(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 117579358 B
(45)授权公告日 2024.09.06

(21)申请号 202311586285.7
(22)申请日 2023.11.24
(65)同一申请的已公布的文献号
申请公布号 CN 117579358 A
(43)申请公布日 2024.02.20
(73)专利权人 中国科学院自动化研究所
地址 100190 北京市海淀区中关村东路95号
(72)发明人 张俊格 乔丹 陈皓
(74)专利代理机构 北京华夏泰和知识产权代理有限公司 11662
专利代理师 邓菊香
(51)Int.Cl.
H04L 9/40(2022.01)
G06N 3/092(2023.01)

(56)对比文件
CN 112801731 A,2021.05.14
CN 113592101 A,2021.11.02

审查员 曾康玲

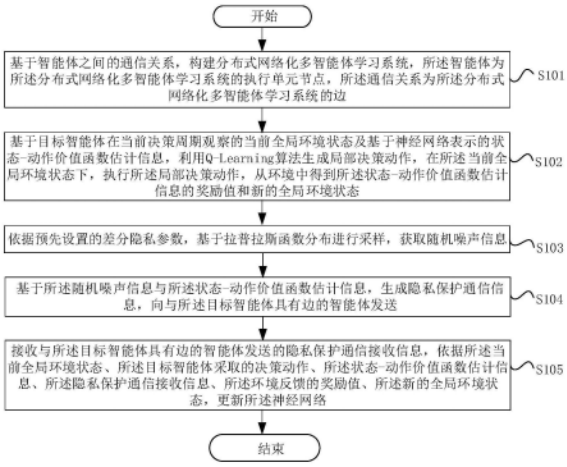
权利要求书2页 说明书13页 附图3页

(54)发明名称

多智能体通信方法、装置、存储介质和电子设备

(57)摘要

本发明涉及一种多智能体通信方法、装置、存储介质和电子设备，包括：基于智能体之间的通信关系，构建分布式网络化多智能体学习系统，智能体作为任务执行节点，通信关系描述为边；智能体基于观察到的当前全局环境状态及自身神经网络，执行局部决策动作，获取状态-动作的奖励值及更新的全局环境状态；基于拉普拉斯函数分布进行采样，获取随机噪声信息；将价值函数估计信息与随机噪声信息结合，生成隐私保护通信信息，与智能体的邻居智能体建立双向通信信道通信；依据当前状态-动作价值函数估计信息、接收的隐私保护通信接收信息、环境反馈的奖励值及新的全局环境状态，对神经网络进行迭代更新，具有严格理论保证的通信安全性能提升。



CN 117579358 B

1.一种多智能体通信方法,其特征在于,包括:

基于智能体之间的通信关系,构建分布式网络化多智能体学习系统,所述智能体为所述分布式网络化多智能体学习系统的执行单元节点,所述通信关系为所述分布式网络化多智能体学习系统的边;

基于目标智能体在当前决策周期观察的当前全局环境状态及基于神经网络表示的状态-动作价值函数估计信息,利用Q-Learning算法生成局部决策动作,在所述当前全局环境状态下,执行所述局部决策动作,从环境中得到所述状态-动作价值函数估计信息的奖励值和新的全局环境状态;

依据预先设置的差分隐私参数,基于拉普拉斯函数分布进行采样,获取随机噪声信息;

基于所述随机噪声信息与所述状态-动作价值函数估计信息,生成隐私保护通信信息,向与所述目标智能体具有边的智能体发送;

接收与所述目标智能体具有边的智能体发送的隐私保护通信信息,依据所述当前全局环境状态、所述目标智能体采取的决策动作、所述状态-动作价值函数估计信息、所述隐私保护通信信息、所述奖励值、所述新的全局环境状态,更新所述神经网络。

2.根据权利要求1所述的多智能体通信方法,其特征在于,所述依据预先设置的差分隐私参数,基于拉普拉斯函数分布进行采样,获取随机噪声信息,包括:

计算采样系数、第一差分隐私参数值及第二差分隐私参数值的乘积;

以零为所述拉普拉斯函数分布的位置参数,以所述乘积为所述拉普拉斯函数分布的尺度参数进行采样,获取所述随机噪声信息。

3.根据权利要求1所述的多智能体通信方法,其特征在于,所述基于所述随机噪声信息与所述状态-动作价值函数估计信息,生成隐私保护通信信息,包括:

将所述随机噪声信息与所述状态-动作价值函数估计信息相加,生成所述隐私保护通信信息。

4.根据权利要求1至3任一项所述的多智能体通信方法,其特征在于,所述方法还包括:

在下一决策周期,所述目标智能体基于所述新的全局环境状态及更新的神经网络,执行所述生成局部决策动作的步骤。

5.一种多智能体通信装置,其特征在于,所述多智能体通信装置包括:

系统构建模块,用于基于智能体之间的通信关系,构建分布式网络化多智能体学习系统,所述智能体为所述分布式网络化多智能体学习系统的执行单元节点,所述通信关系为所述分布式网络化多智能体学习系统的边;

状态动作模块,用于基于目标智能体在当前决策周期观察的当前全局环境状态及基于神经网络表示的状态-动作价值函数估计信息,利用Q-Learning算法生成局部决策动作,在所述当前全局环境状态下,执行所述局部决策动作,从环境中得到所述状态-动作价值函数估计信息的奖励值和新的全局环境状态;

噪声获取模块,用于依据预先设置的差分隐私参数,基于拉普拉斯函数分布进行采样,获取随机噪声信息;

隐私保护模块,用于基于所述随机噪声信息与所述状态-动作价值函数估计信息,生成隐私保护通信信息,向与所述目标智能体具有边的智能体发送;

策略更新模块,用于接收与所述目标智能体具有边的智能体发送的隐私保护通信信

息,依据所述目标智能体采取的决策动作、所述状态-动作价值函数估计信息、所述隐私保护通信信息、所述奖励值、所述新的全局环境状态及所述当前全局环境状态,更新所述神经网络。

6.根据权利要求5所述的多智能体通信装置,其特征在于,所述噪声获取模块包括:

计算单元,用于计算采样系数、第一差分隐私参数值及第二差分隐私参数值的乘积;

采样单元,用于以零为所述拉普拉斯函数分布的位置参数,以所述乘积为所述拉普拉斯函数分布的尺度参数进行采样,获取所述随机噪声信息。

7.根据权利要求5所述的多智能体通信装置,其特征在于,所述隐私保护模块包括:

隐私信息生成单元,用于将所述随机噪声信息与所述状态-动作价值函数估计信息相加,生成所述隐私保护通信信息;

隐私信息发送单元,用于将所述隐私保护通信信息向与所述目标智能体具有边的智能体发送。

8.根据权利要求5至7任一项所述的多智能体通信装置,其特征在于,所述状态动作模块还用于:

在下一决策周期,所述目标智能体基于所述新的全局环境状态及更新的神经网络,执行所述生成局部决策动作的步骤。

9.一种存储介质,其特征在于,存储介质上存储程序或指令,程序或指令被处理器运行时实现如权利要求1至4中任一项所述的多智能体通信方法的步骤。

10.一种电子设备,包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序,其特征在于,所述处理器执行所述程序时实现权利要求1至4中任一项所述的多智能体通信方法的步骤。

多智能体通信方法、装置、存储介质和电子设备

技术领域

[0001] 本发明涉及通信技术领域,尤其涉及一种多智能体通信方法、装置、存储介质和电子设备。

背景技术

[0002] 近年来,随着通信技术和人工智能的高速发展,许多现实生活中的系统都可以被建模成多智能体系统(MAS,Multi-agent System),如传感器网络、网联自动驾驶车辆、智能电网、无人仓储系统等。为了提升智能体的自主决策和协同能力,多智能体强化学习(MARL,Multi-agent Reinforcement Learning)为这些场景提供了有效框架和训练范式。

[0003] 为了解决多智能体通信、决策和学习引起的非平稳问题,基于MARL的框架主要采用集中训练和分散执行(CTDE,Centralized Training with Decentralized Execution)算法。CTDE算法通常假设在训练过程中存在一个强大的中心,收集每个智能体的所有局部观测和个体动作,基于环境状态、联合动作对应的奖励函数学习出所有智能体的联合最优策略并分配给每个智能体。在执行阶段,每个智能体仅根据自身局部观察做出决策。受益于CTDE算法的集中式信息架构发展出了一系列代表性方法,主要有信用分配、通信学习、策略分解等,智能体可以充分利用集中训练阶段的交互信息来更好地了解环境和其他智能体的行为,极大地缓解了训练非平稳问题,有助于智能体做出更有利于团队的行动,代表性算法包括:价值分解网络(VDN,Value Decomposition Networks)、QMIX、多智能体深度确定性策略梯度(MADDPG,Multi Agent Deep Deterministic Policy Gradient)、可微分交互学习(DIAL,Differentiable Inter Agent Learning)、BiCNet等。

[0004] 然而,CTDE算法无法处理呈指数增长的状态-动作空间,即中心控制器中的维数诅咒。此外,在训练过程中,中心控制器和智能体之间海量的信息交换也给通信带来了巨大的压力,而集中式架构的方式也增加了单点故障的系统性风险。因此,放松CTDE算法限制的另一种方法是利用网络化系统的分布式结构,开发去中心化训练去中心化执行(DTDE,Decentralized Training and Decentralized Execution)算法的MARL。在DTDE算法的训练过程中,智能体的可用信息仅限于通信范围内的局部邻居智能体,而不是CTDE算法中的所有智能体,从而可以避免潜在的信息泄露和其他智能体信息的过拟合。在执行过程中,邻居信息的使用促使智能体更多关注彼此之间的策略协调,而不是仅仅根据自身的局部观察做出决策。

[0005] 在DTDE算法中,利用通信网络扩散局部信息,可以提高MARL的可部署性、灵活性、系统鲁棒性和弹性,但也面临着一个独特的问题,即网络通信下邻居信息的可靠性。相关技术中,大多数DTDE算法框架的MARL都假设团队中的通信信道和成员足够安全和可信,而忽略了网络攻击和恶意行为对多智能体强化学习系统安全性的灾难性破坏,使得多智能体强化学习方法中,智能体之间的通信安全性不高。

发明内容

[0006] 有鉴于此,本发明提供一种多智能体通信方法、装置、存储介质和电子设备。

[0007] 具体地,本发明是通过如下技术方案实现的:

[0008] 根据本发明的第一方面,提供一种多智能体通信方法,多智能体通信方法包括:

[0009] 基于智能体之间的通信关系,构建分布式网络化多智能体学习系统,所述智能体为所述分布式网络化多智能体学习系统的执行单元节点,所述通信关系为所述分布式网络化多智能体学习系统的边;

[0010] 基于目标智能体在当前决策周期观察的当前全局环境状态及基于神经网络表示的状态-动作价值函数估计信息,利用Q-Learning算法生成局部决策动作,在所述当前全局环境状态下,执行所述局部决策动作,从环境中得到所述状态-动作价值函数估计信息的奖励值和新的全局环境状态;

[0011] 依据预先设置的差分隐私参数,基于拉普拉斯函数分布进行采样,获取随机噪声信息;

[0012] 基于所述随机噪声信息与所述状态-动作价值函数估计信息,生成隐私保护通信信息,向与所述目标智能体具有边的智能体发送;

[0013] 接收与所述目标智能体具有边的智能体发送的隐私保护通信接收信息,依据所述目标智能体采取的决策动作、所述状态-动作价值函数估计信息、所述隐私保护通信接收信息、所述奖励值、所述新的全局环境状态及所述当前全局环境状态,更新所述神经网络。

[0014] 本技术方案中的多智能体通信方法,通过基于智能体之间的通信关系,以智能体为节点,通信关系为边构建分布式网络化多智能体学习系统;基于目标智能体观察的当前全局环境状态及神经网络,获取局部决策动作,执行局部决策动作,从环境中得到所述状态-动作价值函数估计信息的奖励值和新的全局环境状态;依据预先设置的差分隐私参数,基于拉普拉斯函数分布进行采样,获取随机噪声信息;基于随机噪声信息与状态-动作价值函数估计信息,生成隐私保护通信信息,向与目标智能体具有边的智能体发送;基于状态-动作价值函数估计信息、隐私保护通信接收信息、奖励值、新的全局环境状态及当前全局环境状态,更新神经网络。这样,利用采样得到的随机噪声,对状态-动作价值函数估计信息相加,从而保护用于通信的状态-动作价值函数估计信息无法被还原,提升了通信的安全性;同时,通过神经网络更新,在接收信息被干扰的情况下,仍然能保证加噪的状态-动作价值函数估计信息的收敛性与隐私保护性,实现多智能体之间的高质量策略协同与合作通信。

[0015] 根据本发明的第二方面,提供一种多智能体通信装置,多智能体通信装置包括:

[0016] 系统构建模块,用于基于智能体之间的通信关系,构建分布式网络化多智能体学习系统,所述智能体为所述分布式网络化多智能体学习系统的执行单元节点,所述通信关系为所述分布式网络化多智能体学习系统的边;

[0017] 状态动作模块,用于基于目标智能体在当前决策周期观察的当前全局环境状态及基于神经网络表示的状态-动作价值函数估计信息,利用Q-Learning算法生成局部决策动作,在所述当前全局环境状态下,执行所述局部决策动作,从环境中得到所述状态-动作价值函数估计信息的奖励值和新的全局环境状态;

[0018] 噪声获取模块,用于依据预先设置的差分隐私参数,基于拉普拉斯函数分布进行采样,获取随机噪声信息;

[0019] 隐私保护模块,用于基于所述随机噪声信息与所述状态-动作价值函数估计信息,生成隐私保护通信信息,向与所述目标智能体具有边的智能体发送;

[0020] 策略更新模块,用于接收与所述目标智能体具有边的智能体发送的隐私保护通信接收信息,依据所述目标智能体采取的决策动作、所述状态-动作价值函数估计信息、所述隐私保护通信接收信息、所述奖励值、所述新的全局环境状态及所述当前全局环境状态,更新所述神经网络。

[0021] 根据本发明的第三方面,提供一种存储介质,其上存储有计算机程序,程序被处理器执行时实现第一方面的任意可能的实现方式中的多智能体通信方法的步骤。

[0022] 根据本发明的第四方面,提供一种电子设备,包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序,处理器执行程序时实现第一方面的任意可能的实现方式中的多智能体通信方法的步骤。

附图说明

[0023] 此处的附图被并入说明书中并构成本说明书的一部分,示出了符合本发明的实施例,并与说明书一起用于解释本发明的原理。

[0024] 为了更清楚地说明本发明实施例或现有技术中的技术方案,下面将对实施例或相关技术描述中所需要使用的附图作简单地介绍,显而易见地,对于本领域普通技术人员而言,在不付出创造性劳动性的前提下,还可以根据这些附图获得其他的附图。

[0025] 图1为本发明实施例提供的一种多智能体通信方法的流程示意图;

[0026] 图2为本发明实施例提供的一种多智能体通信方法中分布式网络化多智能体学习系统示意图;

[0027] 图3为本发明实施例提供的一种多智能体通信方法中单车道减速跟车场景下的性能示意图;

[0028] 图4为本发明实施例提供的一种多智能体通信方法中单车道加速跟车场景下的性能示意图;

[0029] 图5为本发明实施例提供的一种多智能体通信处理装置示意图;

[0030] 图6为本发明实施例提供的一种电子设备的结构示意图。

具体实施方式

[0031] 为使本发明实施例的目的、技术方案和优点更加清楚,下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例是本发明的一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有做出创造性劳动的前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0032] 相关技术中,大多数DTDE算法框架的MARL都假设团队中的通信信道和成员足够安全和可信,而忽略了网络攻击和恶意行为对多智能体强化学习系统安全性的灾难性破坏,使得多智能体强化学习方法中,智能体之间的通信安全性不高。

[0033] 本实施例中,提出了一种基于差分隐私(DP,Differential Privacy)保护的多智能体通信方法,可以应用在网络化多智能体强化学习方法或中网络化多智能体强化学习系统。其中,差分隐私来自网络安全和机器学习领域,通过向通信信道中的通信信息添加不相

关随机噪声信息,在不影响多智能体强化学习系统运行的前提下,来保证通信信息在被第三方恶意窃听后,第三方仍然无法将窃听的信息还原成真实的通信信息,从而提高网络化多智能体强化学习系统的安全性和用户隐私性。因而,本实施例利用差分隐私保护机制,通过设计随时间衰减的加性噪声,从时变的Laplace分布中,采样随机噪声,将采样的随机噪声与源通信信息相加,从而保护源通信信息无法被还原。同时,本实施例中,还通过设计相应的智能体策略更新机制,在接收信息被干扰的情况下,仍然能保证加噪的通信信息的收敛性与隐私保护性,实现多智能体之间的高质量策略协同与合作。

[0034] 参见图1,本发明实施例提供了一种网络化系统多智能体分布式强化学习通信方法,该方法可以包括如下步骤:

[0035] S101、基于智能体之间的通信关系,构建分布式网络化多智能体学习系统,所述智能体为所述分布式网络化多智能体学习系统的执行单元节点,所述通信关系为所述分布式网络化多智能体学习系统的边;

[0036] 本实施例中,针对某一应用场景,为该应用场景内的N个智能体分别构建状态动作价值函数,将智能体之间的通信拓扑设定为随机切换且联合联通的无向图G＝{V,E,A},其中,V代表节点,即智能体,E代表节点之间通信的边,A代表邻接矩阵。

[0037] 本实施例中,可通信节点之间的邻接矩阵元素为1,不可通信节点之间的邻接矩阵元素为0。例如,对于某一智能体(节点),与该智能体具有边的智能体,对应邻接矩阵中的邻接矩阵元素为1,与该智能体不具有边的智能体,对应邻接矩阵中的邻接矩阵元素为0。这样,可以缓解集中式训练分布式执行算法的维度诅咒和通信压力。

[0038] 本实施例中,作为一可选实施例,基于网络化多智能体马尔科夫(Markov)决策过程,构建去中心化训练去中心化执行的分布式网络化多智能体学习系统。

[0039] S102、基于目标智能体在当前决策周期观察的当前全局环境状态及基于神经网络表示的状态-动作价值函数估计信息,利用Q-Learning算法生成局部决策动作,在所述当前全局环境状态下,执行所述局部决策动作,从环境中得到所述状态-动作价值函数估计信息的奖励值和新的全局环境状态;

[0040] 本实施例中,执行所述局部决策动作,得到环境反馈的奖励函数和更新的全局环境状态,维护本地的状态-动作价值函数估计,并将“状态-动作”对的状态动作价值函数信息Q(s,a)作为真实信息,即获取环境反馈的奖励函数和更新后的全局环境状态,并维护本地的状态-动作价值函数估计信息。

[0041] 本实施例中,在初始通信时,每个智能体可观察到全局环境状态S,并根据全局环境状态和自身的神经网络生成独立局部决策动作a_i,在环境中执行独立局部决策动作a_i,维护本地的状态-动作价值函数估计信息,以便发送至其他智能体以进行相互通信和协同合作。以网联协同自动驾驶汽车场景为例,每一个智能体是一辆具有车对车(Vehicle to Vehicle,V2V)通信功能的自动驾驶汽车,多个智能体共同行使在道路上并可以与其他智能体进行实时通信,所有车辆的联合驾驶行为会影响该区域的全局环境状态,全局环境状态包括但不限于:各车道的汽车位置、前后车辆的车距、各车辆的速度和加速度、各车道的拥堵情况、各车道的信号灯状态、行驶路线等,局部决策动作包括但不限于:车距控制系数、车速增益系数(包括加速、减速、匀速)等。关于进行独立局部决策动作,具体可参见相关技术文献并根据实际场景需要进行设定,在此略去详述。

[0042] S103、依据预先设置的差分隐私参数,基于拉普拉斯函数分布进行采样,获取随机噪声信息;

[0043] 本实施例中,基于拉普拉斯函数分布和隐私保护机制进行采样,获取随机噪声信息。设定差分隐私保护机制,以对通信信息进行隐私保护。每个智能体从时变的拉普拉斯函数分布中进行采样,得到随机噪声信息。

[0044] 本实施例中,为保护通信信息,采用随时间衰减的加性拉普拉斯噪声机制,通过获取随时间衰减的加性拉普拉斯噪声信息(随机噪声信息),可以对真实的通信信息进行隐私保护。在每次通信时,利用下式,从如下的拉普拉斯(Laplace)分布中采样得到随机噪声信息$\eta_i(t)$:

[0045] $$\eta_i(t) \sim Lap(0, \iota_i(t))$$

[0046] $$\iota_i(t) = s_{gain} s_i q_i^t$$

[0047] 其中,$\iota_i(t)$为噪声分布的方差参数,用于决定Laplace噪声的分布;s_i和q_i为差分隐私参数,分别决定噪声初始分布和衰减速率;s_{gain}为增益系数,用于调节噪声大小的尺度。各参数是可配置的正常数,需要满足如下条件:s_i、q_i、$s_{gain} \in (0,1)$;s_i、$q_i > 0$,$s_{gain} \geq 0$。

[0048] S104、基于所述随机噪声信息与所述状态-动作价值函数估计信息,生成隐私保护通信信息,向与所述目标智能体具有边的智能体发送;

[0049] 本实施例中,将随机噪声信息添加到真实的通信信息中以构建隐私通信信息,即隐私保护通信信息,以与通信范围内的邻居智能体进行隐私通信信息的交换。

[0050] 本实施例中,作为一可选实施例,利用下式生成隐私保护通信信息:

[0051] $$\widehat{Q}_{s,a}^{i}(t) = Q_{s,a}^{i}(t) + \eta_i(t)$$

[0052] 式中,$Q_{s,a}^{i}(t)$为状态-动作价值函数估计信息,$\widehat{Q}_{s,a}^{i}(t)$为隐私保护通信信息。

[0053] 本实施例中,目标智能体与当前时刻通信范围内的邻居智能体以广播形式建立双向通信信道,交换各自进行隐私保护后的信息,即目标智能体通过设置通信范围,该通信范围内的智能体均与目标智能体具有通信关系,即各智能体与目标智能体均通过边相连接,通信信息仅限于通信范围内的局部邻居智能体,从而在一定程度上,避免过时信息的干扰以及潜在的其他智能体信息的过拟合。

[0054] 图2为本发明实施例提供的一种多智能体通信方法中分布式网络化多智能体学习系统示意图,在该分布式网络化多智能体学习系统中,包括8个节点(智能体),分别为A1-A8,节点与节点之间通过边进行连接,对于目标智能体A1,与目标智能体A1具有边连接的智能体包括:A2、A3、A4,则智能体A2、A3、A4所围成的区域(虚线内的区域)为目标智能体A1的通信范围。

[0055] S105、接收与所述目标智能体具有边的智能体发送的隐私保护通信接收信息,依据所述当前全局环境状态、所述目标智能体采取的决策动作、所述状态-动作价值函数估计信息、所述隐私保护通信接收信息、所述环境反馈的奖励值、所述新的全局环境状态,更新所述神经网络。

[0056] 本实施例中,环境反馈的奖励值为在当前全局环境状态,执行局部决策动作能够获得的奖励。隐私保护通信接收信息为与目标智能体具有边的智能体发送的隐私保护通信信息。作为一可选实施例,在每次获取状态-动作价值函数估计信息后,对智能体的神经网

络算法进行参数更新,以进行策略更新,以在下一决策周期,利用更新的神经网络决策局部决策动作。这样,通过依据当前状态-动作价值函数估计、接收的隐私保护通信信息、环境反馈的奖励值及新的全局环境状态,对状态-动作价值函数估计进行迭代更新,最终训练得到稳定的分布式协同决策策略,并具有严格理论保证的通信安全性能提升。

[0057] 本实施例中,在QD-Learning算法的基础上,对通信信息,即状态-动作价值函数估计信息进行随机噪声处理,将原有的通信信息替换为隐私保护通信信息,提高了网络化多智能体强化系统的安全性。

[0058] 本实施例中,依据通信信息、隐私保护通信接收信息、奖励值、新的全局环境状态、当前全局环境状态,对神经网络进行更新,从而通过设计与隐私保护机制相适应的智能体更新策略,在获取局部信息和被噪音扰动的隐私保护通信信息情况下,实现智能体的神经网络更新和状态-动作价值函数估计信息的渐近收敛,并有效保证了算法(p,r)-准确度和数据的∈-差分隐私度。作为一可选实施例,利用下式更新神经网络:

$$Q_{s,a}^{i}(t+1)=Q_{s,a}^{i}(t)-\beta_{s,a}(t)\sum_{j\in N_i(t)}\left(Q_{s,a}^{j}(t)-\widetilde{Q}_{s,a}^{j}(t)\right)$$

[0059]

$$+\alpha_{s,a}(t)\left(r_i\left(s_t,a_t\right)+\gamma\max_{a'\in\mathcal{A}}Q_{s',a'}^{i}(t)-Q_{s,a}^{i}(t)\right)$$

[0060] 其中,参数满足如下条件:

[0061] $$\alpha_{s,a}(k)=\begin{cases}\dfrac{a}{(k+1)^{\tau_2}} & \text{如果 } t=T_{s,a}(k)\\ 0 & \text{其他}\end{cases}$$

[0062] $$\beta_{s,a}(k)=\begin{cases}\dfrac{b}{(k+1)^{\tau_1}} & \text{如果 } t=T_{s,a}(k)\\ 0 & \text{其他}\end{cases}$$

[0063] $$T_{s,a}(k)=\inf\{t\geq 0|\sum_{s=0}^{t}\mathbb{I}_{(s_t,a_t)=(s,a)}=k+1\}$$

[0064] 其中a、b、$\tau_1\in(0.5,1]$,$\tau_2\in\left(0,\tau_1-\frac{1}{2+\epsilon_1}\right)$,$\epsilon_1>0$为正常数,$T_{s,a}(k)$代表在整个随机事件序列中,对于某一状态-动作对(s,a)的第k+1次采样发生的时间,r_i为依据奖励函数计算执行局部决策动作得到的奖励值,S′为新的全局环境状态,S为当前全局环境状态,$\widehat{Q}_{s,a}^{j}(t)$为与目标智能体具有边的第j个智能体发送的隐私保护通信接收信息,$Q_{s,a}^{i}(t+1)$为下一决策周期对应的状态-动作价值函数估计信息,用于目标智能体在下一决策周期向与目标智能体具有边的各智能体发送,$Q_{s,a}^{i}(t)$为目标智能体当前发送的状态-动作价值函数估计信息,s_t为当前全局环境状态,a_t为局部决策动作,A为依据新的全局环境状态得到的局部决策动作集合,a'为依据新的全局环境状态获取的局部决策动作,$Q_{s',a'}^{i}(t)$为依据新的全局环境状态及依据新的全局环境状态获取的局部决策动作得到的状态-动作价值函数估计

信息，γ为遗忘因子，为常系数。

[0065] 本实施例中，神经网络是预先训练得到的，在训练中，以一致性损失和时序差分损失相加作为最终的优化目标，使用去中心化训练去中心化执行(DTDE)框架进行端到端训练以优化一致性损失和时序差分损失相加的和值。作为一可选实施例，目标智能体通过与通信范围内的邻居智能体进行隐私通信信息的交换，进而更新自身的神经网络(策略神经网络)。

[0066] 本实施例中，在下一决策周期，目标智能体基于新的全局环境状态及更新的神经网络，执行获取局部决策动作的步骤，如此循环，直至执行完所有决策周期。

[0067] 本实施例的方法，可以在真实通信信息被噪声扰动的情况下，保证每个智能体的状态-动作价值函数估计信息(通信信息)满足均方一致性和期望一致性，下面进行详细描述。

[0068] 本实施例中，对于每个状态-动作价值函数估计信息$Q^{i}_{s,a}(t)$，在通信拓扑满足连通性的情况下，可以达到以均方渐近一致性：

[0069] $$\lim_{n\to\infty} E\left[\left(Q^{i}_{s,a}(t)-Q^{j}_{s,a}(t)\right)^{2}\right]=0, i,j=1,\ldots,N$$

[0070] 以及，期望渐近一致性：

[0071] $$\lim_{n\to\infty} E\left[Q^{i}_{s,a}(t)-\frac{1}{N}\sum_{i=1}^{N}Q^{i}_{s,a}(t)\right]=0, i=1,\ldots,N$$

[0072] 其中，$\sum_{i=1}^{N}Q^{i}_{s,a}(t)$为与目标智能体具有边的所有智能体的状态动作价值函数的和值，本实施例的方法，由于智能体的状态动作价值函数满足均方一致性和期望一致性，可以保证系统内所有智能体学习到近似一致的状态动作价值函数，进而保证了分布式训练分布式执行框架下的协调能力。

[0073] 同时，本实施例的方法，可以实现准确率和召回率(p,r)-准确度的隐私保护性能，以及∈-差分隐私度。

[0074] 本实施例中，对于每个状态-动作价值函数估计信息，与目标智能体具有边连接的所有智能体的状态动作价值函数的平均值$\overline{Q}_{s,a}(t)=\frac{1}{N}\sum_{i=1}^{N}Q^{i}_{s,a}(t)$相对于最优状态-动作价值函数估计信息$Q^{*}_{s,a}$的误差满足(p,r)-准确度：

[0075] $$P\left(\left|\overline{Q}_{s,a}(t)-Q^{*}_{s,a}\right|\le r\right)\ge 1-\frac{2\,\mathrm{var}(\Delta(t))}{r^{2}}$$

[0076] $$r=\frac{\sqrt{2\,\mathrm{var}(\Delta(t))}}{\sqrt{p}}$$

[0077] 其中，p为精确率，r为召回率，随机变量Δ(t)的方差满足有界性：

[0078] $$\mathrm{var}(\Delta)\le W_0 s_i^2 q_i^{2t-2}\frac{1-\left(\frac{M_t}{q_i^2}\right)^{t}}{1-\frac{M_t}{q_i^2}}$$

[0079] 其中，$M_t=(1-\alpha_{s,a}(t)+\gamma\alpha_{s,a}(t))^2\in(0,1)$，$W_0=\frac{\beta_{s,a}(0)^2}{N^2}\lambda_N(D)$，取决于差分隐私噪音的参数设置和网络拓扑的度矩阵$D=\mathrm{diag}\{d_1,d_2,\ldots,d_N\},d_i=\sum_{j=1}^{N}a_{ij}$。

[0080] 本实施例中，对于真实通信信息的被保护程度，遵循差分隐私机制的定义，可以用∈-差分隐私度进行衡量。具体来说，考虑两个满足δ-相邻的数据集D和D′，该对数据集仅有一个数据点不同且误差值在δ以内，在该对数据集输入到预先设置的随机算法M:D→O后，随机算法的输出满足如下概率关系：

[0081] $P(M(D)\in O)\leqslant\exp(\in)P(M(D')\in O)$

[0082] 本实施例中，算法隐私度满足：$\epsilon=\max\{\frac{|\Delta\eta_i(t)|}{s_iq_i^t}\}$，$\Delta\eta_i(t)$为从相同的拉普拉斯函数分布$\mathrm{Lap}(0,\iota_i(t))$中独立采样两次的随机噪声$\eta_i^1(t)$和$\eta_i^2(t)$之差。

[0083] 本实施例的方法，对网络化多智能体强化学习系统的通信信息进行有效的隐私保护，提高了多智能体强化学习系统的隐私性和安全性。

[0084] 本实施例中，以交通仿真环境(SUMO,Simulation of Urban Mobility)为例，通过构建基于分布式网络化多智能体学习系统的网联多车自动驾驶场景(Cooperative Adaptive Platoon Control,CAPC)，其中，作为一可选实施例，网联多车自动驾驶场景包含一个单车道车辆队列的加速跟车场景和减速跟车场景。每一车辆对应一智能体，每个智能体在训练和执行阶段，只能接收到通信范围内有限邻居智能体的通信信息，例如，对于其中一智能体，只能与该智能体相邻的智能体(前车和后车)进行通信，该智能体与相邻的智能体围成的区域为该智能体的通信范围。接下来，初始化每个智能体的差分隐私参数s_i和q_i，本实施例中，初始化$s_i=0.01$和$q_i=0.99$，根据差分隐私保护机制采样，得到加性噪声$\eta_i(t)$，将加性噪声与真实通信信息(状态-动作价值函数估计信息)相加得到隐私保护通信信息，将隐私保护通信信息传递给通信范围内的邻居。

[0085] 智能体从当前策略网络(神经网络)采样动作(执行局部决策动作)并与环境交互，得到环境的新的全局环境状态和奖励值，接收邻居智能体发送的隐私保护通信信息，根据智能体策略更新方式更新当前策略网络。重复初始化步骤和环境交互步骤，直到策略网络收敛，训练过程结束。

[0086] 为说明本实施例的有效性，本实施例构建了两种常见的协同自动驾驶场景：单车道加速跟车场景(CAPC Catch-up)和单车道减速跟车场景(CAPC slow-down)，本实施例场景均采用SUMO交通仿真器构建。对于智能体(车辆)参数设置，考虑车辆模型采用仿真软件内置最优速度模型(Optimal Velocity Model,OVM)，车辆的状态集合包括距前车距离h_i，当前速度v_i，当前加速度a_i，纵向控制动作集合为前车车距增益和速度增益共同构成的组合(α_i^o,β_i^o)，此处考虑四种最优等级{(0,0) (0,0.5) (0.5,0) (0.5,0.5)}，控制周期为0.1s，总控制时长为60s，奖励(代价)函数设置为$r_{i,t}=(h_{i,t}-h_t^*)^2+(v_{i,t}-v_t^*)^2+0.1a_{i,t}^2$。CAPC的详细场景参数设置如下表0。具体而言，在CAPC加速跟车场景中，随机初始化所有车辆的行驶速度和前车车距，除领头车辆外，满足所有车辆的速度小于最优行驶速度$v_t^*=15m/s$且前车车距大于最优保持车距$h_t^*=20m$，理想目标是所有跟车车辆都能学会提速并缩短车距的协同策略；在CAPC减速跟车场景中，随机初始化所有车辆的行驶速度和前车车距，除领头

车辆外,满足所有车辆的速度大于最优行驶速度$v_t^* = 15m/s$且前车车距略小于最优保持车距$h_t^* = 20m$,理想目标是所有跟车车辆都能学会降低车速并保持避免碰撞的协同策略,由于碰撞的可能性存在,该场景下的决策策略更为复杂。

[0087] 表0:SUMO仿真软件中CPAC场景详细参数设置

[0088]

实验场景参数名	参数值
距前车安全距离	$h_l \geqslant 1m$
安全驾驶速度	$v_i \leqslant 30m/s$
安全加速度	$\lvert a_i \rvert \leqslant 2.5m/s^2$
OVM中的停车前车距离	$h_{stoo} = 5m$
OVM中的全速前车距离	$h_{full} = 35m$
碰撞惩罚(前车距离小于1m)	1000
发生碰撞的额外惩罚代价	$5(2h_{stop}-h_{i,t})^2$

[0089] 设置了三种不同程度的差分隐私增益系数s_{gain}用以对照性能,每组实验均统计训练1,000,000步后得到的累计奖励值(Rewards)作为性能指标,结果如表1所示。

[0090] 表1:CAPC场景下不同程度噪声保护的多智能体学习算法得到的奖励值

[0091]

Noise Level	Rewards	
	CAPC Catch-up	CAPC Slow-down
Non-DP MARL $s_{gain} = 0$	-519.68 ± 58.20	-1585.82 ± 190.37
DP-MARL $s_{gain} = 0.01$	-187.77 ± 64.82	-1450.16 ± 386.11
DP-MARL $s_{gain} = 0.1$	-1296.81 ± 347.86	-2875.79 ± 291.53
DP-MARL $s_{gain} = 1$	None	None

[0092] 图3为本发明实施例提供的一种多智能体通信方法中单车道减速跟车场景下的性能示意图,图4为本发明实施例提供的一种多智能体通信方法中单车道加速跟车场景下的性能示意图,图3和图4中,s_{gain}分别等于0、0.1、0.01,总步数(Steps)为1,000,000。通过本实施例的方法,能够显著提高算法在多智能体合作任务上的性能表现和安全性。

[0093] 本实施例提出的基于差分隐私保护的网络化多智能体强化学习的多智能体通信方法,通过构建去中心化的分布式网络化多智能体学习系统,使得每个智能体在训练和执行的过程中,只能与通信范围内的邻居智能体交换局部信息,降低了中心化系统的高额训练资源需求和单点故障系统性风险,同时,通过向通信信息中添加拉普拉斯加性噪声,从而保护真实通信信息无法被第三方恶意节点窃取和还原,得到安全性、隐私性、鲁棒性更好的多智能体协作策略,能够在DTDE学习框架下保证通信信道中的通信信息无法被第三方恶意节点窃听后还原出真实信息,从而有效提高多智能体合作强化通信的学习效率、可靠性和安全性。具有如下明显优点:

[0094] 1)本实施例通过构建DTDE算法学习框架,相比主流CTDE算法学习框架,具有更低的系统风险,更好的扩展性和灵活性,以及更高的学习效率和智能体之间的协作能力。

[0095] 2)本实施例的方法,能够显著提高多智能体合作算法通信信息的隐私性和安全性,避免通信信息被第三方恶意节点窃听并还原用户真实信息的风险,在数据安全性上超过了目前主流的无隐私保护功能的多智能体合作算法。

[0096] 基于同一发明构思,如图5所示,本发明实施例还提供了一种多智能体通信装置,装置包括:

[0097] 系统构建模块501,用于基于智能体之间的通信关系,构建分布式网络化多智能体学习系统,所述智能体为所述分布式网络化多智能体学习系统的执行单元节点,所述通信关系为所述分布式网络化多智能体学习系统的边;

[0098] 本实施例中,作为一可选实施例,基于网络化多智能体马尔科夫(Markov)决策过程,构建去中心化训练去中心化执行的分布式网络化多智能体学习系统。

[0099] 状态动作模块502,用于基于目标智能体在当前决策周期观察的当前全局环境状态及基于神经网络表示的状态-动作价值函数估计信息,利用Q-Learning算法生成局部决策动作,在所述当前全局环境状态下,执行所述局部决策动作,从环境中得到所述状态-动作价值函数估计信息的奖励值和新的全局环境状态;

[0100] 本实施例中,每个智能体根据当前全局环境状态进行独立的局部决策动作,并收到自身的局部奖励,并将“状态-动作”的状态-动作价值函数估计信息作为真实通信信息。

[0101] 本实施例中,作为一可选实施例,状态动作模块502还用于:

[0102] 在下一决策周期,所述目标智能体基于所述新的全局环境状态及更新的神经网络,执行所述获取局部决策动作的步骤。

[0103] 噪声获取模块503,用于依据预先设置的差分隐私参数,基于拉普拉斯函数分布进行采样,获取随机噪声信息;

[0104] 本实施例中,作为一可选实施例,噪声获取模块503包括:

[0105] 计算单元,用于计算采样系数、第一差分隐私参数值及第二差分隐私参数值的乘积;

[0106] 采样单元,用于以零为所述拉普拉斯函数分布的位置参数,以所述乘积为所述拉普拉斯函数分布的尺度参数进行采样,获取所述随机噪声信息。

[0107] 本实施例中,作为一可选实施例,利用下式获取随机噪声信息:

[0108] $\eta_i(t) \sim \mathrm{Lap}(0, \iota_i(t))$

[0109] $\iota_i(t) = s_{gain} s_i q_i^t$

[0110] 隐私保护模块504,用于基于所述随机噪声信息与所述状态-动作价值函数估计信息,生成隐私保护通信信息,向与所述目标智能体具有边的智能体发送;

[0111] 本实施例中,将随机噪声信息添加到真实的通信信息中以构建隐私通信信息,以与通信范围内的邻居智能体进行隐私通信信息的交换。

[0112] 本实施例中,作为一可选实施例,隐私保护模块504包括:

[0113] 隐私信息生成单元,用于将所述随机噪声信息与所述状态-动作价值函数估计信息相加,生成所述隐私保护通信信息;

[0114] 隐私信息发送单元,用于将所述隐私保护通信信息向与所述目标智能体具有边的智能体发送。

[0115] 本实施例中,作为一可选实施例,利用下式生成隐私保护通信信息:

[0116] $\widehat{Q}^i_{s,a}(t) = Q^i_{s,a}(t) + \eta_i(t)$

[0117] 策略更新模块505,用于接收与所述目标智能体具有边的智能体发送的隐私保护

通信接收信息,依据所述目标智能体采取的决策动作、所述状态-动作价值函数估计信息、所述隐私保护通信接收信息、所述奖励值、所述新的全局环境状态及所述当前全局环境状态,更新所述神经网络。

[0118] 本实施例中,作为一可选实施例,利用下式更新神经网络:

[0119]
$$Q^{i}_{s,a}(t+1) = Q^{i}_{s,a}(t) - \beta_{s,a}(t) \sum_{j \in N_i(t)} \left(Q^{i}_{s,a}(t) - \widehat{Q}^{j}_{s,a}(t)\right) + \alpha_{s,a}(t)\left(r_i(s_t, a_t) + \gamma \max_{a' \in \mathcal{A}} Q^{i}_{s',a'}(t) - Q^{i}_{s,a}(t)\right)$$

[0120] 其中,参数满足如下条件:

[0121]
$$\alpha_{s,a}(k) = \begin{cases} \dfrac{a}{(k+1)^{\tau_2}} & \text{如果 } t = T_{s,a}(k) \\ 0 & \text{其他} \end{cases}$$

[0122]
$$\beta_{s,a}(k) = \begin{cases} \dfrac{b}{(k+1)^{\tau_1}} & \text{如果 } t = T_{s,a}(k) \\ 0 & \text{其他} \end{cases}$$

[0123] 基于同一发明构思,本发明实施例还提供了一种存储介质,其上存储有计算机程序,程序被处理器执行时实现上述任意可能的实现方式中的多智能体通信方法的步骤。

[0124] 可选地,存储介质可以是非临时性计算机可读存储介质,例如,非临时性计算机可读存储介质可以是ROM、随机存取存储器(RAM)、CD-ROM、磁带、软盘和光数据存储设备等。

[0125] 基于同一发明构思,参见图6,本发明实施例还提供了一种电子设备,包括存储器101(例如非易失性存储器)、处理器102及存储在存储器101上并可在处理器102上运行的计算机程序,处理器102执行程序时实现上述任意可能的实现方式中的多智能体通信方法的步骤,可相当于如前的多智能体通信装置,当然,该处理器还可以用来处理其他数据或运算。该电子设备可以是PC、服务器、终端等设备。

[0126] 如图6所示,该电子设备一般还可以包括:内存103、网络接口104、以及内部总线105。除了这些部件外,还可以包括其他硬件,对此不再赘述。

[0127] 需要指出的是,上述多智能体通信装置可以通过软件实现,其作为一个逻辑意义上的装置,是通过其所在的电子设备的处理器102将非易失性存储器中存储的计算机程序指令读取到内存103中运行形成的。

[0128] 本说明书中描述的主题及功能操作的实施例可以在以下中实现:数字电子电路、有形体现的计算机软件或固件、包括本说明书中公开的结构及其结构性等同物的计算机硬件、或者它们中的一个或多个的组合。本说明书中描述的主题的实施例可以实现为一个或多个计算机程序,即编码在有形非暂时性程序载体上以被数据处理装置执行或控制数据处理装置的操作的计算机程序指令中的一个或多个模块。可替代地或附加地,程序指令可以被编码在人工生成的传播信号上,例如机器生成的电、光或电磁信号,该信号被生成以将信息编码并传输到合适的接收机装置以由数据处理装置执行。计算机存储介质可以是机器可读存储设备、机器可读存储基板、随机或串行存取存储器设备、或它们中的一个或多个的组合。

[0129] 本说明书中描述的处理及逻辑流程可以由执行一个或多个计算机程序的一个或多个可编程计算机执行,以通过根据输入数据进行操作并生成输出来执行相应的功能。处理及逻辑流程还可以由专用逻辑电路—例如FPGA(现场可编程门阵列)或ASIC(专用集成电路)来执行,并且装置也可以实现为专用逻辑电路。

[0130] 适合用于执行计算机程序的计算机包括,例如通用和/或专用微处理器,或任何其他类型的中央处理单元。通常,中央处理单元将从只读存储器和/或随机存取存储器接收指令和数据。计算机的基本组件包括用于实施或执行指令的中央处理单元以及用于存储指令和数据的一个或多个存储器设备。通常,计算机还将包括用于存储数据的一个或多个大容量存储设备,例如磁盘、磁光盘或光盘等,或者计算机将可操作地与此大容量存储设备耦接以从其接收数据或向其传送数据,抑或两种情况兼而有之。然而,计算机不是必须具有这样的设备。此外,计算机可以嵌入在另一设备中,例如移动电话、个人数字助理(PDA)、移动音频或视频播放器、游戏操纵台、全球定位系统(GPS)接收机、或例如通用串行总线(USB)闪存驱动器的便携式存储设备,仅举几例。

[0131] 适合于存储计算机程序指令和数据的计算机可读介质包括所有形式的非易失性存储器、媒介和存储器设备,例如包括半导体存储器设备(例如EPROM、EEPROM和闪存设备)、磁盘(例如内部硬盘或可移动盘)、磁光盘以及CD-ROM和DVD-ROM盘。处理器和存储器可由专用逻辑电路补充或并入专用逻辑电路中。

[0132] 虽然本说明书包含许多具体实施细节,但是这些不应被解释为限制任何发明的范围或所要求保护的范围,而是主要用于描述特定发明的具体实施例的特征。本说明书内在多个实施例中描述的某些特征也可以在单个实施例中被组合实施。另一方面,在单个实施例中描述的各种特征也可以在多个实施例中分开实施或以任何合适的子组合来实施。此外,虽然特征可以如上在某些组合中起作用并且甚至最初如此要求保护,但是来自所要求保护的组合中的一个或多个特征在一些情况下可以从该组合中去除,并且所要求保护的组合可以指向子组合或子组合的变型。

[0133] 类似地,虽然在附图中以特定顺序描绘了操作,但是这不应被理解为要求这些操作以所示的特定顺序执行或顺次执行、或者要求所有例示的操作被执行,以实现期望的结果。在某些情况下,多任务和并行处理可能是有利的。此外,上述实施例中的各种系统模块和组件的分离不应被理解为在所有实施例中均需要这样的分离,并且应当理解,所描述的程序组件和系统通常可以一起集成在单个软件产品中,或者封装成多个软件产品。

[0134] 由此,主题的特定实施例已被描述。其他实施例在所附权利要求书的范围以内。在某些情况下,权利要求书中记载的动作可以以不同的顺序执行并且仍实现期望的结果。此外,附图中描绘的处理并非必需所示的特定顺序或顺次顺序,以实现期望的结果。在某些实现中,多任务和并行处理可能是有利的。

[0135] 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。而且,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

在包括要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0136] 以上仅是本发明的具体实施方式,使本领域技术人员能够理解或实现本发明。对这些实施例的多种修改对本领域的技术人员来说将是显而易见的,本文中所定义的一般原理可以在不脱离本发明的精神或范围的情况下,在其它实施例中实现。因此,本发明将不会被限制于本文所示的这些实施例,而是要符合与本文所申请的原理和新颖特点相一致的最宽的范围。

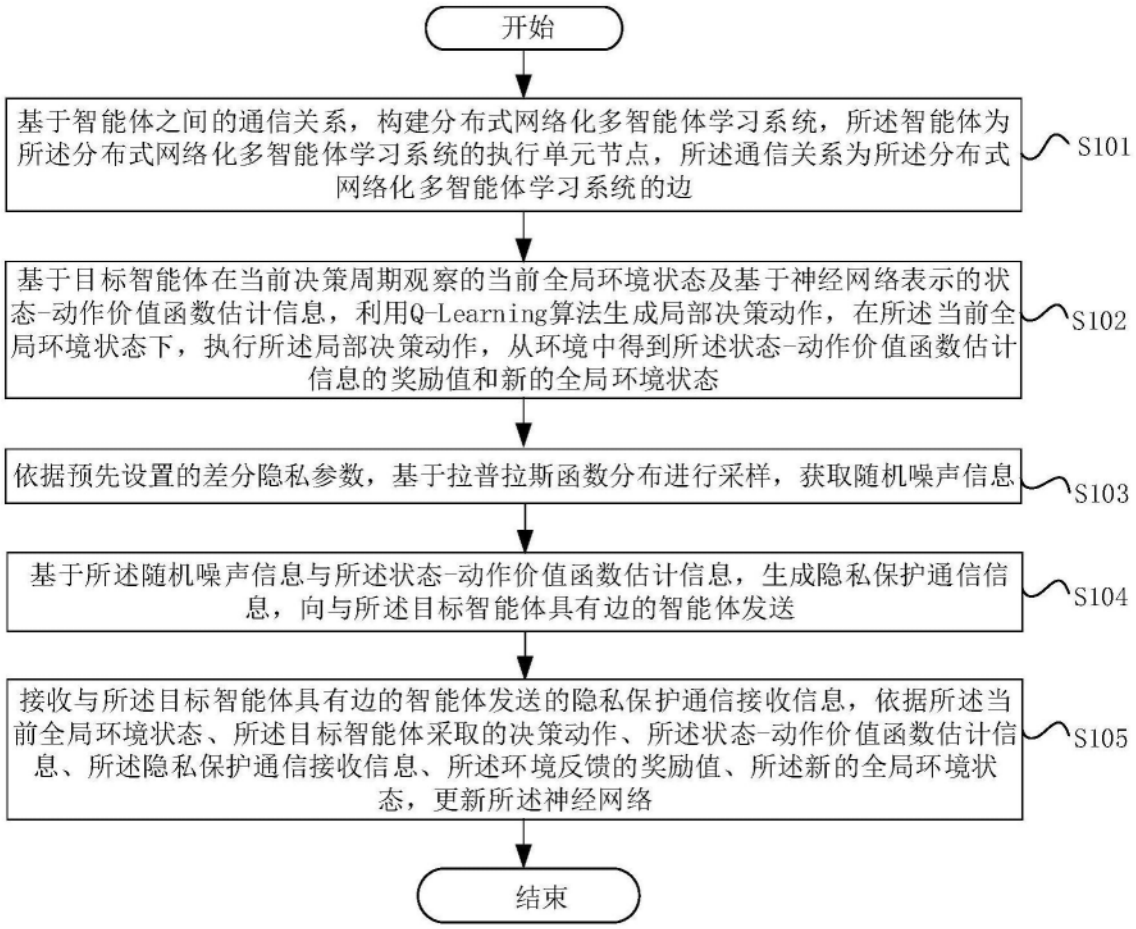

开始
基于智能体之间的通信关系，构建分布式网络化多智能体学习系统，所述智能体为所述分布式网络化多智能体学习系统的执行单元节点，所述通信关系为所述分布式网络化多智能体学习系统的边
S101
基于目标智能体在当前决策周期观察的当前全局环境状态及基于神经网络表示的状态-动作价值函数估计信息，利用Q-Learning算法生成局部决策动作，在所述当前全局环境状态下，执行所述局部决策动作，从环境中得到所述状态-动作价值函数估计信息的奖励值和新的全局环境状态
S102
依据预先设置的差分隐私参数，基于拉普拉斯函数分布进行采样，获取随机噪声信息
S103
基于所述随机噪声信息与所述状态-动作价值函数估计信息，生成隐私保护通信信息，向与所述目标智能体具有边的智能体发送
S104
接收与所述目标智能体具有边的智能体发送的隐私保护通信接收信息，依据所述当前全局环境状态、所述目标智能体采取的决策动作、所述状态-动作价值函数估计信息、所述隐私保护通信接收信息、所述环境反馈的奖励值、所述新的全局环境状态，更新所述神经网络
S105
结束

图1

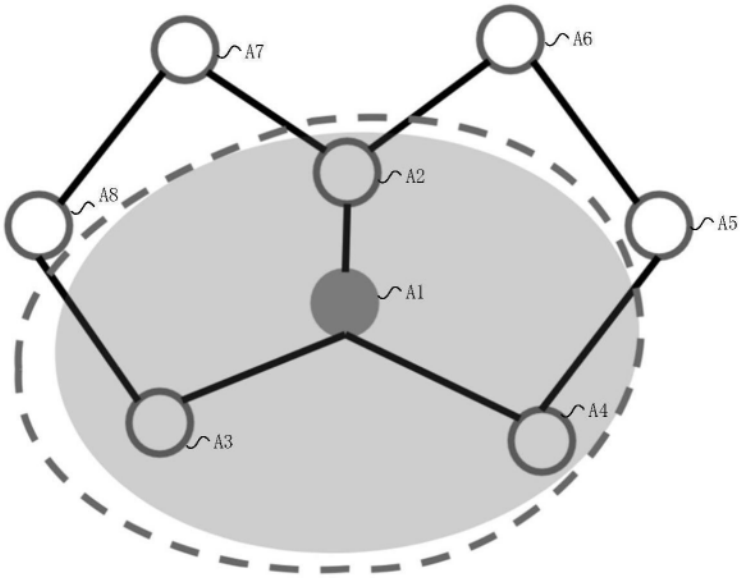

A7
A6
A8
A2
A5
A1
A4
A3

图2

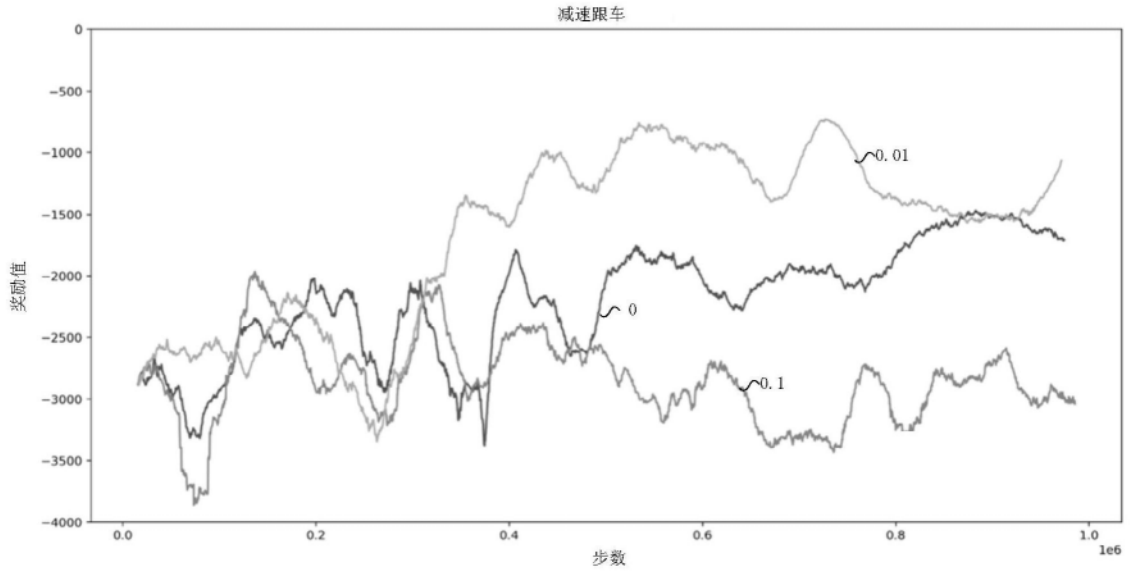


图3

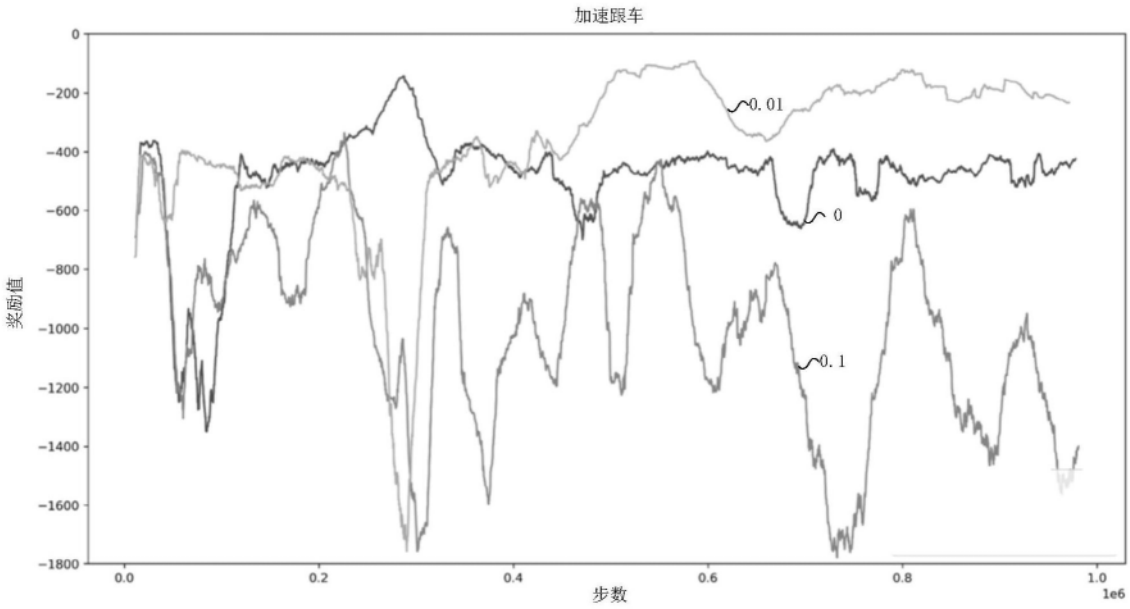


图4

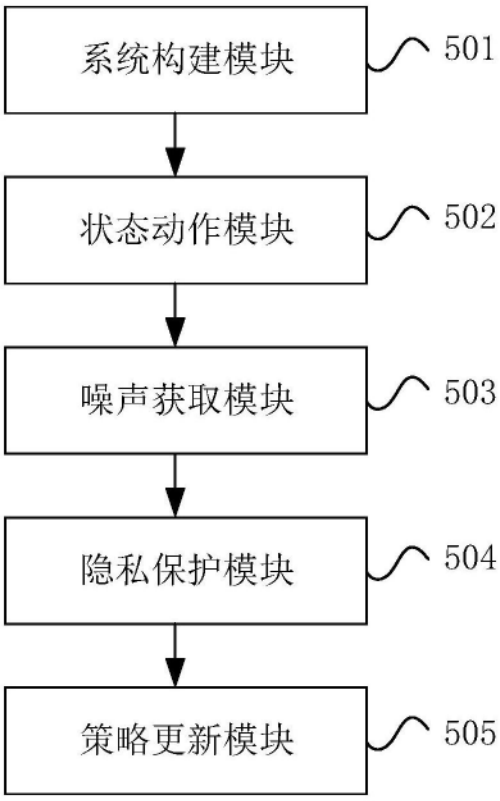

系统构建模块
501
状态动作模块
502
噪声获取模块
503
隐私保护模块
504
策略更新模块
505

图5

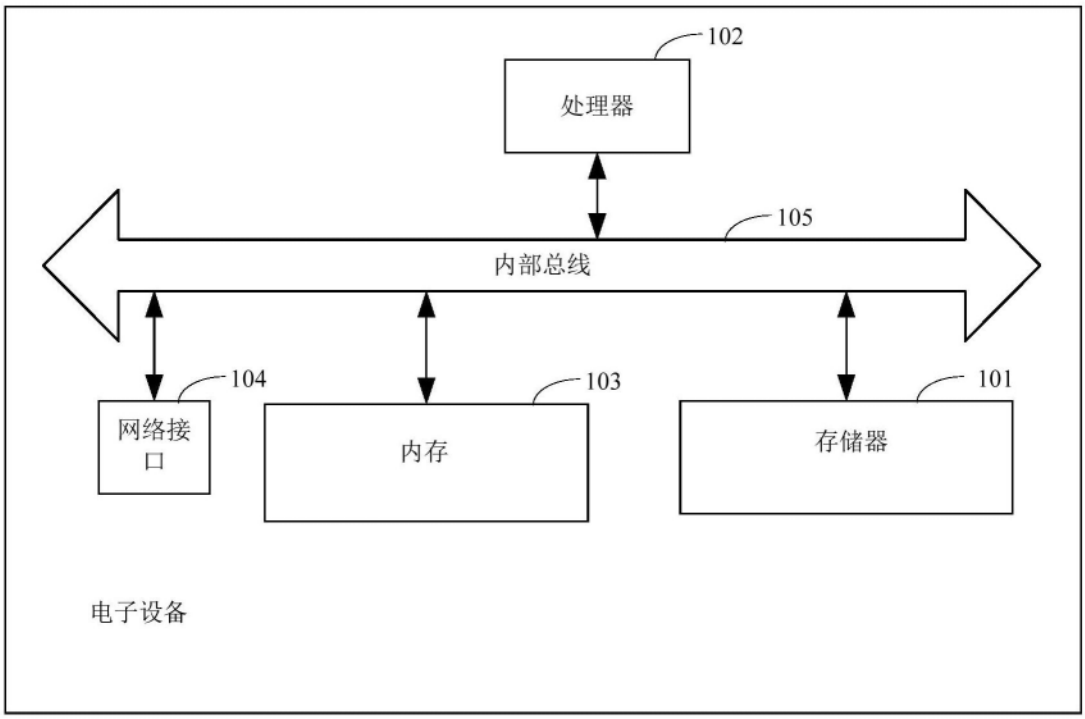

102
处理器
105
内部总线
104
网络接口
103
内存
101
存储器
电子设备

图6