(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 115273502 B
(45)授权公告日 2023.06.30

(21)申请号 202210903865.3
(22)申请日 2022.07.28
(65)同一申请的已公布的文献号
申请公布号 CN 115273502 A
(43)申请公布日 2022.11.01
(73)专利权人 西安电子科技大学
地址 710071 陕西省西安市太白南路2号
(72)发明人 李长乐 王硕 岳文伟 陈新洋 陈越 计星怡
(74)专利代理机构 陕西电子工业专利中心 61205
专利代理师 王品华
(51)Int.Cl.
G08G 1/081(2006.01)
G06F 30/27(2020.01)
G06N 3/04(2023.01)
G06N 3/08(2023.01)
(56)对比文件
CN 113487860 A,2021.10.08
WO 2022126940 A1,2022.06.23
戴福青;庞笔照;袁婕;赵元棣.基于超级网络的空铁联合交通流分布模型.武汉理工大学学报(交通科学与工程版).2017,(第05期),全文.
审查员 王晓红

权利要求书4页 说明书8页 附图3页

(54)发明名称

一种交通信号协同控制方法

(57)摘要

本发明提出了一种交通信号协同控制方法，主要解决现有集中式评价MARL方法在交通信号协同控制中各智能体具有相同信誉导致合作效率低的问题。其实现方案为:构建路网交通信号控制仿真环境,获取训练样本集;构建由Critic神经网和Actor神经网络并行排布的DRMA网络模型;设计该网络模型的目标优化函数,为各智能体分配不同的信誉并计算各自在协作中的差异贡献;用训练样本集和目标优化函数对DRMA网络模型进行迭代更新,获得训练好的DRMA模型;用训练好的网络模型从环境中获取交通信号协同控制方案。本发明提高了路网的交通信号协同控制效率,降低了路网的平均车辆行程延迟,可用于城市路网的自适应交通信号控制。

CN 115273502 B

1.一种交通信号协同控制方法,其特征在于,包括如下步骤:

(1)构建路网交通信号控制仿真环境:

构建由路口集合I＝{I_1,I_2,…,I_n,…,I_N}及与其对应智能体集合a＝{a_1,a_2,…,a_n,…,a_N}组成的交通信号协同控制场景,其中,N为路口集合中的路口总数,a_n表示对应N个智能体中的第n个智能体,I_n表示N个路口中的第n个路口,每个路口中均存在车道集合$L_n=\{L_n^1,L_n^2,\cdots,L_n^m,\cdots,L_N^M\}$,$L_n^m$表示路口$I_n$的M个车道中的第m个车道,M为车道集合中的车道总数,0≤n≤N,0≤m≤M,N≥2,M≥2;

(2)获取训练样本集Y:

每个智能体a_n采集T个时间步长的交通状态信息,每个时间步长的交通状态信息包括:各路口智能体在每个时刻t下的交通状态观察s_t^n、执行动作u_t^n、执行动作后得到的协作奖励$\hat{r}_t^n$,$0\le t\le T$,用T个时间步长的交通状态信息构成样本规模为N×T的训练样本集Y;

(3)构建DRMA网络模型H:

(3a)建立由7个全连接层依次级联组成的Critic神经网络;

(3b)建立由5个全连接层依次级联组成的Actor神经网络;

(3c)将Critic神经网络和Actor神经网络并行排布构成DRMA网络模型H;

(4)设计DRMA网络模型H的目标优化函数J:

根据路网中的信号灯智能体通过Actor网络π输出策略执行相应动作的机制,采用动作价值Q^π评估智能体在交通状态s_t下执行动作u_t的价值,设计DRMA网络模型H的如下目标优化函数,以使智能体的动作价值Q^π的期望达到最大:

$$\max_{\theta_\pi} J(\theta_\pi)=\max_{\theta_\pi} E\left[Q^\pi \mid \pi(\theta_\pi)\right]$$

其中,$J(\theta_\pi)$为Actor网络π输出策略条件下路网中智能体动作价值Q^π的期望值,θ_π为Actor网络π的神经网络参数;

(5)对DRMA网络模型H进行迭代训练:

(5a)初始化迭代次数为e,最大迭代次数为E,E≥2000,e＝1;

(5b)Critic网络通过训练集Y计算每个智能体的个体贡献A_t^n并反馈给Actor网络,以计算Actor网络的参数优化方向$\nabla_{\theta_\pi} J(\theta_\pi)$;

(5c)采用梯度上升法并行更新Actor网络参数θ_π和Critic网络参数θ_c,更新过程按照时间步长依次进行,每T个时间步长更新记作一次迭代训练,其中,T≥3000,为一次迭代训练的最大时间步长,执行一次迭代训练后,e＝e+1;

(5d)重复执行(5b)和(5c),直到满足e≥E,则训练结束,得到的训练好的DRMA网络模型为H^*;

(6)获取交通信号协同控制方案:

(6a)采集路网当前最新的交通状态信息,构建与(2)中训练样本Y结构相同的测试样本F;

(6b)将测试样本F输入至训练好的DRMA网络模型H^*,该模型中的Actor网络根据测试样本F输出每个时刻全局智能体的动作概率分布;

(6c)每个智能体根据最大概率原则输出各自最优的协作动作u_n^*，得到全局智能体的最优协作动作集合$\{u_1^*,u_2^*,\cdots,u_n^*,\cdots,u_N^*,\}$，该集合为该路网的交通信号协同控制方案。

2.根据权利要求1所述的方法，其特征在于，步骤(2)中用T个时间步长的交通状态信息构成样本规模为N×T的训练样本集Y，实现如下：

(2a)采集t时刻路口I_n第m条车道上的车辆数$v_t^{n,m}$，计算路口I_n所有车道上的车辆总数s_t^n：

$$s_t^n=\sum_{m=1}^{M}v_t^{n,m}$$

其中，M为每个路口具有的车道总数，s_t^n记作智能体a_n在时刻t下的交通状态观察；

(2b)采集智能体a_n在时刻t下的执行动作u_t^n，即该时刻交通灯的相位动作；

(2c)采集t时刻路口I_n处的车辆流出量out_t^n和流入量in_t^n，计算该时刻路口I_n处的车辆净流出量r_t^n，计算公式如下：

$$r_t^n=out_t^n-in_t^n$$

其中，r_t^n记作智能体a_n在t时刻执行动作后收到的奖励；

(2d)对(2c)中智能体a_n的奖励r_t^n进行空间加权，获得智能体a_n的协作奖励$\hat{r}_t^n$，其计算公式如下：

$$\hat{r}_t^n=r_t^n+\alpha\sum_{i\in K(n)}r_t^i$$

其中，K(n)表示智能体a_n的相邻智能体集合，α为空间加权因子，协作奖励$\hat{r}_t^n$使智能体能够考虑周围邻居的执行动作和奖励以加强彼此间的协作；

(2e)将t时刻在(2a)、(2b)、(2d)中得到的智能体a_n的交通状态观察s_t^n、执行动作u_t^n和协作奖励$\hat{r}_t^n$三者集合构成一个训练样本$y_{n,t}$：

$$y_{n,t}=\{s_t^n,u_t^n,\hat{r}_t^n\};$$

(2f)对N个智能体重复进行(2a)至(2e)，按照时间步长共进行T步，获得N×T个训练样本，构成训练样本集Y，形式如下：

$$Y=\begin{bmatrix} y_{1,1} & \cdots & y_{1,t} & \cdots & y_{1,T} \\ \vdots & \ddots & & & \vdots \\ y_{n,1} & & \ddots & & y_{n,T} \\ \vdots & & & \ddots & \vdots \\ y_{n,1} & \cdots & y_{n,t} & \cdots & y_{N,T} \end{bmatrix}_{N\times T}$$

其中，$y_{n,t}$表示智能体a_n在t时刻构建的一个训练样本。

3.根据权利要求1所述的方法，其特征在于，步骤(3a)中建立由7个全连接层依次级联组成的Critic神经网络，具体结构和参数如下：

该Critic神经网络中顺次级联的7个全连接层为：输入层→第一隐藏层→第二隐藏层

→第三隐藏层→第四隐藏层→第五隐藏层→线性输出层；

该Critic神经网络的输入数据是规模为N×M+N维的联合状态向量，五个隐藏层的神经元规模依次为380、250、160、80、20，每个隐藏层的输出均使用ReLU函数激活，输出层输出的数据为N维的联合价值向量。

4.根据权利要求1所述的方法，其特征在于，步骤(3b)中建立由5个全连接层依次级联组成的Actor神经网络，具体结构和参数如下：

该Actor神经网络中顺次级联的5个全连接层为：输入层→第一隐藏层→第二隐藏层→第三隐藏层→SoftMax输出层；

该Actor神经网络的输入数据是M维的局部交通状态观察向量，三个隐藏层的神经元规模依次为256、128、64，每个隐藏层的输出均使用ReLU函数激活，SoftMax层输出智能体执行动作的概率分布向量。

5.根据权利要求1所述的方法，其特征在于，步骤(5b)中所述的Critic网络通过训练集Y计算每个智能体的个体贡献A_t^n并反馈给Actor网络，以计算Actor网络的参数优化方向$\nabla_{\theta_\pi} J(\theta_\pi)$，公式表示如下：

$$\nabla_{\theta_\pi} J(\theta_\pi)=\mathrm{E}\left[\sum_{t=0}^{T}\nabla\theta_\pi \log p_{\theta_\pi}^n\left(u_t^n \mid s_t^n\right) A_t^n\left(s_t,u_t\right)\right]$$

其中，θ_π为Actor网络π的神经网络参数，∇_{θ_π}为对θ_π求梯度的运算，$p_{\theta_\pi}^n\left(u_t^n \mid s_t^n\right)$为t时刻智能体$a_n$在状态$s_t^n$的条件下使用Actor网络π执行动作$u_t^n$的概率；

$A_t^n\left(s_t,u_t\right)$表示智能体$a_n$在t时刻根据全局的交通状态观察$s_t$和全局的执行动作$u_t$计算得出的自身个体贡献，其计算为$A_t^n\left(s_t,u_t\right)=Q^\pi\left(s_t,u_t\right)-\sum_{u_t^n} p_{\theta_\pi}^n\left(u_t^n \mid s_t^n\right) Q^\pi\left(s_t,\left(u_t^{-n},u_t^n\right)\right)$，式中，$u_t=(u_t^n,u_t^{-n})$表示全局的执行动作$u_t$分为两个部分：自身执行动作$u_t^n$和其他智能体的执行动作集合$u_t^{-n}$，$Q^\pi(s_t,u_t)$表示对全局智能体在t时刻后执行动作的价值估计，$\sum_{u_t^n} p_{\theta_\pi}^n\left(u_t^n \mid s_t^n\right) Q^\pi\left(s_t,\left(u_t^{-n},u_t^n\right)\right)$表示对除智能体$a_n$自身外其他智能体在t时刻后执行动作的价值估计，两者之差即为t时刻智能体a_n在协同控制中的个体贡献A_t^n；

$Q^\pi(s_t,u_t)$表示Critic网络根据输入的全局状态s_t和联合动作u_t计算得出的全局动作价值，其计算为$Q^\pi\left(s_t,u_t\right)=E\left[\sum_{\xi=t}^{T}\gamma^{\xi-t}\hat{r}_t\left(s_\xi,u_\xi\right) \mid s_t,u_t\right]$，式中，$\hat{r}_t\left(s_t,u_t\right)$为全局智能体的空间加权协作奖励，γ为未来回报折扣因子。

6.如权利要求1所述的方法，其特征在于，步骤(5c)中对Actor网络参数θ_π和Critic网络参数θ_c进行更新，公式如下：

$$\theta'_\pi=\theta_\pi+\beta_\pi\nabla_{\theta_\pi} J(\theta_\pi)$$

$$\theta'_c=\theta_c+\beta_c\nabla_{\theta_c}\left(\delta_t\right)^2$$

其中，θ'_π为更新后的Actor网络参数，θ'_c为更新后的Critic网络参数，$\nabla_\theta J(\theta)$为智能体

a_n获得的未来折扣回报期望的梯度，作为Actor网络的优化步长，β_π为Actor网络的学习率；$\nabla_{\theta_c}(\delta_t)^2$为Critic网络的优化步长，$\beta_c$为Critic网络的学习率；$\delta_t$为一步时间误差，作为Critic网络进行优化的损失函数，根据空间加权奖励$\hat{r}_t$和动作价值Q^π计算得到：$\delta_t=\hat{r}_t+\gamma Q^\pi(s_t,u_t)-Q^\pi{}_{t-1}(s_t,u_t)$，式中，γ为未来回报折扣因子，$Q^\pi(s_t,u_t)$为Critic网络根据输入的全局状态$s_t$和联合动作$u_t$计算得出的全局动作价值。

一种交通信号协同控制方法

技术领域

[0001] 本发明属于多智能体强化学习技术领域,特别涉及一种交通信号协同控制方法,可用于城市路网的自适应交通信号控制。

背景技术

[0002] 目前我国大型城市交通拥堵问题日益严重,更新缓慢的道路基础设施以及无法适应交通流变化的固定相位交通信号灯使得城市路网中的交通流无法被有效疏导从而造成大面积的交通拥堵。针对这一问题,自适应交通信号控制ATSC技术被提出用于应对实时多变的交通流。传统的自适应交通信号控制方法通常是基于时间间隔或时间损耗的。基于时间损耗的控制方法根据路口驶来车辆的时间损失来控制交通信号的相位状态延长。基于时间间隔的方法选择在检测到连续的车流间有足够的时间间隔时切换交通信号相位。

[0003] 近年来,强化学习RL方法在ATSC领域兴起。与传统的基于时间损失或时间间隔的方法不同,RL采用参数化的网络模型,其输入来自真实的交通场景,输出是通过最大化奖励函数所得到的控制策略。经典的RL方法以Q-learning为代表,采用Q-table存储动作价值,但该方法在高维数据问题中的应用受到限制。针对这一问题,深度神经网络端到端的学习方式被应用于RL算法中,得到改进后的RL算法被称为深度强化学习DRL算法,其在众多复杂的高维数据任务中取得了突破性的表现。深度强化学习DRL可分为两种主要的方法:基于价值的和基于策略的。基于价值的方法,例如深度Q-learning,采用深度神经网络拟合状态价值函数并采用一步时间误差来更新网络参数。基于策略的方法,例如策略迭代和策略梯度,采用深度神经网络对状态价值函数进行参数化,并利用随机梯度下降的优化方法更新其参数。后来,一种AC方法被提出,该方法是基于价值和基于策略两种学习方法的结合体,通过使用Critic网络对每个Actor的动作价值进行评估,并引导他们优化自己的策略。AC方法在价值估计上的方差更小,且比基于策略的方法收敛更快,在交通信号控制方面优于Q-learning方法。

[0004] 申请公布号为CN112201060A的专利中提出了一种基于AC方法的单交叉口交通信号控制方法,其实现步骤为:获取固定时间间隔的路网车辆位置信息和速度信息,以及对应时刻的信号灯状态;对采集的训练数据进行预处理,获得车辆队列-信号灯状态的集合;利用车辆队列-信号灯状态集,更新Actor网络和Critic网络参数;根据最终的收敛模型,可以得到基于AC的单交叉口交通信号最优配时方案,即下一时刻的最优信号。与现有技术相比,该发明通过人工智能方法,获取了交通运行过程中所隐藏的重要交通信息,最终得到了比传统定时方法通行效率更高的配时方案。但该专利只研究了单交叉路口信号控制问题,无法实现多交叉路口的协同控制,不适用于城市路网。

[0005] 尽管DRL方法在交通信号控制中表现良好,但对于城市路网,训练所需的联合动作空间随所控制的交通信号灯数量呈指数级增长,极其高维的联合动作空间对于单一集中式的DRL方法在训练上难以达到收敛。在这种情况下,多智能体强化学习MARL方法被提出。该方法在早期采用分布式独立控制的DRL对城市路网中各路口的交通信号进行独立控制。但

由于各智能体之间没有通信，每个智能体只考虑最大化自己的回报，在同时与环境交互而不相互协作的情况下，这种早期分布式独立控制的MARL算法在收敛性上表现很差。为了获得更好的收敛性，MARL方法得到了改进，即在分布控制的基础上加入了集中评价机制，主要思想是利用集中式的Critic网络和分布式的Actor网络来控制路网中的交通信号，通过提高每个智能体的环境观测能力使智能体能够在控制策略中考虑彼此的动作，从而实现各路口信号灯控制的有限合作。然而，目前集中式评价的MARL方法仍存在信誉分配问题，即中心Critic网络只能根据联合动作策略向所有智能体返回相同的价值，这样每个智能体单独对全局网络的贡献无法被准确地评估，导致每个智能体的策略改进的方向不准确，因此目前的集中式MARL方法在路网交通信号控制中的合作效率低，导致在交通效率上路网的平均车辆行程延迟较高。

发明内容

[0006] 本发明目的在于针对上述现有技术的不足，提出一种交通信号协同控制方法，以集中式Critic网络中高效的协作奖励分配机制设计，为路网中分布控制的信号灯智能体提供准确的个体协作策略改进指导，提高信号灯智能体间的合作效率，降低路网的平均车辆行程延迟，实现路网中交通信号的高效协同控制。

[0007] 为实现上述目的，本发明采取的技术方案包括如下步骤：

[0008] (1)构建路网交通信号控制仿真环境：

[0009] 构建由路口集合$I=\{I_1,I_2,\cdots,I_n,\cdots,I_N\}$及与其对应智能体集合 $a=\{a_1,a_2,\cdots,a_n,\cdots,a_N\}$组成的交通信号协同控制场景，其中，N为路口集合中的路口总数，a_n表示对应N个智能体中的第n个智能体，I_n表示N个路口中的第n个路口，每个路口中均存在车道集合$L_n=\{L_n^1,L_n^2,\cdots,L_n^m,\cdots,L_N^M\}$，$L_n^m$表示路口$I_n$的M个车道中的第m个车道，M为车道集合中的车道总数，$0\leqslant n\leqslant N$，$0\leqslant m\leqslant M$，$N\geqslant 2$，$M\geqslant 2$；

[0010] (2)获取训练样本集Y：

[0011] 每个智能体a_n采集T个时间步长的交通状态信息，每个时间步长的交通状态信息包括：各路口智能体在每个时刻t下的交通状态观察s_t^n、执行动作u_t^n、执行动作后得到的协作奖励$\hat{r}_t^n$，$0\leqslant t\leqslant T$，用T个时间步长的交通状态信息构成样本规模为N×T的训练样本集Y；

[0012] (3)构建DRMA网络模型H：

[0013] (3a)建立由7个全连接层依次级联组成的Critic神经网络；

[0014] (3b)建立由5个全连接层依次级联组成的Actor神经网络；

[0015] (3c)将Critic神经网络和Actor神经网络并行排布构成DRMA网络模型H；

[0016] (4)设计DRMA网络模型H的目标优化函数J：

[0017] 根据路网中的信号灯智能体通过Actor网络π输出策略执行相应动作的机制，采用动作价值Q^{π}评估智能体在交通状态s_t下执行动作u_t的价值，设计DRMA网络模型H的如下目标优化函数，以使智能体的动作价值Q^{π}的期望达到最大：

[0018] $$\max_{\theta_\pi} J(\theta_\pi)=\max_{\theta_\pi} E\left[Q^{\pi}\mid\pi(\theta_\pi)\right]$$

[0019] 其中，$J(\theta_\pi)$为Actor网络π输出策略条件下路网中智能体动作价值Q^{π}的期望值，θ_π

为Actor网络π的神经网络参数；

[0020] (5)对DRMA网络模型H进行迭代训练：

[0021] (5a)初始化迭代次数为e，最大迭代次数为E，E≥2000，e＝1；

[0022] (5b)将训练集Y作为DRMA网络模型H的输入，Actor网络根据当前时刻t的联合交通状态信息s_t输出每个智能体要执行的动作概率分布$p_{\theta\pi}$，同时Critic网络根据当前时刻 t每个智能体选择执行的动作u_t^n和联合交通状态信息s_t评估智能体执行动作u_t^n后获得的价值Q^{π}，随后Critic网络根据Q^{π}得到每个智能体在合作中的个体贡献A_t^n并反馈给 Actor网络，Actor网络根据每个时刻的A_t^n得到其参数θ_{π}的更新方向$\nabla_{\theta_{\pi}}J(\theta_{\pi})$；

[0023] (5c)采用梯度上升法并行更新Actor网络参数θ_{π}和Critic网络参数θ_c，更新过程按照时间步长依次进行，每T个时间步长更新记作一次迭代训练，其中，T≥3000，为一次迭代训练的最大时间步长，执行一次迭代训练后，e＝e+1；

[0024] (5d)重复执行(5b)和(5c)，直到满足e≥E，则训练结束，得到的训练好的DRMA 网络模型为H^*；

[0025] (6)获取交通信号协同控制方案：

[0026] (6a)采集路网当前最新的交通状态信息，构建与(2)中训练样本Y结构相同的测试样本F；

[0027] (6b)将测试样本F输入至训练好的DRMA网络模型H^*，该模型中的Actor网络根据测试样本F输出每个时刻全局智能体的动作概率分布；

[0028] (6c)每个智能体根据最大概率原则输出各自最优的协作动作u_n^*，得到全局智能体的最优协作动作集合$\{u_1^*,u_2^*,\cdots,u_n^*,\cdots,u_N^*,\}$，该集合为该路网的交通信号协同控制方案。

[0029] 本发明与现有技术相比，具有以下优点：

[0030] 1)本发明通过集中式Critic网络评估每个智能体在路网交通信号协同控制中不同的个体贡献，对各智能体的Actor网络参数反馈各自相应的改进方向，能够激励各智能体高效地进行协作学习，克服了现有集中式评价方法存在的信誉分配问题，提高了路网中信号灯智能体间的合作效率，降低了路网的平均车辆行程延迟。

[0031] 2)本发明在所构建的训练样本中对各智能体的奖励进行了空间加权以加强各彼此间的合作，通过该空间加权奖励，各智能体能够接收周围邻居在同一时刻所执行的动作以及返回的奖励，将彼此独立的奖励机制相互耦合，进一步加强了智能体在路网交通信号控制中的协作效率。

附图说明

[0032] 图1为本发明的实现流程图；

[0033] 图2为本发明中DRMA网络模型H的结构示意图；

[0034] 图3为分别用本发明和现有方法对目标路网进行交通信号控制的仿真对比图；

[0035] 图4为分别用本发明和现有方法对目标路网进行车辆行程延迟的仿真对比图。

具体实施方式

[0036] 具体实现方式

[0037] 以下结合附图对本发明的实施例和效果进一步详细描述。

[0038] 参照图1,本实例的实现步骤如下:

[0039] 步骤1,构建路网交通信号控制仿真环境。

[0040] 构建由路口集合I及与其对应智能体集合a组成的交通信号协同控制场景,公式表示如下:

[0041] $I=\{I_1,I_2,\cdots,I_n,\cdots,I_N\}$

[0042] $a=\{a_1,a_2,\cdots,a_n,\cdots,a_N\}$

[0043] 其中,N为路口集合中的路口总数,a_n表示对应N个智能体中的第n个智能体,I_n表示N个路口中的第n个路口,每个路口中均存在车道集合L_n,公式表示如下:

[0044] $$L_n=\{L_n^1,L_n^2,\cdots,L_n^m,\cdots,L_N^M\}$$

[0045] 其中,L_n^m表示路口I_n的M个车道中的第m个车道,M为车道集合中的车道总数,$0\leqslant n\leqslant N$,$0\leqslant m\leqslant M$,$N\geqslant 2$,$M\geqslant 2$;

[0046] 本实施例中,采用LuST城市路网作为交通信号控制场景,K=22,M=24。

[0047] 步骤2,获取训练样本集Y。

[0048] 2.1)采集t时刻路口I_n第m条车道上的车辆数$v_t^{n,m}$,计算路口I_n所有车道上的车辆总数s_t^n:

[0049] $$s_t^n=\sum_{m=1}^{M}v_t^{n,m}$$

[0050] 式中,s_t^n记作智能体a_n在时刻t下的交通状态观察;

[0051] 2.2)采集智能体a_n在时刻t下的执行动作u_t^n,即该时刻交通灯的相位信号动作,本实施例中,交通灯的信号动作采用8相位模式;

[0052] 2.3)采集t时刻路口I_n处的车辆流出量out_t^n和流入量in_t^n,计算该时刻路口I_n处的车辆净流出量r_t^n:

[0053] $$r_t^n=out_t^n-in_t^n$$

[0054] 式中,r_t^n记作智能体a_n在t时刻执行动作后收到的奖励;

[0055] 2.4)对2.3)中智能体a_n的奖励r_t^n进行空间加权,获得智能体a_n的协作奖励$\hat{r}_t^n$:

[0056] $$\hat{r}_t^n=r_t^n+\alpha\sum_{i\in K(n)}r_t^i$$

[0057] 其中,K(n)表示智能体a_n的相邻智能体集合,α为空间加权因子,协作奖励$\hat{r}_t^n$使智能体能够考虑周围邻居的执行动作和奖励以加强彼此间的协作,本实施例中,α=0.8;

[0058] 2.5)将t时刻在步骤2.1)、2.2)、2.4)中得到的智能体a_n的交通状态观察s_t^n、执行动作u_t^n和协作奖励$\hat{r}_t^n$三者集合构成一个训练样本$y_{n,t}$:

[0059] $y_{n,t}=\{s_t^n,u_t^n,\hat{r}_t^n\}$；

[0060] 2.6)对N个智能体重复进行步骤2.1)至步骤2.5)，按照时间步长共进行T步，本实施例中，T＝3600，获得N×T个训练样本，构成训练样本集Y，形式如下：

[0061] $$Y=\begin{bmatrix} y_{1,1} & \cdots & y_{1,t} & \cdots & y_{1,T} \\ \vdots & \ddots & & & \vdots \\ y_{n,1} & & \ddots & & y_{n,T} \\ \vdots & & & \ddots & \vdots \\ y_{n,1} & \cdots & y_{n,t} & \cdots & y_{N,T} \end{bmatrix}_{N\times T}$$

[0062] 其中，$y_{n,t}$表示智能体a_n在t时刻构建的一个训练样本。

[0063] 步骤3，构建DRMA网络模型H。

[0064] 参照图2，本步骤的具体实现如下：

[0065] 3.1)建立由7个全连接层依次级联组成的Critic神经网络；

[0066] 该Critic神经网络中顺次级联的7个全连接层为：输入层→第一隐藏层→第二隐藏层→第三隐藏层→第四隐藏层→第五隐藏层→线性输出层；

[0067] 该Critic神经网络的输入数据是规模为N×M+N维的联合状态向量，五个隐藏层的神经元规模依次为380、250、160、80、20，每个隐藏层的输出均使用ReLU函数激活，输出层输出的数据为N维的联合价值向量。

[0068] 3.2)建立由5个全连接层依次级联组成的Actor神经网络；

[0069] 该Actor神经网络中顺次级联的5个全连接层为：输入层→第一隐藏层→第二隐藏层→第三隐藏层→SoftMax输出层；

[0070] 该Actor神经网络的输入数据是M维的局部交通状态观察向量，三个隐藏层的神经元规模依次为256、128、64，每个隐藏层的输出均使用ReLU函数激活，SoftMax层输出智能体执行动作的概率分布向量；

[0071] 3.3)将Critic神经网络和Actor神经网络并行排布构成DRMA网络模型H，其中：

[0072] Actor网络负责收集局部交通状态并执行局部最优协作控制动作；

[0073] Critic网络负责根据全局交通状态对Actor网络输出的动作策略进行价值评估并反馈给Actor网络，为Actor网络提供参数优化方案。

[0074] 步骤4，设计DRMA网络模型H的目标优化函数J。

[0075] 根据路网中的信号灯智能体通过Actor网络π输出策略执行相应动作的机制，采用动作价值Q^{π}评估智能体在交通状态s_t下执行动作u_t的价值，设计DRMA网络模型H的如下目标优化函数，以使智能体的动作价值Q^{π}的期望达到最大：

[0076] $$\max_{\theta_\pi} J(\theta_\pi)=\max_{\theta_\pi} E\left[Q^{\pi}\mid\pi(\theta_\pi)\right]$$

[0077] 其中，$J(\theta_\pi)$为Actor网络π输出策略条件下路网中智能体动作价值Q^{π}的期望值，θ_π为Actor网络π的神经网络参数。

[0078] 步骤5，对DRMA网络模型H进行迭代训练。

[0079] 5.1)初始化迭代次数为e，最大迭代次数为E，E≥2000，e＝1，本实施例中，E＝2000；

[0080] 5.2)Critic网络通过训练集Y计算每个智能体的个体贡献A_t^n并反馈给Actor网络，以计算Actor网络的参数优化方向$\nabla_{\theta_\pi} J(\theta_\pi)$：

[0081] 5.2.1)将训练集Y作为DRMA网络模型H的输入，Critic网络根据当前时刻t路网中智能体的联合动作u_t和全局交通状态信息s_t计算全局动作价值Q^π：

[0082] $$Q^\pi(s_t,u_t)=E\left[\sum_{\xi=t}^{T}\gamma^{\xi-t}\hat{r}_t(s_\xi,u_\xi)\,|\,s_t,u_t\right]$$

[0083] 式中，$\hat{r}_t(s_t,u_t)$为全局智能体的空间加权协作奖励，γ为未来回报折扣因子，本实施例中，γ＝0.99；

[0084] 5.2.2)Critic网络根据全局动作价值Q^π和全局智能体动作u_t计算得出每个智能体在交通信号协同控制中的个体贡献A_t^n，并反馈给Actor网络：

[0085] $$A_t^n=Q^\pi(s_t,u_t)-\sum_{u_t^n}p_{\theta_\pi}^n(u_t^n\,|\,s_t^n)Q^\pi\left(s_t,(u_t^{-n},u_t^n)\right)$$

[0086] 式中，$\sum_{u_t^n}p_{\theta_\pi}^n(u_t^n\,|\,s_t^n)Q^\pi\left(s_t,(u_t^{-n},u_t^n)\right)$表示对除智能体$a_n$自身外其他智能体在t时刻后的联合动作价值估计，$u_t=(u_t^n,u_t^{-n})$表示全局的执行动作分为两个部分：自身执行动作$u_t^n$和其他智能体的执行动作集合$u_t^{-n}$，$p_{\theta_\pi}^n(u_t^n\,|\,s_t^n)$表示Actor网络根据当前时刻t智能体$a_n$的局部观察$s_t^n$输出执行动作$u_t^n$的概率分布；

[0087] 5.2.3)Actor网络根据Critic网络在每个时刻t输出的个体贡献A_t^n得出其参数θ_π的更新方向$\nabla_{\theta_\pi} J(\theta_\pi)$：

[0088] $$\nabla_{\theta_\pi} J(\theta_\pi)=\mathrm{E}\left[\sum_{t=0}^{T}\nabla\theta_\pi\log p_{\theta_\pi}^n(u_t^n\,|\,s_t^n)A_t^n(s_t,u_t)\right]$$

[0089] 式中，θ_π为Actor网络π的神经网络参数，∇_{θ_π}为对θ_π求梯度的运算。

[0090] 5.3)采用梯度上升法并行更新Actor网络参数θ_π和Critic网络参数θ_c，公式如下：

[0091] $$\theta'_\pi=\theta_\pi+\beta_\pi\nabla_{\theta_\pi}J(\theta_\pi)$$

[0092] $$\theta'_c=\theta_c+\beta_c\nabla_{\theta_c}(\delta_t)^2$$

[0093] 其中，θ'_π为更新后的Actor网络参数，θ'_c为更新后的Critic网络参数；$\nabla_\theta J(\theta)$为智能体a_n获得的未来折扣回报期望的梯度，其作为Actor网络的优化步长，β_π为Actor 网络的学习率，$\nabla_{\theta_c}(\delta_t)^2$为Critic网络的优化步长，$\beta_c$为Critic网络的学习率，本实施例中，$\beta_\pi$＝0.05，$\beta_c$＝0.001；$\delta_t$表示一步时间误差，作为Critic网络进行优化的损失函数，根据空间加权奖励$\hat{r}_t$和动作价值Q^π计算得到：$\delta_t=\hat{r}_t+\gamma Q^\pi(s_t,u_t)-Q^\pi{}_{t-1}(s_t,u_t)$；

[0094] 本步骤的更新过程按照时间步长依次进行，每T个时间步长更新记作一次迭代训练，本实施例中，T＝3600，为一次迭代训练的最大时间步长，执行一次迭代训练后，e＝e+1；

[0095] 5.4)重复执行5.2)和5.3),直到满足e≥E,则训练结束,得到的训练好的DRMA 网络模型为H^*。

[0096] 步骤6,获取交通信号协同控制方案。

[0097] 6.1)采集路网当前最新的交通状态信息,构建与(2)中训练样本Y结构相同的测试样本F;

[0098] 6.2)将测试样本F输入至训练好的DRMA网络模型H^*,该模型中的Actor网络根据测试样本F输出每个时刻全局智能体的动作概率分布;

[0099] 6.3)每个智能体根据最大概率原则输出各自最优的协作动作u_n^*,得到全局智能体的最优协作动作集合$\{u_1^*,u_2^*,\cdots,u_n^*,\cdots,u_N^*,\}$,该集合为该路网的交通信号协同控制方案。

[0100] 以下结合仿真对本发明的效果作进一步说明:

[0101] 一、仿真条件

[0102] 本发明仿真实验的硬件条件为:Intel Xeon Gold 5218CPU和GEFORCE RTX 2080Ti GPU。

[0103] 本发明仿真实验的软件条件为:Ubuntu20.04操作系统和SUMO1.14.1交通仿真平台。

[0104] 仿真实验的具体参数如表1所示:

[0105] 表1:仿真实验参数表

[0106]

参数名称	参数值
目标路网	LuST 城市路网

[0107]

交通信号控制路口数 N	22
路口包含车道数 M	24
车道限速	30m/s
交通灯信号模式	8 相位
路网导入最大车辆数 V	4700
单次相位动作持续时间	10s
相位动作间隔时间	5s
模型迭代训练次数 E	2000
一次迭代训练的时间步长 T	3600

[0108] 二、仿真实验内容及结果分析:

[0109] 仿真实验1:在上述仿真条件下,分别使用本发明和现有方法IA2C和MA2C,在目标路网中获取交通信号协同控制方案,结果如图3,其中,纵坐标为全局动作价值Q^{π}表示所使用方法对路网交通效率提升的收益,横坐标为迭代训练次数;

[0110] 现有的IA2C方法是一种分布式独立控制的交通信号控制方法,该方法中的每个智能体在路网中彼此独立,只负责优化各自局部的交通信号控制方案,同时与环境交互而不相互协作。

[0111] 现有的MA2C方法是一种集中式评价的交通信号控制方法,该方法中的每个智能体能够在控制策略中考虑彼此的动作并以此进行协作,每个智能体通过合作收到统一的奖励

回报来更新参数，即每个智能体被分配相同的信誉。

[0112] 从图3可以看出，本发明的训练曲线最先收敛，且收敛后的全局动作价值Q^{π}最高，表明本发明在交通信号协同控制中的智能体协作效率是最高的，且获得的交通效率提升收益是最大的。

[0113] 仿真实验2，在上述仿真条件下，分别使用本发明和现有方法IA2C、MA2C和Fixed Phase，在目标路网中进行车辆行程延迟对比，结果如图4，其中左侧纵坐标为车辆平均行程延迟D，计算公式为：$D=\frac{1}{V}\sum_{i=1}^{V}\frac{TT_i^o}{TT_i^f}$，式中V为路网中导入的最大车辆数，$TT_i^o$为每辆车的实际行程时间，$TT_i^f$为每辆车的理想行程时间；右侧纵坐标为车辆数，横坐标为时刻。

[0114] 现有的Fixed Phase方法是一种采用固定相位模型信号的交通灯控制方案。

[0115] 从图4可以看出，随着路网中车辆数的变化趋势，本发明的交通信号协同控制方案在目标路网中的车辆平均行程延迟是最低的，表明本发明对目标路网的交通疏导是最有效的。

[0116] 以上描述仅使本发明的一个具体实例，并未构成对本发明的任何限制，显然对于本领域的专业人员来说，在了解了本发明内容和原理后，都可以在不背离本发明原理、结构的请看下，进行形式和细节上的各中修改和改变，但是这些基于本发明思想的修正和改变仍然在本发明的权利要求保护范围之内。

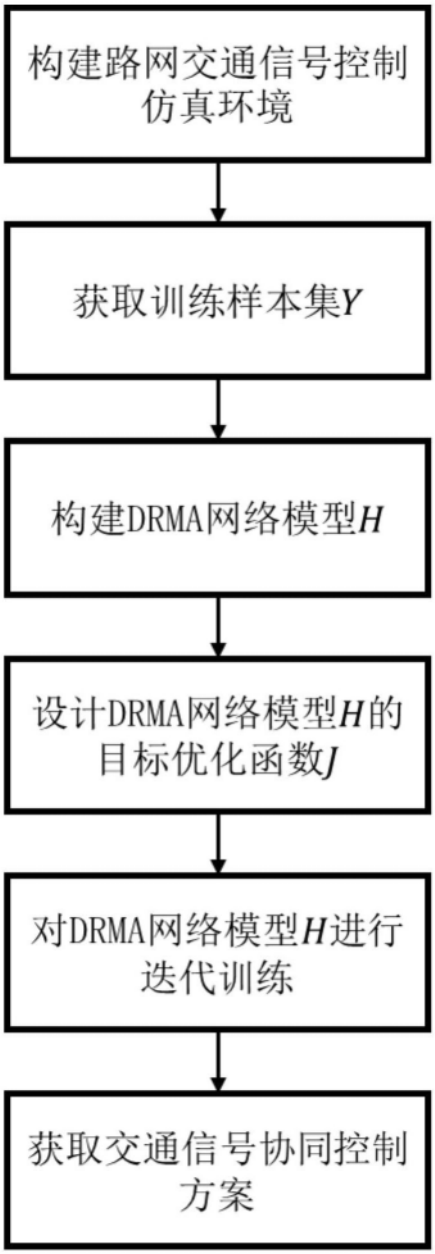
构建路网交通信号控制仿真环境
获取训练样本集Y
构建DRMA网络模型H
设计DRMA网络模型H的目标优化函数J
对DRMA网络模型H进行迭代训练
获取交通信号协同控制方案

图1

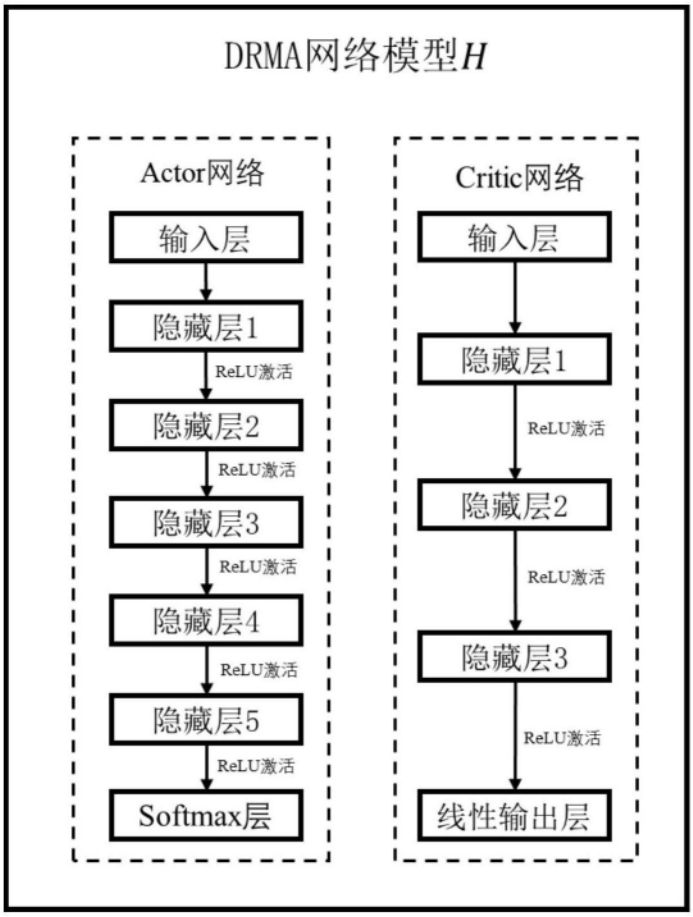
DRMA网络模型H
Actor网络
输入层
隐藏层1
ReLU激活
隐藏层2
ReLU激活
隐藏层3
ReLU激活
隐藏层4
ReLU激活
隐藏层5
ReLU激活
Softmax层
Critic网络
输入层
隐藏层1
ReLU激活
隐藏层2
ReLU激活
隐藏层3
ReLU激活
线性输出层

图2

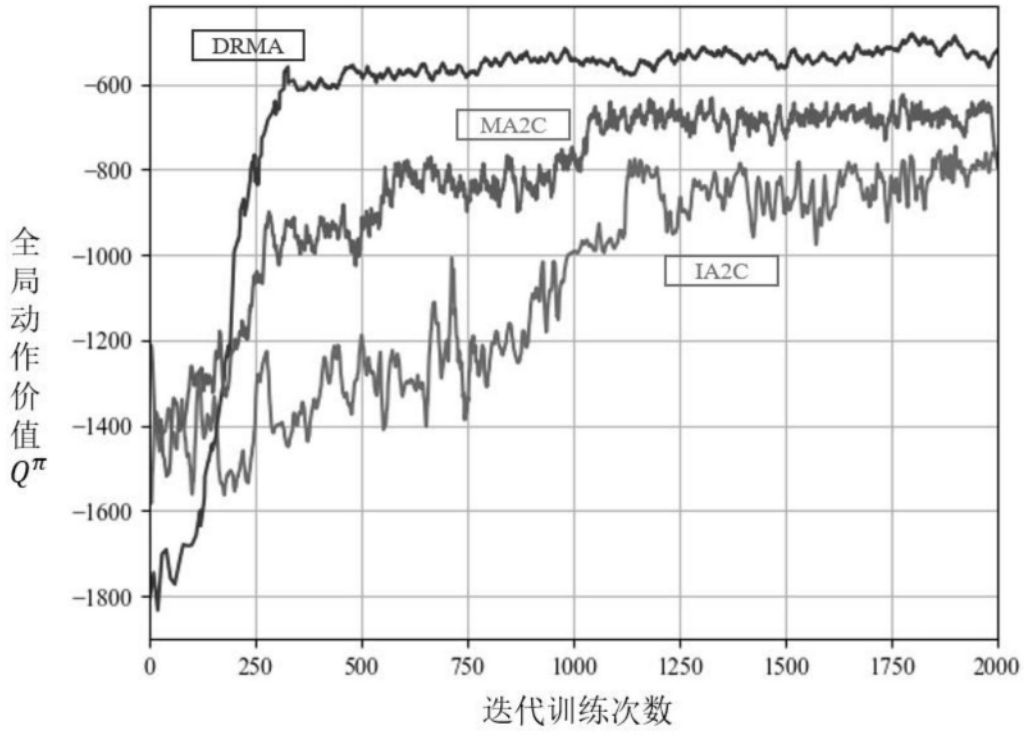
DRMA
MA2C
IA2C
全局动作价值 Q^{π}
-600
-800
-1000
-1200
-1400
-1600
-1800
0
250
500
750
1000
1250
1500
1750
2000
迭代训练次数

图3

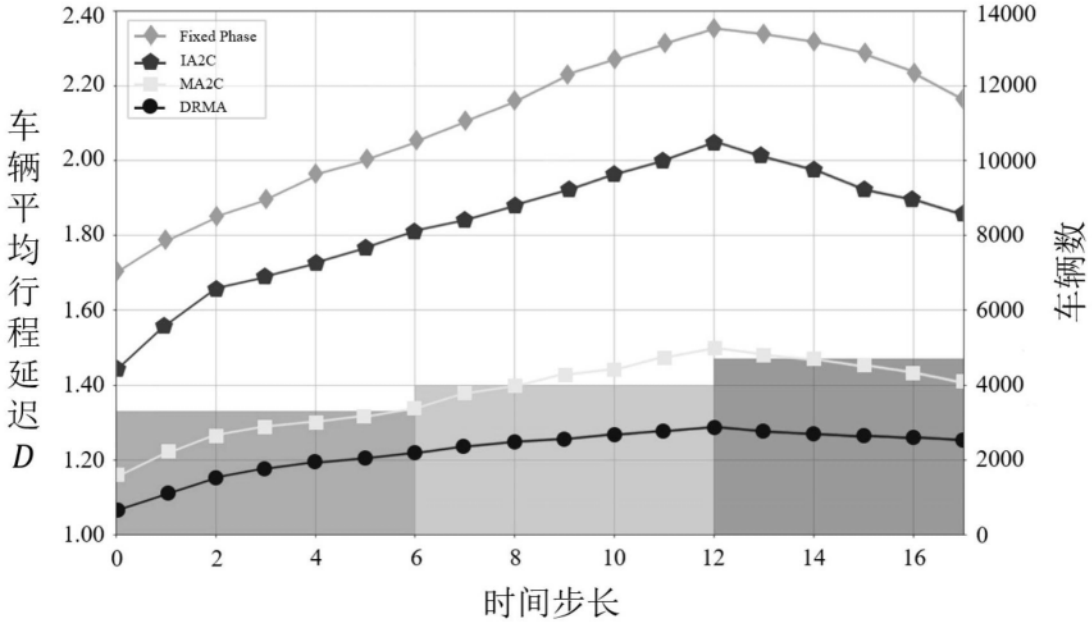
Fixed Phase
IA2C
MA2C
DRMA
车辆平均行程延迟 D
车辆数
2.40
2.20
2.00
1.80
1.60
1.40
1.20
1.00
14000
12000
10000
8000
6000
4000
2000
0
0
2
4
6
8
10
12
14
16
时间步长

图4