(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113128705 B
(45)授权公告日 2024.02.09

(21)申请号 202110314960.5

(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 113128705 A

(43)申请公布日 2021.07.16

(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 殷楠 段世红 何昊

(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.
G06N 20/20(2019.01)

(56)对比文件
CN 107045655 A,2017.08.15
CN 109760046 A,2019.05.17
CN 110119844 A,2019.08.13
CN 110648049 A,2020.01.03
CN 110726990 A,2020.01.24
CN 110958135 A,2020.04.03
CN 111062359 A,2020.04.24
CN 111291890 A,2020.06.16
CN 111525955 A,2020.08.11
CN 111582311 A,2020.08.25
CN 111695617 A,2020.09.22
CN 111708355 A,2020.09.25
CN 112329948 A,2021.02.05
CN 112364984 A,2021.02.12
US 6591146 B1,2003.07.08
程晓北;沈晶;刘海波;顾国昌;张国印.分层强化学习研究进展.计算机工程与应用.2008,(13),全文.
许诺;杨振伟.稀疏奖励下基于MADDPG算法的多智能体协同.现代计算机.2020,(15),全文.
任红格;向迎帆;李福进.基于内在动机的智能机器人自主发育算法.计算机应用.2015,(09),全文.
席磊;余涛;张孝顺;张泽宇;谭敏.基于狼爬山快速多智能体学习策略的电力系统智能发电控制方法.电工技术学报.30(23),全文.
祁文凯;桑国明.基于延迟策略的最大熵优势演员评论家算法.小型微型计算机系统.2020,(08),全文.

审查员 许晓燕

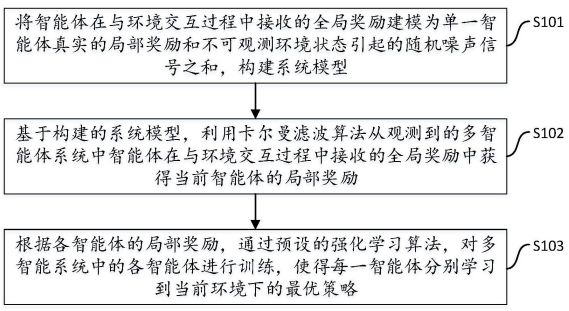
权利要求书4页 说明书9页 附图2页

(54)发明名称

一种智能体最优策略获取方法及装置

(57)摘要

本发明公开了一种智能体最优策略获取方法及装置,该方法包括:将智能体在与环境交互过程中接收的全局奖励建模为单一智能体真实的局部奖励和不可观测环境状态引起的随机噪声信号之和,构建系统模型;利用卡尔曼滤波算法,从观测到的多智能体系统中智能体在与环境交互过程中接收的全局奖励中获得当前智能体对应的局部奖励;根据各智能体对应的局部奖励,通过预设的强化学习算法,对多智能系统中的各智能体进行训练,使得每一智能体分别学习到当前环境下的最优策略。本发明解决了不平稳环境下的多智能体系统信用分配问题,能提高多智能体系统在完成协作任务过程中计算的高效性和鲁棒性。



CN 113128705 B

1.一种智能体最优策略获取方法,适用于多智能系统,其特征在于,包括:

将智能体在与环境交互过程中接收的全局奖励建模为单一智能体真实的局部奖励和不可观测环境状态引起的随机噪声信号之和,构建系统模型;

基于构建的系统模型,利用卡尔曼滤波算法从观测到的多智能体系统中智能体在与环境交互过程中接收的全局奖励中获得当前智能体对应的局部奖励;

根据各智能体所对应的局部奖励,通过预设的强化学习算法,对多智能系统中的各智能体进行训练,使得每一智能体分别学习到当前环境下的最优策略;

所述将智能体在与环境交互过程中接收的全局奖励建模为单一智能体真实的局部奖励和不可观测环境状态引起的随机噪声信号之和,构建系统模型,包括:

将智能体在与环境交互过程中接收的全局奖励建模为单一智能体真实的局部奖励和不可观测环境状态引起的随机噪声信号之和,表达式如下:

$g_t=r(i)_t+b_t$

其中,g_t表示t时刻的全局奖励,$r(i)_t$表示t时刻智能体在状态i下获得的真实奖励,b_t表示t时刻的噪声,而且,$b_{t+1}=b_t+z_t$,$z_t\sim N(\mu,\sigma_\omega^2)$,$z_t$为服从均值为μ,方差为$\sigma_\omega^2$的t时刻的高斯随机变量;

系统模型的表达式如下:

$$\begin{cases} x_t = Ax_{t-1} + \omega_t, & \omega_t \sim N(0, \sum_1) \\ g_t = Cx_t + v_t, & v_t \sim N(0, \sum_2) \end{cases}$$

其中,x_t表示t时刻的状态向量;

$$x_t = \begin{pmatrix} r(1)_t \\ r(2)_t \\ . \\ . \\ . \\ r(s)_t \\ b_t \end{pmatrix}_{(s+1)\times 1}$$

其中,s表示状态数量;

ω_t表示服从0均值高斯分布的系统噪声且协方差矩阵Σ_1如下:

$$\sum_1 = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_\omega^2 \end{pmatrix}_{(s+1)\times(s+1)}$$

其中,σ_ω^2表示系统噪声的协方差;

v_t表示观测误差,属于高斯白噪声,假设无观测误差,$\Sigma_2=0$;A表示状态转移矩阵为,A=I,C表示观测矩阵,$C=(0\ \cdots\ 0\ 1_i\ 0\ \cdots\ 0\ 1)_{1\times(s+1)}$,若处于状态i,则C的第i个元素和最后一个元素值为1,其余为0;

所述基于构建的系统模型,利用卡尔曼滤波算法从观测到的多智能体系统中智能体在与环境交互过程中接收的全局奖励中获得当前智能体对应的局部奖励,包括:

将所述系统模型转换为如下表示：

$$\begin{cases} x_t = Ax_{t-1} + Bu + \varepsilon_t\,, & \varepsilon_t \sim N(0,\ \Sigma_1) \\ g_t = Cx_t + v_t\,, & v_t \sim N(0,\ \Sigma_2) \end{cases}$$

其中，$B=\begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}_{(s+1)\times(s+1)}$，$u=\begin{pmatrix} 0 \\ 0 \\ \cdots \\ 0 \\ \mu \end{pmatrix}$，$\varepsilon_t$为零均值高斯白噪声，$\Sigma_2=0$；

估计由环境的不平稳性引起的噪声的协方差σ^2的值；

根据估计的σ^2的值，基于转换后的系统模型，利用卡尔曼滤波算法从观测到的智能体在与环境交互过程中接收的全局奖励中获得智能体对应的局部奖励；

所述估计由环境的不平稳性引起的噪声的协方差σ^2的值，包括：

初始化$\mu_0=0$，$\sigma_{w0}^2=0.1$，$t=0$；

使用连续的μ_0，σ_{w0}^2运行卡尔曼滤波，迭代n次，记录$x_{t+1}(s+1)$，$x_{t+2}(s+1)$，...，$x_{t+n}(s+1)$；其中，$n>200$；

估计噪声的均值和方差：

$$\mu_t = \frac{1}{n-1}\sum_{i=2}^{n}(x_{t+i}(s+1) - x_{t+i-1}(s+1))$$

$$\sigma_{wt}^2 = \frac{1}{n-1}\sum_{i=2}^{n}(x_{t+i}(s+1) - x_{t+i-1}(s+1) - \mu_t)^2$$

其中，μ_t表示t时刻噪声的均值，σ_{wt}^2表示t时刻噪声的方差。

2.如权利要求1所述的智能体最优策略获取方法，其特征在于，所述预设的强化学习算法为MADDPG算法。

3.一种智能体最优策略获取装置，适用于多智能系统，其特征在于，包括：

系统模型构建模块，用于将智能体在与环境交互过程中接收的全局奖励建模为单一智能体真实的局部奖励和不可观测环境状态引起的随机噪声信号之和，构建系统模型；

卡尔曼滤波模块，用于基于所述系统模型构建模块所构建的系统模型，利用卡尔曼滤波算法从观测到的多智能体系统中智能体在与环境交互过程中接收的全局奖励中获得当前智能体对应的局部奖励；

强化学习训练模块，用于根据所述卡尔曼滤波模块计算出的各智能体所对应的局部奖励，通过预设的强化学习算法，对多智能系统中的各智能体进行训练，使得每一智能体分别学习到当前环境下的最优策略；

所述系统模型构建模块具体用于：

将智能体在与环境交互过程中接收的全局奖励建模为单一智能体真实的局部奖励和不可观测环境状态引起的随机噪声信号之和，表达式如下：

$g_t=r(i)_t+b_t$

其中，g_t表示t时刻的全局奖励，$r(i)_t$表示t时刻智能体在状态i下获得的真实奖励，b_t

表示t时刻的噪声,而且,$b_{t+1}=b_t+z_t$,$z_t\sim N(\mu,\sigma_\omega^2)$,$z_t$为服从均值为μ,方差为$\sigma_\omega^2$的t时刻的高斯随机变量;

系统模型的表达式如下:

$$\begin{cases} x_t = Ax_{t-1} + \omega_t, & \omega_t \sim N(0, \sum_1) \\ g_t = Cx_t + v_t, & v_t \sim N(0, \sum_2) \end{cases}$$

其中,x_t表示t时刻的状态向量;

$$x_t = \begin{pmatrix} r(1)_t \\ r(2)_t \\ . \\ . \\ . \\ r(s)_t \\ b_t \end{pmatrix}_{(s+1)\times 1}$$

其中,s表示状态数量;

ω_t表示服从0均值高斯分布的系统噪声且协方差矩阵Σ_1如下:

$$\sum_1 = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_\omega^2 \end{pmatrix}_{(s+1)\times(s+1)}$$

其中,σ_ω^2表示系统噪声的协方差;

v_t表示观测误差,属于高斯白噪声,假设无观测误差,$\Sigma_2=0$;A表示状态转移矩阵为,A=I,C表示观测矩阵,$C=(0\ \cdots\ 0\ 1_i\ 0\ \cdots\ 0\ 1)_{1\times(s+1)}$,若处于状态i,则C的第i个元素和最后一个元素值为1,其余为0;

所述卡尔曼滤波模块具体用于:

将所述系统模型转换为如下表示:

$$\begin{cases} x_t = Ax_{t-1} + Bu + \varepsilon_t, & \varepsilon_t \sim N(0, \Sigma_1) \\ g_t = Cx_t + v_t, & v_t \sim N(0, \Sigma_2) \end{cases}$$

其中,$B=\begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}_{(s+1)\times(s+1)}$,$u=\begin{pmatrix} 0 \\ 0 \\ \cdots \\ 0 \\ \mu \end{pmatrix}$,$\varepsilon_t$为零均值高斯白噪声,$\Sigma_2=0$;

估计由环境的不平稳性引起的噪声的协方差σ^2的值;

根据估计的σ^2的值,基于转换后的系统模型,利用卡尔曼滤波算法从观测到的智能体在与环境交互过程中接收的全局奖励中获得智能体对应的局部奖励;

所述卡尔曼滤波模块具体还用于:

初始化$\mu_0=0$,$\sigma_{w0}^2=0.1$,t=0;

使用连续的μ_0,σ_{w0}^2运行卡尔曼滤波,迭代n次,记录$x_{t+1}(s+1)$,$x_{t+2}(s+1)$,...,$x_{t+n}(s+1)$;其中,n>200;

估计噪声的均值和方差：

$$\mu_t = \frac{1}{n-1}\sum_{i=2}^{n}(x_{t+i}(s+1) - x_{t+i-1}(s+1))$$

$$\sigma_{wt}^2 = \frac{1}{n-1}\sum_{i=2}^{n}(x_{t+i}(s+1) - x_{t+i-1}(s+1) - \mu_t)^2$$

其中，μ_t表示t时刻噪声的均值，σ_{wt}^2表示t时刻噪声的方差。

4.如权利要求3所述的智能体最优策略获取装置，其特征在于，所述强化学习训练模块所使用的强化学习算法为MADDPG算法。

一种智能体最优策略获取方法及装置

技术领域

[0001] 本发明涉及多智能体强化学习技术领域,特别涉及一种智能体最优策略获取方法及装置。

背景技术

[0002] 随着强化学习在工业、通信等领域的广泛应用,依据现实任务的需求,多个智能体在合作或竞争的环境中如何高效交互成为重点讨论的问题。在部分可观测环境下,相比单智能体,引入多智能体使得状态空间的马尔可夫性失效,即环境不平稳。这是由于智能体不具备全局观测的能力,从而无法获取其他智能体对环境做出的改变,得不到真实的环境状态。

[0003] 当使用同样的全局奖励信号来训练多个智能体时,尤其是在合作任务情况下,要明确单个智能体对团队做出的贡献,降低环境不平稳性对智能体学习的影响,即需要解决多智能体系统中的信用分配问题。对于多智能体系统的单个智能体学习环境下的最优策略,使用个人奖励而非全局奖励训练则会更加有效。

[0004] 针对上述需求,现有技术多从博弈论和广义马尔可夫决策过程MDP角度出发,假设我们对于所处环境模式有足够的领域知识,从而证实了这些方法在降低环境不平稳对智能体学习的影响是可行的。然而当环境中其他智能体的行为产生变化,或者智能体无法观察到的全部的环境变化时,这些技术则会失效。除此之外,也有将滤波方法应用于强化学习框架。证实了在简单的网格环境下,滤波与基于值函数的强化学习方法结合,对提高智能体学习效率是有效的。

[0005] 综上所述,现有的多智能体信用分配强化学习方法难以满足复杂的部分可观测环境下的智能体协作需求,解决效率较低且应用能力较差。

发明内容

[0006] 本发明提供了一种智能体最优策略获取方法及装置,以解决现有的多智能体信用分配强化学习方法难以满足复杂的部分可观测环境下的智能体协作需求,解决效率较低且应用能力较差的技术问题。

[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,本发明提供了一种智能体最优策略获取方法,适用于多智能体系统,所述智能体最优策略获取方法包括:

[0009] 将智能体在与环境交互过程中接收的全局奖励建模为单一智能体真实的局部奖励和不可观测环境状态引起的随机噪声信号之和,构建系统模型;

[0010] 基于构建的系统模型,利用卡尔曼滤波算法从观测到的多智能体系统中智能体在与环境交互过程中接收的全局奖励中获得当前智能体对应的局部奖励;

[0011] 根据各智能体所对应的局部奖励,通过预设的强化学习算法,对多智能系统中的各智能体进行训练,使得每一智能体分别学习到当前环境下的最优策略。

[0012] 进一步地,所述将智能体在与环境交互过程中接收的全局奖励建模为单一智能体真实的局部奖励和不可观测环境状态引起的随机噪声信号之和,构建系统模型,包括:

[0013] 将智能体在与环境交互过程中接收的全局奖励建模为单一智能体真实的局部奖励和不可观测环境状态引起的随机噪声信号之和,表达式如下:

[0014] $g_t=r(i)_t+b_t$

[0015] 其中,g_t表示t时刻的全局奖励,$r(i)_t$表示t时刻智能体在状态i下获得的真实奖励,b_t表示t时刻的噪声,而且,$b_{t+1}=b_t+z_t$,$z_t\sim N(\mu,\sigma_\omega^2)$,$z_t$为服从均值为μ,方差为$\sigma_\omega^2$的t时刻的高斯随机变量;

[0016] 系统模型的表达式如下:

[0017] $$\begin{cases} x_t = Ax_{t-1} + \omega_t, & \omega_t \sim N(0, \sum_1) \\ g_t = Cx_t + v_t, & v_t \sim N(0, \sum_2) \end{cases}$$

[0018] 其中,x_t表示t时刻的状态向量;

[0019] $$x_t = \begin{pmatrix} r(1)_t \\ r(2)_t \\ . \\ . \\ . \\ r(s)_t \\ b_t \end{pmatrix}_{(s+1)\times 1}$$

[0020] 其中,s表示状态数量;

[0021] ω_t表示服从0均值高斯分布的系统噪声且协方差矩阵Σ_1如下:

[0022] $$\sum_1 = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_\omega^2 \end{pmatrix}_{(s+1)\times(s+1)}$$

[0023] 其中,σ_ω^2表示系统噪声的协方差;

[0024] v_t表示观测误差,属于高斯白噪声,假设无观测误差,$\Sigma_2=0$;A表示状态转移矩阵为,A=I,C表示观测矩阵,$C=(0\ \cdots\ 0\ 1_i\ 0\ \cdots\ 0\ 1)_{1\times(s+1)}$,若处于状态i,则C的第i个元素和最后一个元素值为1,其余为0。

[0025] 进一步地,所述基于构建的系统模型,利用卡尔曼滤波算法从观测到的多智能体系统中智能体在与环境交互过程中接收的全局奖励中获得当前智能体对应的局部奖励,包括:

[0026] 将所述系统模型转换为如下表示:

[0027] $$\begin{cases} x_t = Ax_{t-1} + Bu + \varepsilon_t, & \varepsilon_t \sim N(0, \Sigma_1) \\ g_t = Cx_t + v_t, & v_t \sim N(0, \Sigma_2) \end{cases}$$

[0028] 其中,$B=\begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}_{(s+1)\times(s+1)}$,$u=\begin{pmatrix} 0 \\ 0 \\ \cdots \\ 0 \\ \mu \end{pmatrix}$,$\varepsilon_t$为零均值高斯白噪声,$\Sigma_2=0$;

[0029] 估计由环境的不平稳性引起的噪声的协方差σ^2的值;

[0030] 根据估计的σ^2的值,基于转换后的系统模型,利用卡尔曼滤波算法从观测到的智能体在与环境交互过程中接收的全局奖励中获得智能体对应的局部奖励。

[0031] 进一步地,所述估计由环境的不平稳性引起的噪声的协方差σ^2的值,包括:

[0032] 初始化$\mu_0=0$,$\sigma_{w0}^2=0.1$,t=0;

[0033] 使用连续的μ_0,σ_{w0}^2运行卡尔曼滤波,迭代n次,记录$x_{t+1}(s+1)$,$x_{t+2}(s+1)$,…,$x_{t+n}(s+1)$;其中,n>200;

[0034] 估计噪声的均值和方差:

[0035] $$\mu_t=\frac{1}{n-1}\sum_{i=2}^{n}(x_{t+i}(s+1)-x_{t+i-1}(s+1))$$

[0036] $$\sigma_{wt}^2=\frac{1}{n-1}\sum_{i=2}^{n}(x_{t+i}(s+1)-x_{t+i-1}(s+1)-\mu_t)^2$$

[0037] 其中,μ_t表示t时刻噪声的均值,σ_{wt}^2表示t时刻噪声的方差。

[0038] 进一步地,所述预设的强化学习算法为MADDPG算法。

[0039] 另一方面,本发明还提供了一种智能体最优策略获取装置,适用于多智能体系统,所述智能体最优策略获取装置包括:

[0040] 系统模型构建模块,用于将智能体在与环境交互过程中接收的全局奖励建模为单一智能体真实的局部奖励和不可观测环境状态引起的随机噪声信号之和,构建系统模型:

[0041] 卡尔曼滤波模块,用于基于所述系统模型构建模块所构建的系统模型,利用卡尔曼滤波算法从观测到的多智能体系统中智能体在与环境交互过程中接收的全局奖励中获得当前智能体对应的局部奖励;

[0042] 强化学习训练模块,用于根据所述卡尔曼滤波模块计算出的各智能体所对应的局部奖励,通过预设的强化学习算法,对多智能系统中的各智能体进行训练,使得每一智能体分别学习到当前环境下的最优策略。

[0043] 进一步地,所述系统模型构建模块具体用于:

[0044] 将智能体在与环境交互过程中接收的全局奖励建模为单一智能体真实的局部奖励和不可观测环境状态引起的随机噪声信号之和,表达式如下:

[0045] $g_t=r(i)_t+b_t$

[0046] 其中,g_t表示t时刻的全局奖励,$r(i)_t$表示t时刻智能体在状态i下获得的真实奖励,b_t表示t时刻的噪声,而且,$b_{t+1}=b_t+z_t$,$z_t\sim N(\mu,\sigma_\omega^2)$,$z_t$为服从均值为μ,方差为$\sigma_\omega^2$的t时刻的高斯随机变量;

[0047] 系统模型的表达式如下:

[0048] $$\begin{cases}x_t=Ax_{t-1}+\omega_t, & \omega_t\sim N(0,\sum_1)\\ g_t=Cx_t+v_t, & v_t\sim N(0,\sum_2)\end{cases}$$

[0049] 其中,x_t表示t时刻的状态向量;

[0050] $$x_t = \begin{pmatrix} r(1)_t \\ r(2)_t \\ . \\ . \\ . \\ r(s)_t \\ b_t \end{pmatrix}_{(s+1)\times 1}$$

[0051] 其中,s表示状态数量;

[0052] ω_t表示服从0均值高斯分布的系统噪声且协方差矩阵Σ_1如下:

[0053] $$\sum_1 = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_\omega^2 \end{pmatrix}_{(s+1)\times(s+1)}$$

[0054] 其中,σ_ω^2表示系统噪声的协方差;

[0055] v_t表示观测误差,属于高斯白噪声,假设无观测误差,$\Sigma_2=0$;A表示状态转移矩阵为,A=I,C表示观测矩阵,$C=(0\ \cdots\ 0\ 1_i\ 0\ \cdots\ 0\ 1)_{1\times(s+1)}$,若处于状态i,则C的第i个元素和最后一个元素值为1,其余为0。

[0056] 进一步地,所述卡尔曼滤波模块具体用于:

[0057] 将所述系统模型转换为如下表示:

[0058] $$\begin{cases} x_t = Ax_{t-1} + Bu + \varepsilon_t, & \varepsilon_t \sim N(0,\ \Sigma_1) \\ g_t = Cx_t + v_t, & v_t \sim N(0,\ \Sigma_2) \end{cases}$$

[0059] 其中,$B=\begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}_{(s+1)\times(s+1)}$,$u=\begin{pmatrix} 0 \\ 0 \\ \cdots \\ 0 \\ \mu \end{pmatrix}$,$\varepsilon_t$为零均值高斯白噪声,$\Sigma_2=0$;

[0060] 估计由环境的不平稳性引起的噪声的协方差σ^2的值;

[0061] 根据估计的σ^2的值,基于转换后的系统模型,利用卡尔曼滤波算法从观测到的智能体在与环境交互过程中接收的全局奖励中获得智能体对应的局部奖励。

[0062] 进一步地,所述卡尔曼滤波模块具体还用于:

[0063] 初始化$\mu_0=0$,$\sigma_{w0}^2=0.1$,t=0;

[0064] 使用连续的μ_0,σ_{w0}^2运行卡尔曼滤波,迭代n次,记录$x_{t+1}(s+1)$,$x_{t+2}(s+1)$,…,$x_{t+n}(s+1)$;其中,n>200;

[0065] 估计噪声的均值和方差:

[0066] $$\mu_t = \frac{1}{n-1}\sum_{i=2}^{n}(x_{t+i}(s+1) - x_{t+i-1}(s+1))$$

[0067] $$\sigma_{wt}^2 = \frac{1}{n-1}\sum_{i=2}^{n}(x_{t+i}(s+1) - x_{t+i-1}(s+1) - \mu_t)^2$$

[0068] 其中，μ_t表示t时刻噪声的均值，σ_{wt}^2表示t时刻噪声的方差。

[0069] 进一步地，所述强化学习训练模块所使用的强化学习算法为MADDPG算法。

[0070] 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。

[0071] 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。

[0072] 本发明提供的技术方案带来的有益效果至少包括：

[0073] 本发明提出了部分可观测环境下的多智能体系统全局奖励信号的设定方法，并建立了全局奖励与个人贡献和环境噪声的关系；使用卡尔曼滤波方法从带噪声的全局奖励中获取智能体真实的训练回报，使其学习到不同场景下的最优策略。从而实现了部分可观测环境下多智能体执行协作任务过程中的信用分配；适用于部分可观测环境下，降低不平稳环境的影响，使多智能体高效学习在不同环境下完成任务的最优策略；解决了不平稳环境下的多智能体系统信用分配问题，可提高多智能体系统在完成协作任务过程中计算的高效性和鲁棒性。

附图说明

[0074] 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。

[0075] 图1为本发明实施例提供的智能体最优策略获取方法的流程示意图；

[0076] 图2为本发明实施例提供的引入滤波的MADDPG算法框架示意图；

[0077] 图3为本发明实施例提供的智能体最优策略获取方法的具体流程图。

具体实施方式

[0078] 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。

[0079] 第一实施例

[0080] 本实施例提供了一种智能体最优策略获取方法，适用于多智能体系统，可用于解决复杂部分可观测环境下的多智能体协作，针对环境不平稳而出现的智能体信用分配问题。解决实际场景下，未知复杂环境的目标定位和路径导航问题，适用于火场救援的受伤人员位置定位、矿场环境中机器人协作勘探路线导航等领域。该方法可由电子设备实现，该电子设备可以是终端或者服务器。

[0081] 本实施例的方法结合环境的非平稳性和部分可观测性对多智能体系统的影响，建立起具有普遍意义的全局奖励和智能体自身贡献的联系。最终在多智能体训练过程中，使用基于卡尔曼滤波的信用分配方法，从观测到的带噪声的全局奖励中获得智能体局部奖励，用于单个智能体学习最优策略。

[0082] 具体地，本实施例所提供的方法的执行流程如图1所示，包括以下步骤：

[0083] S101，将智能体在与环境交互过程中接收的全局奖励建模为单一智能体真实的局

部奖励和不可观测环境状态引起的随机噪声信号之和,构建系统模型;

[0084] S102,基于构建的系统模型,利用卡尔曼滤波算法从观测到的多智能体系统中智能体在与环境交互过程中接收的全局奖励中获得当前智能体的局部奖励;

[0085] S103,根据各智能体的局部奖励,通过预设的强化学习算法,对多智能系统中的各智能体进行训练,使得每一智能体分别学习到当前环境下的最优策略。

[0086] 进一步地,需要说明的是,在合作任务情况下,要明确单个智能体对团队做出的贡献,对于系统中的单个智能体学习环境下的最优策略,使用个人奖励而非全局奖励训练则会更加有效。环境的非平稳性、部分可观测性和全局奖励的影响可以认为是用原始奖励设置的非平稳函数替代真实奖励值。对于一个智能体来说,环境中存在一个或多个不可观测的状态变量都会影响到全局奖励信号,而其他智能体的存在或者环境变化都属于不可观测的状态,故将这些不可观测的状态变量对于全局奖励信号的影响当作环境噪声来处理。基于此,本实施例将单一智能体观察到的全局奖励建模为其自身贡献(即个人奖励信号)和随机马尔可夫过程(即由其他智能体或外部因素引起的奖励信号)的总和,如果智能体在时间t处于状态i,并且它接收到全局奖励为g_t,那么它可以表示为:

[0087] $g_t=r(i)_t+b_t$

[0088] 其中,g_t表示t时刻的全局奖励,$r(i)_t$表示t时刻智能体在状态i下获得的真实奖励,b_t表示t时刻的噪声,代表不可观测的环境状态对全局奖励的影响,

[0089] 而且:

[0090] $b_{t+1}=b_t+z_t, z_t\sim N(\mu,\sigma_\omega^2)$

[0091] 其中,z_t为服从均值为μ,方差为σ_ω^2的高斯随机变量;

[0092] 基于上述,系统模型可以表示为:

[0093] $$\begin{cases} x_t = Ax_{t-1}+\omega_t, & \omega_t\sim N(0,\sum_1) \\ g_t = Cx_t+v_t, & v_t\sim N(0,\sum_2) \end{cases}$$

[0094] 其中,x_t表示t时刻的状态向量;

[0095] $$x_t=\begin{pmatrix} r(1)_t \\ r(2)_t \\ \cdot \\ \cdot \\ \cdot \\ r(s)_t \\ b_t \end{pmatrix}_{(s+1)\times 1}$$

[0096] 其中,s表示状态数量;

[0097] ω_t表示服从0均值高斯分布的系统噪声且协方差矩阵Σ_1如下:

[0098] $$\sum_1=\begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_\omega^2 \end{pmatrix}_{(s+1)\times(s+1)}$$

[0099] 其中,σ_ω^2表示系统噪声的协方差;

[0100] v_t表示观测误差,属于高斯白噪声,假设无观测误差,$\Sigma_2=0$;A表示状态转移矩阵

为,A=I,C表示观测矩阵,$C=(0\ \cdots\ 0\ 1_i\ 0\ \cdots\ 0\ 1)_{1\times(s+1)}$,若处于状态i,则C的第i个元素和最后一个元素值为1,其余为0。

[0101] 卡尔曼滤波是一种利用线性系统状态方程,通过输入观测数据,对系统状态进行最优估计的算法。卡尔曼滤波的优势在于它只需要部分数据,而非全部的历史状态和观测值。而且,能够在每次更新期间计算得到足够的统计数据。所以在训练过程中,可以在线运行该算法,且速度不会随着时间的推移而恶化。

[0102] 标准卡尔曼滤波算法基于具有零均值高斯白噪声的系统模型。为了使用标准卡尔曼滤波算法,将上面的系统模型转换为如下表示:

[0103] $$\begin{cases} x_t = Ax_{t-1} + Bu + \varepsilon_t\ ,\ \ \varepsilon_t \sim N(0,\ \Sigma_1) \\ g_t = Cx_t + v_t, \qquad\qquad v_t \sim N(0,\ \Sigma_2) \end{cases}$$

[0104] 其中,$B=\begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}_{(s+1)\times(s+1)}$,$u=\begin{pmatrix} 0 \\ 0 \\ \cdots \\ 0 \\ \mu \end{pmatrix}$,$\varepsilon_t$为零均值高斯白噪声,$\Sigma_2=0$;

[0105] 进一步地,在利用卡尔曼滤波算法从观测到的多智能体系统中智能体在与环境交互过程中接收的全局奖励中获得当前智能体的局部奖励之前,必须猜测由环境的不平稳性引起的噪声的协方差σ^2的值。然而,在一个真正的多智能体系统中,获取这个协方差值是不实际的。因为在上面给出的系统模型中噪声不为零,故本实施例采用如下的噪声在线估计方法,具体实现过程如下:

[0106] 步骤1:初始化$\mu_0=0$,$\sigma_{w0}^2=0.1$,t=0。

[0107] 步骤2:使用连续的μ_0,σ_{w0}^2运行卡尔曼滤波,迭代n次(n>200),记录$x_{t+1}(s+1)$,$x_{t+2}(s+1)$,…,$x_{t+n}(s+1)$。

[0108] 步骤3:开始循环

[0109] ①估计噪声的均值和方差:

[0110] $$\mu_t = \frac{1}{n-1}\sum_{i=2}^{n}(x_{t+i}(s+1) - x_{t+i-1}(s+1))$$

[0111] $$\sigma_{wt}^2 = \frac{1}{n-1}\sum_{i=2}^{n}(x_{t+i}(s+1) - x_{t+i-1}(s+1) - \mu_t)^2$$

[0112] ②使用μ_t和σ_{wt}^2运行基于滤波的信用分配算法,记录$x_{t+n+1}(s+1)$的值。

[0113] ③t←t+1

[0114] 结束循环。

[0115] 其中,μ_t表示t时刻噪声的均值,σ_{wt}^2表示t时刻噪声的方差。

[0116] 通过上述过程,即可根据估计的σ^2的值,基于转换后的系统模型,利用卡尔曼滤波算法从观测到的全局奖励中获得单一智能体对应的局部奖励。

[0117] 进一步地,本实施例所采用的强化学习算法为MADDPG算法,本实施例通过在MADDPG算法中引入基于卡尔曼滤波的信用分配实现智能体训练,其框架如图2所示。在MADDPG框架的基础上引入滤波,即将智能体接收到的全局奖励信号滤除不稳定环境因素影

响后得到智能体局部奖励信号,并将得到的奖励值用于对应的智能体训练,从而降低环境不平稳对智能体学习过程的影响。

[0118] 结合上述状态模型和步骤,本实施例的智能体最优策略获取方法的执行过程如图3所示,其包括如下步骤:

[0119] 步骤1:初始化强化学习环境、智能体网络参数和卡尔曼滤波参数。

[0120] 步骤2:循环迭代

[0121] ①在当前的状态i下,基于强化学习算法选择动作a并执行,到达新的状态k,得到全局奖励g_t。

[0122] ②对状态进行预测,更新估计值$\widehat{x_t}$和协方差矩阵$\widehat{P_t}$;

[0123] ③使用得到的全局奖励g_t更新状态后验估计值,校正模型。

[0124] ④根据$x_t(i)$获得局部奖励r(i),将探索得到的$<s_i,a_i,r_i,s'_i>$存入经验回放池。

[0125] ⑤重新估计噪声过程的均值μ和方差σ^2,更新u和Σ_1。

[0126] ⑥对于每一个智能体a,a∈(1,N)从Replay buffer中随机采样数据用于更新对应智能体的actor网络和critic网络。

[0127] ⑦t←t+1,i←k

[0128] 停止迭代。

[0129] 综上,本实施例提出了部分可观测环境下的多智能体系统全局奖励信号的设定方法并建立了全局奖励与个人贡献和环境噪声的关系;使用卡尔曼滤波方法从带噪声的全局奖励中获取智能体真实的训练回报,使其学习到不同场景下的最优策略。实现了部分可观测环境下多智能体执行协作任务过程中的信用分配;适用于部分可观测环境下,降低不平稳环境的影响,使多智能体高效学习在不同环境下完成任务的最优策略;解决了不平稳环境下的多智能体系统信用分配问题,可提高多智能体系统在完成协作任务过程中计算的高效性和鲁棒性。

[0130] 第二实施例

[0131] 本实施例提供了一种智能体最优策略获取装置,该装置包括以下模块:

[0132] 系统模型构建模块,用于将智能体在与环境交互过程中接收的全局奖励建模为单一智能体真实的局部奖励和不可观测环境状态引起的随机噪声信号之和,构建系统模型;

[0133] 卡尔曼滤波模块,用于基于所述系统模型构建模块所构建的系统模型,利用卡尔曼滤波算法从观测到的多智能体系统中智能体在与环境交互过程中接收的全局奖励中获得当前智能体对应的局部奖励;

[0134] 强化学习训练模块,用于根据所述卡尔曼滤波模块计算出的各智能体所对应的局部奖励,通过预设的强化学习算法,对多智能系统中的各智能体进行训练,使得每一智能体分别学习到当前环境下的最优策略。

[0135] 本实施例的智能体最优策略获取装置与上述第一实施例的智能体最优策略获取方法相对应;其中,本装置中的各功能模块所实现的功能与上述第一实施例的智能体最优策略获取方法中的各流程步骤一一对应;故,在此不再赘述。

[0136] 第三实施例

[0137] 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。

[0138] 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。

[0139] 第四实施例

[0140] 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。

[0141] 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。

[0142] 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。

[0143] 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。

[0144] 还需要说明的是,在本文中,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。

[0145] 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。

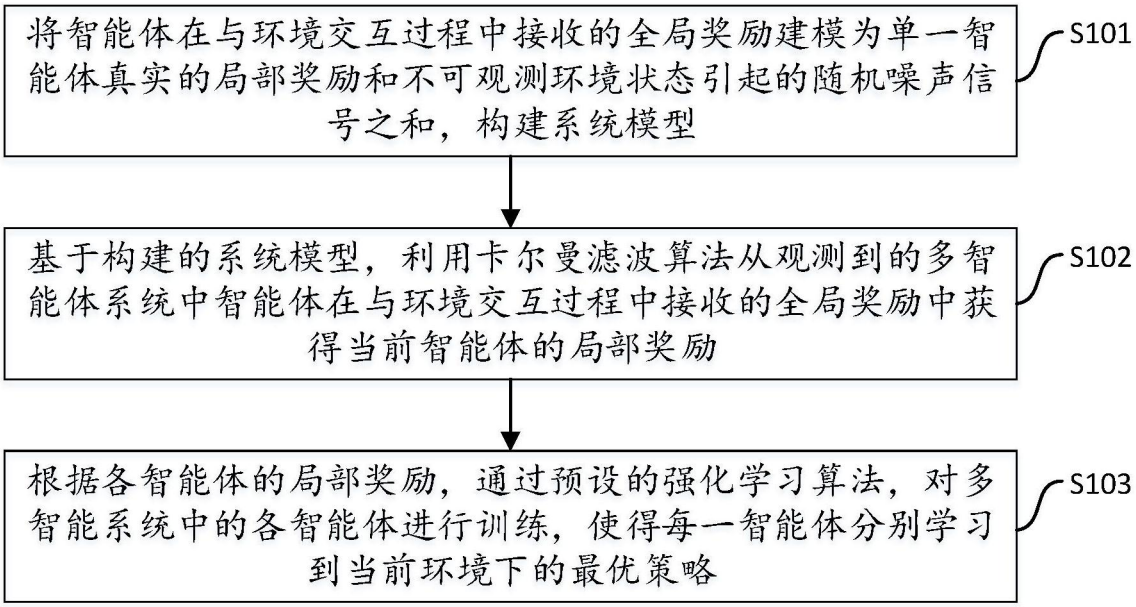

将智能体在与环境交互过程中接收的全局奖励建模为单一智能体真实的局部奖励和不可观测环境状态引起的随机噪声信号之和，构建系统模型
S101
基于构建的系统模型，利用卡尔曼滤波算法从观测到的多智能体系统中智能体在与环境交互过程中接收的全局奖励中获得当前智能体的局部奖励
S102
根据各智能体的局部奖励，通过预设的强化学习算法，对多智能系统中的各智能体进行训练，使得每一智能体分别学习到当前环境下的最优策略
S103

图1

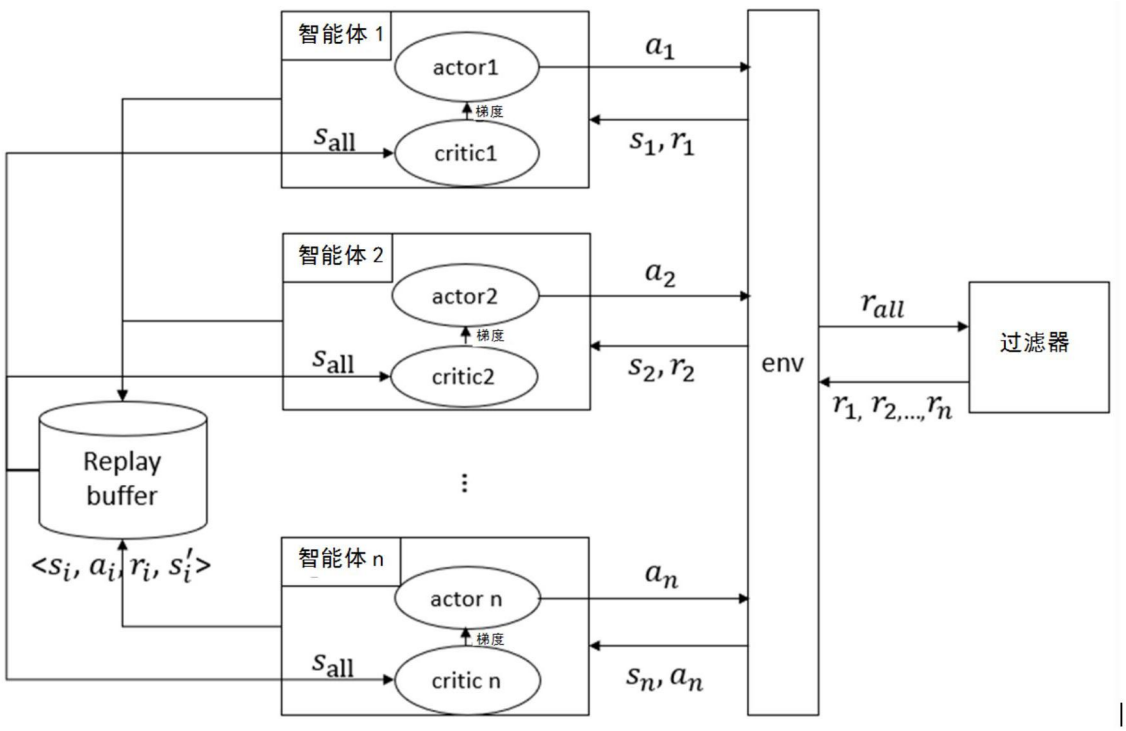

智能体 1
actor1
梯度
critic1
s_{all}
a_1
s_1, r_1
智能体 2
actor2
梯度
critic2
s_{all}
a_2
s_2, r_2
智能体 n
actor n
梯度
critic n
s_{all}
a_n
s_n, a_n
env
r_{all}
过滤器
$r_1, r_2, \ldots, r_n$
Replay buffer
$<s_i, a_i, r_i, s_i'>$

图2

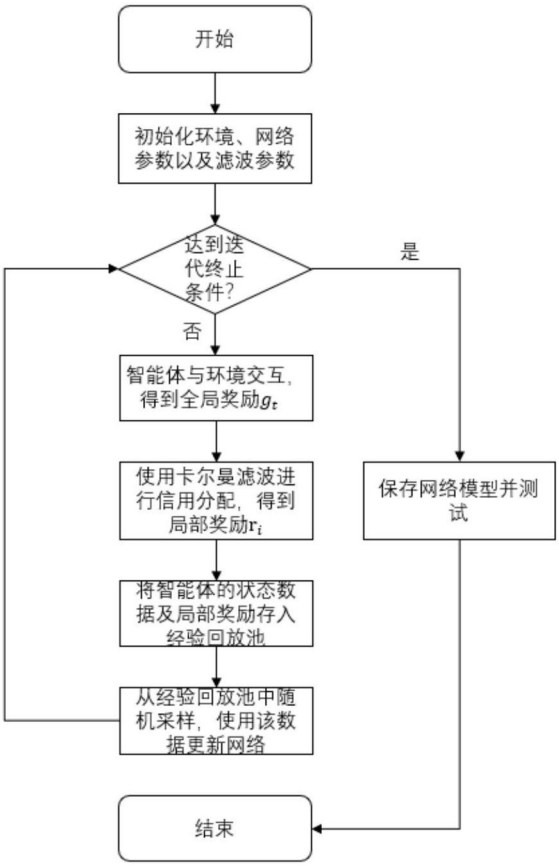

开始
初始化环境、网络参数以及滤波参数
达到迭代终止条件?
是
否
智能体与环境交互，得到全局奖励g_t
使用卡尔曼滤波进行信用分配，得到局部奖励r_i
保存网络模型并测试
将智能体的状态数据及局部奖励存入经验回放池
从经验回放池中随机采样，使用该数据更新网络
结束

图3