迭代训练的步骤；其中，

所述初始化无人飞行器群组任务分配网络的步骤中，基于Q-learning强化学习方法构造部署在单个无人飞行器的强化学习任务分配网络，包括状态集合S、动作集合A、动作选择策略π、奖励函数R、奖励衰减因子γ、探索率ε和学习速率α；

所述状态集合S是由状态s_t构成的集合空间，所述状态s_t包括在时间切片t时探测任务分配场景中的m个探测区域的状态和n架无人飞行器所处的环境状态；

所述动作集合A是由的动作a_t构成的集合空间，所述动作a_t包括时间切片t时每个无人飞行器根据状态s_t选择的动作和动作执行后的结果；

所述动作选择策略π表示选择动作的策略，所述动作选择策略π是在状态为s时无人飞行器采取动作a的策略，记为$\pi(a \mid s), a \in A, s \in S$；

所述奖励函数R是探测任务分配的奖励函数，所述奖励函数R是在时间切片t时，在状态s_t下采取动作a_t变化到状态s_{t+1}对于下一时间切片$t+1$时的奖励R_{t+1}；

所述奖励衰减因子γ表示下一状态和奖励的相关性，$\gamma \in [0,1]$；

所述探索率ε是用于避免陷入局部最优的参数，$\varepsilon \in (0,1)$；

所述重置无人飞行器群组任务分配网络环境参数的步骤中，设置迭代回合数的初

始值e为1，设置初始化观察状态为s_1，设置时间切片初始值t为1，设置Q表中的累积奖励的初值为0，所述Q表是一个二维矩阵，Q表以(s_t,a_t)为索引记录相应的累积奖励；

所述单步长无人飞行器群组任务分配训练的步骤中，计算时间切片t对应的奖励以及累积奖励，包括以下子步骤：

步骤step1，若当前状态s_t为s_1，即若当前时间切片$t=1$，从所有动作中随机选择一个动作a_t作为无人飞行器群的执行动作，若当前状态其他状态时，进行步骤step2；

步骤step2，根据当前状态s_t执行动作a_t后得到下一时间切片的状态s_{t+1}；根据状态s_t、动作a_t和状态s_{t+1}采用奖励函数R计算奖励R_{t+1}；以探索率ε采用动作选择策略$\pi(a|s_{t+1})$从动作集合A中选择下一时间切片的动作a_{t+1}，即下一时间切片的动作a_{t+1}采用以下公式选择，

$$a_{t+1}=\begin{cases}\pi(a|s_{t+1}),\ a\in A & ,\ p=\varepsilon\\ A\text{不等于}\ \pi(a|s_{t+1})\text{其他任意动作} & ,\ p=1-\varepsilon\end{cases}$$

其中，p表示无人飞行器群组选取下一时间切片动作时，为避免局部最优，按照均匀分布以概率ε采用策略$\pi(a|s_{t+1})$从动作空间A中选择的动作，按照均匀分布以概率1-ε选择动作空间A中不等于$\pi(a|s_{t+1})$其它动作；

步骤step3，更新Q表中对应当前索引(s_t,a_t)的Q值$Q(s_t,a_t)$，Q值$Q(s_t,a_t)$的更新方式为

$$Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha[R_{t+1}+\gamma Q(s_{t+1},a_{t+1})-Q(s_t,a_t)],$$

所述按时间切片序列进行无人飞行器群组任务分配训练的步骤，该步骤中，根据当前状态s_t和动作a_t按时间切片序列依次进行单步长无人飞行器群组任务分配训练的步骤，直到训练时间切片满足最大时长或满足终止条件时，停止当前回合训练；

迭代训练的步骤，该步骤中，完成一回合所有时间切片的无人飞行器群组任务分配训练后，回合数e计数增加1，时间切片进入下一回合的训练，直到回合数达到预设的训练样本数为止。

[0005] 进一步的，所述状态s_t采用如下表示方法：

$$s_t=(E_{plan}(t),A_{plan}(t),\Psi(t))$$

其中，向量$E_{plan}(t)=[p_1,p_2,...,p_n]$表示无人飞行器任务分配情况，$p_i$表示无人飞行器前往探测区域的编号，$p_i\in\{0,1,2,...,m\}$，$i=1,2,3,...,n$，$p_i=0$表示该无人飞行器节点尚未分配任务；向量$A_{plan}(t)=[c_1,c_2,...,c_m]$表示探测区域的侦察任务承载情况，$c_j\in\{0,1,2,3\}$，$j=1,2,3,...,m$，其中$c_j=0$表示探测区域尚未分配侦察任务，$c_j=1$表示该探测区域为部分覆盖侦察，$c_j=2$表示该探测区域为完全覆盖侦察，$c_j=3$表示该探测区域为冗余覆盖侦察；向量$\Psi(t)=[(x_1,y_1,z_1),...,(x_n,y_n,z_n)]$是无人飞行器规划时刻的

三维坐标数据。

[0006] 进一步的，所述无人飞行器规划时刻的三维坐标数据是原始三维坐标数据或者对原始三维坐标数据归一化处理后得到的数据。

[0007] 进一步的，所有规划时刻的向量$\Psi(t)$相等，即$\Psi(t+1) \equiv \Psi(t)$。

[0008] 进一步的，所述奖励R_{t+1}采用如下表示方法：

$$R_{t+1} = \frac{\sum_{j=1}^{m} (c_j^t - c^{t+1}_j)}{\max\limits_{i=1,2\cdots n} t_{endi} - \min\limits_{i=1,2\cdots n} t_{starti}}$$

其中，c_j^t、c^{t+1}_j分别是状态s_t和状态s_{t+1}中的第j个探测区域任务承载情况，t_{endi}是无人飞行器节点i前往动作a_t的向量$C_{plan}(t)$中p_i所指定探测区域执行探测任务的结束时刻，$\max\limits_{i=1,2\cdots n} t_{endi}$是其中的最大值；$t_{starti}$是无人飞行器节点$i$前往动作$a_t$的向量$C_{plan}(t)$中$p_i$所指定探测区域执行探测任务的开始时刻，$\min\limits_{i=1,2\cdots n} t_{starti}$是其中的最小值。

[0009] 进一步的，所述动作a_t采用如下表示方法：

$$a_t = (A_{stable}(t), C_{plan}(t))$$

其中，向量$A_{stable}(t) = [a_{s1}, a_{s2}, \ldots, a_{sn}]$表示无人飞行器的行动策略，$a_{si} \in \{0,1,2\}$表示无人飞行器$i$的行动策略，其中，$a_{si} = 0$表示无人飞行器节点维持原状，$a_{si} = 1$表示无人飞行器节点放弃当前任务，$a_{si} = 2$表示无人飞行器节点重新选择任务，$C_{plan}(t) = [p_1, p_2, \ldots, p_n]$表示执行行动策略$A_{stable}(t)$后无人飞行器的任务重新分配情况，$p_i \in \{0,1,2,\ldots,m\}$。

[0010] 进一步的，所述动作选择策略$\pi(a|s_{t+1})$为$\underset{a \in A}{argmax}\, Q(s_{t+1}, a)$，表示从动作空间$A$中选取使$Q$值$Q(s_{t+1}, a)$最大的动作。

[0011] 本发明还提供了一种基于强化学习的无人飞行器协同探测规划装置，包括无人飞行器群组任务分配网络初始化模块、无人飞行器群组任务分配网络环境参数重置模块、单步长无人飞行器群组任务分配训练模块、无人飞行器群组任务分配时间切片序列训练控制模块、迭代训练控制模块；其中，

无人飞行器群组任务分配网络初始化模块，该模块用于基于Q-learning强化学习方法构造部署在单个无人飞行器的强化学习任务分配网络，包括状态集合S、动作集合A、动作选择策略π、奖励函数R、奖励衰减因子γ、探索率ε和学习速率α；

所述状态集合S是由状态s_t构成的集合空间，所述状态s_t包括在时间切片t时探测任务分配场景中的m个探测区域的状态和n架无人飞行器所处的环境状态；

所述动作集合A是由的动作a_t构成的集合空间，所述动作a_t包括时间切片t时每个

无人飞行器根据状态s_t选择的动作和动作执行后的结果；

所述动作选择策略π表示选择动作的策略，所述动作选择策略π是在状态为s时无人飞行器采取动作a的策略，记为$\pi(a \mid s), a \in A, s \in S$；

所述奖励函数R是探测任务分配的奖励函数，所述奖励函数R是在时间切片t时，在状态s_t下采取动作a_t变化到状态s_{t+1}对于下一时间切片$t+1$时的奖励R_{t+1}；

所述奖励衰减因子γ表示下一状态和奖励的相关性，$\gamma \in [0,1]$；

所述探索率ε是用于避免陷入局部最优的参数，$\varepsilon \in (0,1)$；

无人飞行器群组任务分配网络环境参数重置模块，该模块用于设置迭代回合数的初始值e为1，设置初始化观察状态为s_1，设置时间切片初始值t为1，设置Q表中的累积奖励的初值为0，所述Q表是一个二维矩阵，Q表以(s_t, a_t)为索引记录相应的累积奖励；

单步长无人飞行器群组任务分配训练模块，该模块用于计算时间切片t对应的奖励以及累积奖励，包括子模块：

子模块step1，若当前状态s_t为s_1，即若当前时间切片$t=1$，从所有动作中随机选择一个动作a_t作为无人飞行器群的执行动作，若当前状态其他状态时，进入子模块step2；

子模块step2，根据当前状态s_t执行动作a_t后得到下一时间切片的状态s_{t+1}；根据状态s_t、动作a_t和状态s_{t+1}采用奖励函数R计算奖励R_{t+1}；以探索率ε采用动作选择策略$\pi(a|s_{t+1})$从动作集合A中选择下一时间切片的动作a_{t+1}，即下一时间切片的动作a_{t+1}采用以下公式选择，

$$a_{t+1}=\begin{cases}\pi(a|s_{t+1}),\ a\in A & ,\ p=\varepsilon\\ A\text{不等于}\ \pi(a|s_{t+1})\text{其他任意动作} & ,\ p=1-\varepsilon\end{cases}$$

其中，p表示无人飞行器群组选取下一时间切片动作时，为避免局部最优，按照均匀分布以概率ε采用策略$\pi(a|s_{t+1})$从动作空间A中选择的动作，按照均匀分布以概率1-ε选择动作空间A中不等于$\pi(a|s_{t+1})$其它动作；

子模块step3，用于更新Q表中对应当前索引(s_t, a_t)的Q值$Q(s_t, a_t)$，Q值$Q(s_t, a_t)$的更新方式为

$$Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha[R_{t+1}+\gamma Q(s_{t+1},a_{t+1})-Q(s_t,a_t)],$$

无人飞行器群组任务分配时间切片序列训练控制模块，该模块用于根据当前状态s_t和动作a_t按时间切片序列依次调用单步长无人飞行器群组任务分配训练模块进行单步长无人飞行器群组任务分配训练，直到训练时间切片满足最大时长或满足终止条件时，停止当前回合训练；

迭代训练控制模块，该模块用于在完成一回合所有时间切片的无人飞行器群组任

务分配训练后，控制无人飞行器群组任务分配时间切片序列训练控制模块进入下一回合的训练，并且回合数e计数增加1，直到回合数达到预设的训练样本数为止。

[0012] 本发明还提供了一种基于强化学习的无人飞行器协同探测规划平台，包括：至少一个处理器；以及，与所述至少一个处理器通信连接的存储器；其中，所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行所述的基于强化学习的无人飞行器协同探测规划方法。

[0013] 本发明还提供了一种计算机可读存储介质，存储有计算机程序，所述计算机程序被处理器执行时实现所述的基于强化学习的无人飞行器协同探测规划方法。

[0014] 有益效果

本发明基于Q-learning强化学习技术，构造了强化学习无人飞行器群组任务分配网络，提出了将强化学习方法用于工程实现的流程框架，实现了一种智能化强化学习无人飞行器协同探测规划方法。本发明能够节约计算资源，适于工程应用。

附图说明

[0015] 图1为本发明基于强化学习的无人飞行器协同探测规划方法的流程图。

[0016] 图2为本发明基于强化学习的无人飞行器协同探测规划装置的组成框图。

[0017] 图3为本发明基于强化学习的无人飞行器协同探测规划平台的组成框图。

具体实施方式

[0018] 下面结合附图对本发明的具体实施方式进行详细的说明。

[0019] 实施例1

本实施例用于详细说明本发明的一种基于强化学习的无人飞行器协同探测规划方法。

[0020] 如图1所示，本实施例中的基于强化学习的无人飞行器协同探测规划方法包括初始化无人飞行器群组任务分配网络的步骤，重置无人飞行器群组任务分配网络环境参数的步骤，单步长无人飞行器群组任务分配训练的步骤，按时间切片序列进行无人飞行器群组任务分配训练的步骤，迭代训练的步骤。

[0021] 初始化无人飞行器群组任务分配网络的步骤S100。

[0022] 该步骤中，基于Q-learning强化学习方法构造部署在单个无人飞行器的强化学习任务分配网络，包括状态集合S、动作集合A、动作选择策略π、奖励函数R、奖励衰减因子γ、探索率ε和学习速率α。

[0023] 所述状态集合S是由状态s_t构成的集合空间，所述状态s_t包括在时间切片t时探测任务分配场景中的m个探测区域的状态和n架无人飞行器所处的环境状态。本发明的一些具体实施例中，所述状态s_t可以采用如下表示方法：

$$s_t = (E_{plan}(t), A_{plan}(t), \Psi(t))$$

其中，向量$E_{plan}(t) = [p_1, p_2, ..., p_n]$表示无人飞行器任务分配情况，$p_i$表示无人飞行器前往探测区域的编号，$p_i \in \{0,1,2,...,m\}$，$i = 1,2,3,...,n$，$p_i = 0$表示该无人飞行

器节点尚未分配任务；向量$A_{plan}(t)=[c_1,c_2,...,c_m]$表示探测区域的侦察任务承载情况，$c_j\in\{0,1,2,3\}$，$j=1,2,3,...,m$，其中$c_j=0$表示探测区域尚未分配侦察任务，$c_j=1$表示该探测区域为部分覆盖侦察，$c_j=2$表示该探测区域为完全覆盖侦察，$c_j=3$表示该探测区域为冗余覆盖侦察；向量$\Psi(t)=[(x_1,y_1,z_1),...,(x_n,y_n,z_n)]$是无人飞行器规划时刻的三维坐标数据。

[0024] 在本发明的一些实施例中，为降低计算量提高规划速度，不考虑运动参数步进，规划时刻$\Psi(t+1)\equiv\Psi(t)$，表示坐标是不变的，这样的规划时刻是锁定某个飞行时间段的，它可以是飞行前的离线规划，也可以是在飞行时某个时刻冻结后的在线规划。

[0025] 所述无人飞行器规划时刻的三维坐标数据可以采用原始三维坐标数据，为避免训练过程的维度爆炸问题，优选的，三维坐标数据采用对原始三维坐标数据归一化处理后得到的数据。将无人飞行器原始三维坐标归一化是各个坐标乘以一个比例系数，该比例系数是对应维度的坐标最大值的倒数，以维度x_1为例子，原始坐标x_1归一化后的数据$x_1'=x_1/\max\limits_{i=1\cdots n}x_i$。

[0026] 所述动作集合A是由的动作a_t构成的集合空间，所述动作a_t包括时间切片t时每个无人飞行器根据状态s_t选择的动作和动作执行后的结果。本发明的一些具体实施例中，所述动作a_t采用如下表示方法：

$$a_t=(A_{stable}(t),C_{plan}(t))$$

其中，向量$A_{stable}(t)=[a_{s1},a_{s2},...,a_{sn}]$表示无人飞行器的行动策略，$a_{si}\in\{0,1,2\}$表示无人飞行器$i$的行动策略，其中，$a_{si}=0$表示无人飞行器节点维持原状，$a_{si}=1$表示无人飞行器节点放弃当前任务，$a_{si}=2$表示无人飞行器节点重新选择任务，$C_{plan}(t)=[p_1,p_2,...,p_n]$表示执行行动策略$A_{stable}(t)$后无人飞行器的任务重新分配情况，$p_i\in\{0,1,2,...,m\}$。

[0027] 所述动作选择策略π表示选择动作的策略，所述动作选择策略π是在状态为s时无人飞行器采取动作a的策略，记为$\pi(a\mid s)$，$a\in A,s\in S$。

[0028] 所述奖励函数R是探测任务分配的奖励函数，所述奖励函数R是在时间切片t时，在状态s_t下采取动作a_t变化到状态s_{t+1}对于下一时间切片$t+1$时的奖励R_{t+1}。本发明的一些具体实施例中，所述奖励R_{t+1}采用如下表示方法：

$$R_{t+1}=\frac{\sum_{j=1}^{m}(c_j^t-c^{t+1}_j)}{\max\limits_{i=1,2\cdots n}t_{endi}-\min\limits_{i=1,2\cdots n}t_{starti}}$$

其中，c_j^t、c^{t+1}_j分别是状态s_t和状态s_{t+1}中的第j个探测区域任务承载情况，t_{endi}是无人飞行器节点i前往动作a_t的向量$C_{plan}(t)$中p_i所指定探测区域执行探测任务的结束

时刻，$\max_{i=1,2\cdots n} t_{endi}$是其中的最大值；$t_{starti}$是无人飞行器节点$i$前往动作$a_t$的向量$C_{plan}(t)$中$p_i$所指定探测区域执行探测任务的开始时刻，$\min_{i=1,2\cdots n} t_{starti}$是其中的最小值。

[0029] 所述奖励衰减因子γ表示下一状态和奖励的相关性，γ∈[0,1]。当γ=0表示价值只与当前获得的奖励相关，当γ=1表示所有的后续状态获得的奖励和当前奖励拥有相同的影响因子。γ越大说明接下来的状态-奖励的相关性越大。

[0030] 所述探索率ε是用于避免陷入局部最优的参数，$\varepsilon\in(0,1)$。探索率ε表示在智能体选择动作时，会有一定的概率ε不执行能使得价值函数最大的动作，而是执行随机动作，用于跳出局部最优的寻优结果。

[0031] 对于所述学习速率α，当学习率α越小时，之前训练的结果的权重就越大。本发明的一些实施例中，α取值为$\alpha = \frac{1}{t+1}$，即随着迭代步长而减小，本发明的其他一些实施例中，也可以取一个较小的正数值，例如0.1，具体可根据求解问题实际进行遍历优化。

[0032] 重置无人飞行器群组任务分配网络环境参数的步骤S200，该步骤中，设置迭代回合数的初始值e为1，设置初始化观察状态为s_1，设置时间切片初始值t为1，设置Q表中的累积奖励的初值为0，所述Q表是一个二维矩阵，Q表以(s_t,a_t)为索引记录相应的累积奖励。

[0033] 单步长无人飞行器群组任务分配训练的步骤S300，该步骤计算时间切片t对应的奖励以及累积奖励，完成以下子步骤：

步骤step1，若当前状态s_t为s_1，从所有动作中随机选择一个动作a_t作为无人飞行器群的执行动作，若当前状态其他状态时，进行步骤step2；

步骤step2，根据当前状态s_t执行动作a_t后得到下一时间切片的状态s_{t+1}；根据状态s_t、动作a_t和状态s_{t+1}采用奖励函数R计算奖励R_{t+1}；以探索率ε采用动作选择策略$\pi(a|s_{t+1})$从动作集合A中选择下一时间切片的动作a_{t+1}，即下一时间切片的动作a_{t+1}采用以下公式选择：

$$a_{t+1}=\begin{cases}\pi(a|s_{t+1}),\ a\in A & ,\ p=\varepsilon\\ A\text{不等于}\ \pi(a|s_{t+1})\text{其他任意动作} & ,\ p=1-\varepsilon\end{cases}$$

其中，p表示无人飞行器群组选取下一时间切片动作时，为避免局部最优，按照均匀分布以概率ε采用策略$\pi(a|s_{t+1})$从动作空间A中选择的动作，按照均匀分布以概率1-ε选择动作空间A中不等于$\pi(a|s_{t+1})$其它动作。

[0034] 优选的，本发明的一些实施例中，所述动作选择策略$\pi(a|s_{t+1})$为$\underset{a\in A}{argmax}\ Q(s_{t+1},a)$，表示从动作空间$A$中选取使$Q$值$Q(s_{t+1},a)$最大的动作。

[0035] 步骤step3，更新Q表中对应当前索引(s_t,a_t)的Q值$Q(s_t,a_t)$，Q值$Q(s_t,a_t)$的更新方式为

$$Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha[\mathrm{R}_{t+1}+\gamma Q(s_{t+1},a_{t+1})-Q(s_t,a_t)]$$

所述按时间切片序列进行无人飞行器群组任务分配训练的步骤，该步骤中，根据当前状态s_t和动作a_t按时间切片序列依次进行单步长无人飞行器群组任务分配训练的步骤，直到训练时间切片满足最大时长或满足终止条件时，停止当前回合训练；

迭代训练的步骤，该步骤中，完成一回合所有时间切片的无人飞行器群组任务分配训练后，回合数e计数增加1，时间切片进入下一回合的训练，直到回合数达到预设的训练样本数为止。

[0036] 实施例2

本实施例用于详细说明本发明的基于强化学习的无人飞行器协同探测规划装置。如图2所示，本实施例中的装置包括无人飞行器群组任务分配网络初始化模块、无人飞行器群组任务分配网络环境参数重置模块、单步长无人飞行器群组任务分配训练模块、无人飞行器群组任务分配时间切片序列训练控制模块、迭代训练控制模块；其中，

无人飞行器群组任务分配网络初始化模块，该模块用于基于Q-learning强化学习方法构造部署在单个无人飞行器的强化学习任务分配网络，包括状态集合S、动作集合A、动作选择策略π、奖励函数R、奖励衰减因子γ、探索率ε和学习速率α；

所述状态集合S是由状态s_t构成的集合空间，所述状态s_t包括在时间切片t时探测任务分配场景中的m个探测区域的状态和n架无人飞行器所处的环境状态。本发明的一些具体实施例中，所述状态s_t可以采用如下表示方法：

$$s_t=(E_{plan}(t),A_{plan}(t),\Psi(t))$$

其中，向量$E_{plan}(t)=[p_1,p_2,...,p_n]$表示无人飞行器任务分配情况，$p_i$表示无人飞行器前往探测区域的编号，$p_i\in\{0,1,2,...,m\}$，$i=1,2,3,...,n$，$p_i=0$表示该无人飞行器节点尚未分配任务；向量$A_{plan}(t)=[c_1,c_2,...,c_m]$表示探测区域的侦察任务承载情况，$c_j\in\{0,1,2,3\}$，$j=1,2,3,...,m$，其中$c_j=0$表示探测区域尚未分配侦察任务，$c_j=1$表示该探测区域为部分覆盖侦察，$c_j=2$表示该探测区域为完全覆盖侦察，$c_j=3$表示该探测区域为冗余覆盖侦察；向量$\Psi(t)=[(x_1,y_1,z_1),...,(x_n,y_n,z_n)]$是无人飞行器规划时刻的三维坐标数据。

[0037] 在本发明的一些实施例中，为降低计算量提高规划速度，不考虑运动参数步进，规划时刻$\Psi(t+1)\equiv\Psi(t)$，表示坐标是不变的，这样的规划时刻是锁定某个飞行时间段的，它可以是飞行前的离线规划，也可以是在飞行时某个时刻冻结后的在线规划。

[0038] 所述无人飞行器规划时刻的三维坐标数据可以采用原始三维坐标数据，为避免训练过程的维度爆炸问题，优选的，三维坐标数据采用对原始三维坐标数据归一化处理后得到的数据。将无人飞行器原始三维坐标归一化是各个坐标乘以一个比例系数，该比例系数是对应维度的坐标最大值的倒数，以维度x_1为例子，原始坐标x_1归一化后的数据

$x_1' = x_1 / \max_{i=1\cdots n} x_i$。

[0039] 所述动作集合A是由的动作a_t构成的集合空间，所述动作a_t包括时间切片t时每个无人飞行器根据状态s_t选择的动作和动作执行后的结果。本发明的一些具体实施例中，所述动作a_t采用如下表示方法：

$$a_t = (A_{stable}(t), C_{plan}(t))$$

其中，向量$A_{stable}(t) = [a_{s1}, a_{s2}, \ldots, a_{sn}]$表示无人飞行器的行动策略，$a_{si} \in \{0,1,2\}$表示无人飞行器$i$的行动策略，其中，$a_{si} = 0$表示无人飞行器节点维持原状，$a_{si} = 1$表示无人飞行器节点放弃当前任务，$a_{si} = 2$表示无人飞行器节点重新选择任务，$C_{plan}(t) = [p_1, p_2, \ldots, p_n]$表示执行行动策略$A_{stable}(t)$后无人飞行器的任务重新分配情况，$p_i \in \{0,1,2,\ldots,m\}$。

[0040] 所述动作选择策略π表示选择动作的策略，所述动作选择策略π是在状态为s时无人飞行器采取动作a的策略，记为$\pi(a \mid s), a \in A, s \in S$；

所述奖励函数R是探测任务分配的奖励函数，所述奖励函数R是在时间切片t时，在状态为状态s_t的情况下采取动作a_t对于下一时间切片$t+1$时的价值，或者说奖励$R_{t+1}(s_t, a_t)$。本发明的一些具体实施例中，所述奖励R_{t+1}采用如下表示方法：

$$R_{t+1} = \frac{\sum_{j=1}^{m} (c_j^t - c_j^{t+1})}{\max_{i=1,2\cdots n} t_{endi} - \min_{i=1,2\cdots n} t_{starti}}$$

其中，c_j^t、c_j^{t+1}分别是状态s_t和状态s_{t+1}中的第j个探测区域任务承载情况，t_{endi}是无人飞行器节点i前往动作a_t的向量$C_{plan}(t)$中p_i所指定探测区域执行探测任务的结束时刻，$\max_{i=1,2\cdots n} t_{endi}$是其中的最大值；$t_{starti}$是无人飞行器节点$i$前往动作$a_t$的向量$C_{plan}(t)$中$p_i$所指定探测区域执行探测任务的开始时刻，$\min_{i=1,2\cdots n} t_{starti}$是其中的最小值。

[0041] 所述奖励衰减因子γ表示下一状态和奖励的相关性，$\gamma \in [0,1]$；

所述探索率ε是用于避免陷入局部最优的参数，$\varepsilon \in (0,1)$；

对于所述学习速率α，当学习率α越小时，之前训练的结果的权重就越大。本发明的一些实施例中，α取值为$\alpha = \frac{1}{t+1}$，即随着迭代步长而减小，本发明的其他一些实施例中，也可以取一个较小的正数值，例如0.1，具体可根据求解问题实际进行遍历优化。

[0042] 无人飞行器群组任务分配网络环境参数重置模块，该模块用于设置迭代回合数的初始值e为1，设置初始化观察状态为s_1，设置时间切片初始值t为1，设置Q表中的累积奖励

的初值为0，所述Q表是一个二维矩阵，Q表以(s_t,a_t)为索引记录相应的累积奖励；

单步长无人飞行器群组任务分配训练模块，该模块用于计算时间切片t对应的奖励以及累积奖励，包括子模块：

子模块step1，若当前状态s_t为s_1，即若当前时间切片$t=1$，从所有动作中随机选择一个动作a_t作为无人飞行器群的执行动作，若当前状态其他状态时，进入子模块step2；

子模块step2，根根据当前状态s_t执行动作a_t后得到下一时间切片的状态s_{t+1}；根据状态s_t、动作a_t和状态s_{t+1}采用奖励函数R计算奖励R_{t+1}；以探索率ε采用动作选择策略$\pi(a|s_{t+1})$从动作集合A中选择下一时间切片的动作a_{t+1}，即下一时间切片的动作a_{t+1}采用以下公式选择，

$$a_{t+1}=\begin{cases}\pi(a|s_{t+1}),\ a\in A & ,\ p=\varepsilon\\ A\text{不等于}\ \pi(a|s_{t+1})\text{其他任意动作} & ,\ p=1-\varepsilon\end{cases}$$

其中，p表示无人飞行器群组选取下一时间切片动作时，为避免局部最优，按照均匀分布以概率ε采用策略$\pi(a|s_{t+1})$从动作空间A中选择的动作，按照均匀分布以概率1-ε选择动作空间A中不等于$\pi(a|s_{t+1})$其它动作。

[0043] 优选的，本发明的一些实施例中，所述动作选择策略$\pi(a|s_{t+1})$为$\underset{a\in A}{argmax}\,Q(s_{t+1},a)$，表示从动作空间$A$中选取使$Q$值$Q(s_{t+1},a)$最大的动作。

[0044] 子模块step3，用于更新Q表中对应当前索引(s_t,a_t)的Q值$Q(s_t,a_t)$，Q值$Q(s_t,a_t)$的更新方式为

$$Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha[R_{t+1}+\gamma Q(s_{t+1},a_{t+1})-Q(s_t,a_t)],$$

无人飞行器群组任务分配时间切片序列训练控制模块，该模块用于根据当前状态s_t和动作a_t按时间切片序列依次调用单步长无人飞行器群组任务分配训练模块进行单步长无人飞行器群组任务分配训练，直到训练时间切片满足最大时长或满足终止条件时，停止当前回合训练；

迭代训练控制模块，该模块用于在完成一回合所有时间切片的无人飞行器群组任务分配训练后，控制无人飞行器群组任务分配时间切片序列训练控制模块进入下一回合的训练，并且回合数e计数增加1，直到回合数达到预设的训练样本数为止。

[0045] 本领域技术人员通过上述说明可知，本发明的实施例可提供为方法、装置、或计算机程序产品。因此，本发明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。

[0046] 实施例3

本实施例用于详细说明本发明采用硬件实现的基于强化学习的无人飞行器协同探测规划平台的具体实施方式。

[0047] 如图3所示，本发明第三实施方式涉及一种基于强化学习的无人飞行器协同探测

规划平台，包括至少一个处理器，以及与至少一个处理器通信连接的存储器；其中，存储器存储有可被至少一个处理器执行的指令，指令被至少一个处理器执行，以使至少一个处理器能够执行上述基于强化学习的无人飞行器协同探测规划方法。

[0048] 其中，存储器和处理器采用总线方式连接，总线可以包括任意数量的互联的总线和桥，总线将一个或多个处理器和存储器的各种电路连接在一起。总线还可以通过接口将诸如外围设备、稳压器和功率管理电路等之类的各种其他电路连接在一起，这些都是本领域所公知的。接口在总线和收发机之间提供接口，例如通信接口、用户接口。收发机可以是一个元件，也可以是多个元件，比如多个接收器和发送器，提供用于在传输介质上与各种其他装置通信的单元。经处理器处理的数据通过天线在无线介质上进行传输，进一步，天线还接收数据并将数据传送给处理器。

[0049] 处理器负责管理总线和通常的处理，还可以提供各种功能，包括定时，外围接口，电压调节、电源管理以及其他控制功能。而存储器可以被用于存储处理器在执行操作时所使用的数据。

[0050] 实施例4

本实施例用于详细说明本发明采用计算机可读存储介质实现的具体实施方式。本实施例涉及一种计算机可读存储介质，存储有计算机程序，计算机程序被处理器执行时实现上述方法实施例。

[0051] 本领域技术人员通过上述说明可以理解，实现上述实施例方法中的全部或部分步骤是可以通过程序来指令相关的硬件来完成，该程序存储在一个存储介质中，包括若干指令用以使得一个设备(可以是单片机，芯片等)或处理器(processor)执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括但不限于U盘、移动硬盘、磁性存储器、光学存储器等各种可以存储程序代码的介质。

[0052] 本发明基于Q-learning强化学习技术，构造了强化学习无人飞行器群组任务分配网络，提出了将强化学习方法用于工程实现的流程框架，实现了一种智能化强化学习无人飞行器协同探测规划方法。本发明能够满足应用需求，节约计算资源，适于工程应用。

[0053] 以上仅为发明的优选实施例而已，并不用以限制本发明，凡在本发明的思想原则内所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

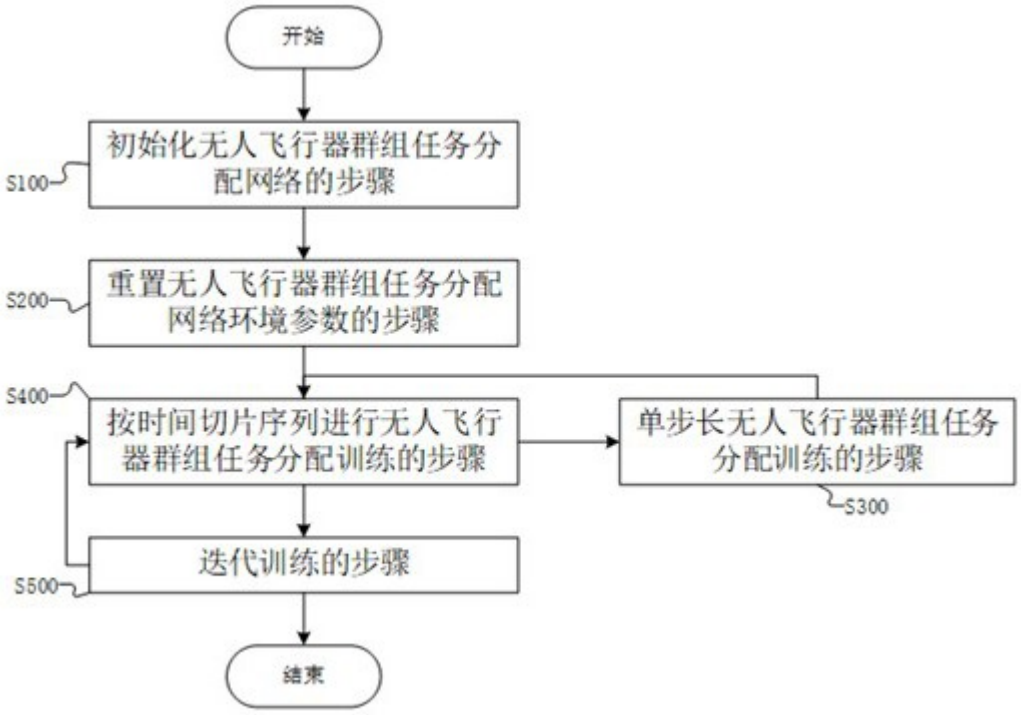

开始
初始化无人飞行器群组任务分配网络的步骤
S100
重置无人飞行器群组任务分配网络环境参数的步骤
S200
S400
按时间切片序列进行无人飞行器群组任务分配训练的步骤
单步长无人飞行器群组任务分配训练的步骤
S300
迭代训练的步骤
S500
结束

图1

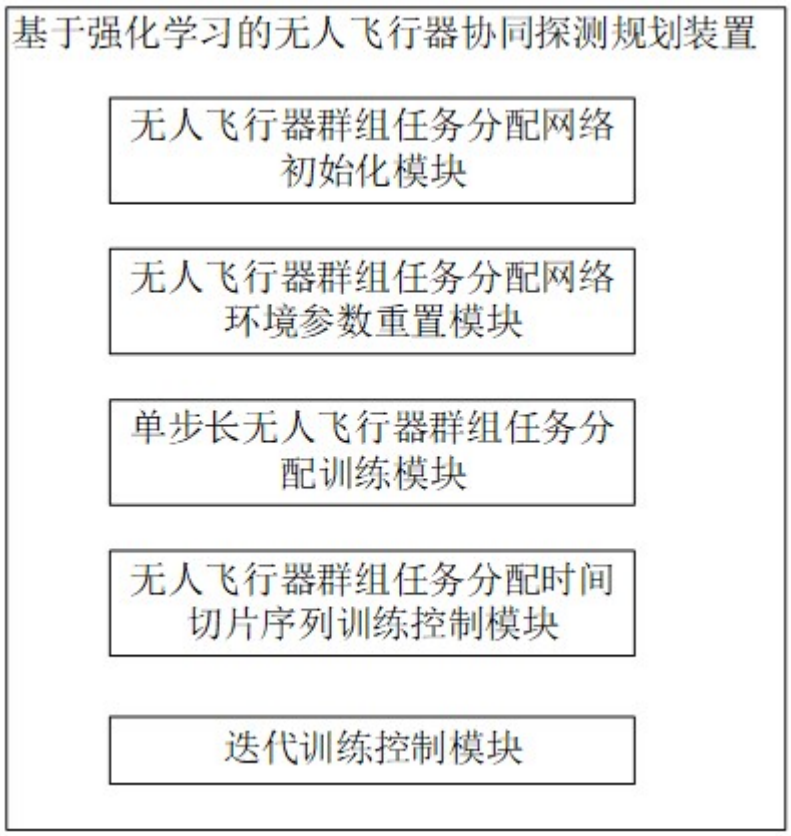

基于强化学习的无人飞行器协同探测规划装置
无人飞行器群组任务分配网络初始化模块
无人飞行器群组任务分配网络环境参数重置模块
单步长无人飞行器群组任务分配训练模块
无人飞行器群组任务分配时间切片序列训练控制模块
迭代训练控制模块

图2

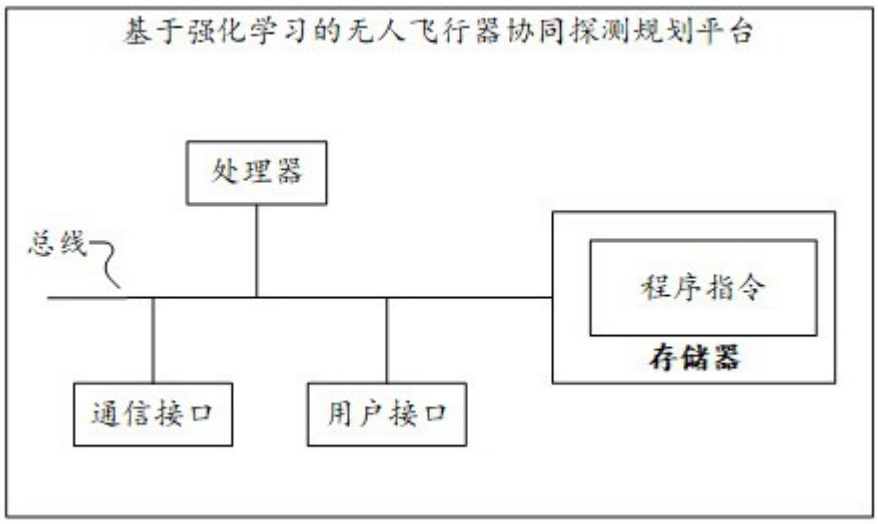


图3