聚类与信息共享的多智能体深度强化学习协同控制交通灯

杜同春[①] 王 波*[①] 程浩然[①] 罗 乐[①] 曾能民[②]

[①](安徽师范大学计算机与信息学院 芜湖 241008)

[②](哈尔滨工程大学经济管理学院 哈尔滨 150001)

摘 要：该文提出一种适用于多路口交通灯实时控制的多智能体深度循环Q-网络(MADRQN)，目的是提高多个路口的联合控制效果。该方法将交通灯控制建模成马尔可夫决策过程，将每个路口的控制器作为智能体，根据位置和观测信息对智能体聚类，然后在聚类内部进行信息共享和中心化训练，并在每个训练过程结束时将评价值最高的值函数网络参数分享给其它智能体。在城市交通仿真软件(SUMO)下的仿真实验结果表明，所提方法能够减少通信的数据量，使得智能体之间的信息共享和中心化训练更加可行和高效，车辆平均等待时长少于当前最优的基于多智能体深度强化学习的交通灯控制方法，能够有效地缓解交通拥堵。

关键词：交通信号灯协同控制；集中训练分散执行；强化学习智能体聚类；生长型神经气；深度循环Q网络

中图分类号：TN929.5 **文献标识码**：A **文章编号**：1009-5896(2024)02-0538-08

DOI: 10.11999/JEIT230857

Multi-Agent Deep Reinforcement Learning with Clustering and Information Sharing for Traffic Light Cooperative Control

DU Tongchun[①] WANG Bo[①] CHENG Haoran[①] LUO Le[①] ZENG Nengmin[②]

[①](*School of Computer and Information, Anhui Normal University, Wuhu* 241008, *China*)

[②](*College of Economics and Management, Harbin Engineering University, Harbin* 150001, *China*)

Abstract: In order to improve the joint control effect of multi-crossing, Multi-Agent Deep Recurrent Q-Network (MADRQN) for real-time control of multi-intersection traffic signals is proposed in this paper. Firstly, the traffic light control is modeled as a Markov decision process, wherein one controller at each crossing is considered as an agent. Secondly, agents are clustered according to their position and observation. Then, information sharing and centralized training are conducted within each cluster. Also the value function network parameters of agents with the highest critic value are shared with other agent at the end of every training process. The simulated experimental results under Simulation of Urban MObility (SUMO) show that the proposed method can reduce the amount of communication data, make information sharing of agents and centralized training more feasible and efficient. The average delay of vehicles is reduced obviously compared with the state-of-the-art traffic light control methods based on multi-agent deep reinforcement learning. The proposed method can effectively alleviate traffic congestion.

Key words: Traffic light cooperative control; Centralized training with decentralized execution; Reinforcement learning agent cluster; Growing neural gas; Deep recurrent Q-network

1 引言

交通拥堵增加能源消耗和碳排放。本文根据交通状况进行多个路口协同地、动态地调节信号灯相位和时长，以减少车辆在路口的延迟、提高整体通行率。交通信号灯智能控制方法大致可分为3类：

第1类是时序控制转化成优化问题的方法，例如线性规划[1]、神经网络[2]等在一定程度上不能适用实时多变的多路口交通灯控制中，在实践中导致有限的可扩展性或最优性。

第2类是强化学习的方法，在设计好奖励函数、动作、状态的前提下，能够从样本数据中学习最优控制策略，工作包括文献[3–5]等。虽然在单路口交通信号控制场景下，减少了车辆等待时间，但是传统的强化学习方法由于对高维复杂状态的表达能力有限，而且通常仅适用于有限离散状态和动作空间的马尔可夫决策过程，因此其局限性很明显。

第3类为结合了深度学习的特征提取[6]能力和

收稿日期：2023-08-08；改回日期：2023-12-04；网络出版：2023-12-14

*通信作者： 王波 924405928@qq.com

强化学习的序列决策能力[7]的深度强化学习方法，是目前最为合适的交通信号灯控制方法。多数工作以深度Q网络(Deep Q-learning Network, DQN)为基础，选择更合理的观测，设计奖励函数，以及设计更好的特征提取网络，例如文献[8–11]。多路口交通信号灯协同控制具有巨大的状态与动作空间，造成维度爆炸、计算缓慢等问题，因此不能仅用单智能体深度强化学习方法解决。多智能体深度强化学习(Multi-Agent Deep Reinforcement Learning, MADRL)则更适用于解决分布式交通信号控制问题。MADRL的目标是提高多个智能体之间的协同效果，当前的主流框架是中心化训练去中心化执行(centralized training and decentralized execution)，其关键是设计协同训练的框架和方法，根据协同方法分为：(1)全局值函数分解方法，例如混合式多智能体Q学习(Q-value MIXtures, QMIX)[12]、Q变形分解多智能体强化学习(Q-value TRANsformation, QTRAN)[13]；(2)基于中心化的评判(Critic)方法，例如多智能体深度确定性策略梯度(Multi-Agent Deep Deterministic Policy Gradient, MADDPG)[14]、反事实多智能体策略梯度(COunterfactual Multi-Agent policy gradient, COMA)[15]等。MADRL用于交通信号灯控制的研究工作包括[16]：因果推理的MADRL[17]，基于生成对抗网络(Generative Adversarial Networks, GAN)的交通数据恢复法和MADRL相结合的信号灯控制[18]，引入注意力机制和领域认知一致性来解决智能体之间的合作问题[19]，广义强化学习提高智能体之间的有效交互[20]等。可见，目前工作的主要目的是提高智能体之间的通信和合作水平。

目前MADRL研究的实验都是在少量智能体的仿真游戏上进行的，对于多智能体马尔可夫决策过程(Markov Decision Process, MDP)问题，智能体之间共享的信息量非常巨大，导致通信延迟。实际上，对于距离较远、观测差异大的智能体共享信息的意义很小，甚至可能是噪声。如何共享有限且高效信息共享以提高协同水平是值得研究的问题。为此，针对建模成多智能体马尔可夫决策过程的多路口交通信号灯控制的实际问题，本文贡献如下：

(1)提出基于生长型神经气(Growing Neural Gas, GNG)的智能体聚类方法，对多路口的智能体进行聚类，目的是减少信息共享的数据量并找出具有相似位置和观测的智能体以实现更好的协同；

(2)提出全局值函数分解与观测共享的多智能体深度循环Q网络(Multi-Agent Deep Recurrent Q-Network, MADRQN)中心化训练算法，实现对聚类内部多路口的交通信号灯控制模型的训练；

(3)提出最优智能体参数分析的方法。在训练过程中，借鉴粒子群算法，将评价值最高的智能体的策略网络的参数分享给其它智能体，目的是加快训练速度并且使得全部智能体向最优参数协同进化。

本文的所有仿真实验均在城市交通仿真软件(Simulation of Urban MObility, SUMO)下进行，更接近真实路况。

2　智能体聚类

具有相似认知的智能体才能更好地协同完成任务。本文以位置和观测构成的混合特征，使用GNG对智能体聚类。GNG是一种动态自组织算法，不需要指定类别数，根据竞争的赫布(Hebbian)学习规则不断地更新网络的节点和连接，以学习输入矢量分布的拓扑结构[21]。交通路网可视为图结构，而车流量是动态变化的，因此每隔1 000个时间步对智能体重新聚类1次。使用迭代的生长型神经气算法对混合特征聚类，得到具有一致或相似认知的多个智能体，如图1(b)中的上面两个灰色圆和下面两个灰色圆分别是两个聚类内的智能体。然后在聚类内部进行信息共享和中心化训练，由此减少共享的数据量、提高协同效果。智能体聚类算法，如算法1中第5～14行所示。

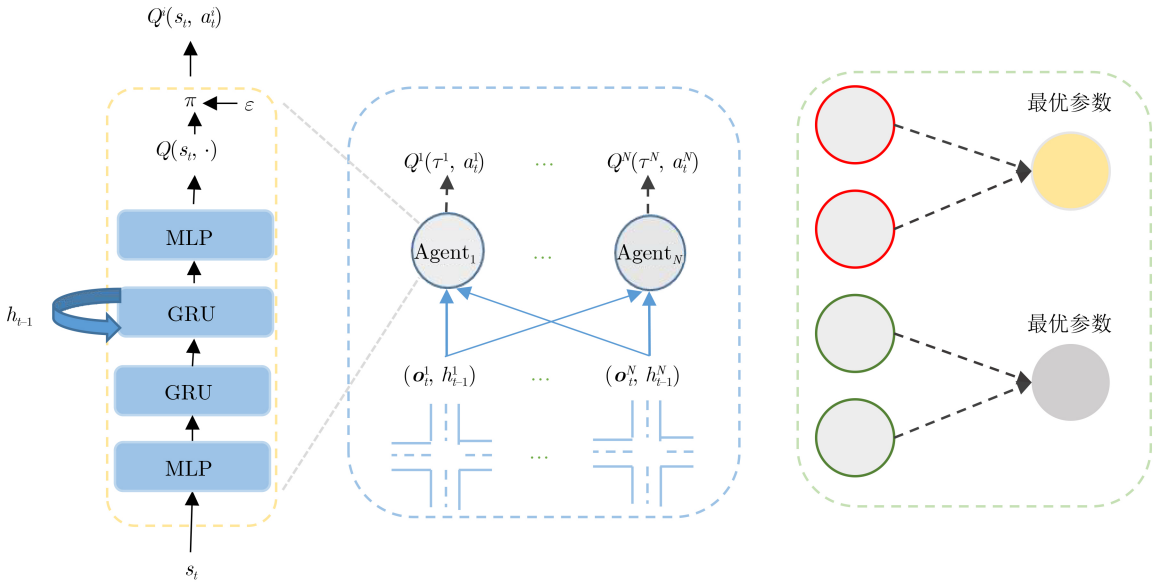
3　观测信息共享与最优参数分享的MADRQN中心化训练

仅在聚类内进行信息贡献极大地缓解了由环境变化造成的算法收敛困难的问题。在中心化训练阶段，设计如图1(a)，在t时刻，智能体能够以全局状态s_t和全部智能体上一时刻的隐状态h_{t-1}为输入，拟合状态动作值$Q^i(\tau^i,a^i)$，其中GRU为门控循环单元，用于提取时间维度上对交通状态的观测特征。智能体采用$\varepsilon-$贪婪策略选择动作a_t^i。图1所示是聚类内智能体的中心化训练，而且多个聚类之间可以并行训练，由此加速了多个智能体训练的速度。s_t为全局状态，即聚类内全部智能体的观测构成的联合观测，如式(1)，a_t为聚类内全部智能体的动作构成的联合动作，如式(2)。在去中心化执行阶段，依然可以采用共享观测的方法，而不用担心数据量大造成的通信延迟问题。得到$Q_t^i(s_t,a_t)$之后，第i个智能体的损失函数表示如式(3)

$$s_t=\{\boldsymbol{o}_t^1,\boldsymbol{o}_t^2,\cdots,\boldsymbol{o}_t^N\} \tag{1}$$

$$a_t=\{a_t^1,a_t^2,\cdots,a_t^N\} \tag{2}$$

$$L(\theta^i)=\frac{1}{M}\sum_{j=1}^{M}\left(y_j-Q^j(s_j,a_j)\right)^2 \tag{3}$$



(a) 中心化训练与去中心化执行　　(b) 参数分享

图 1 MADRQN算法框架

式(3)中M为训练样本的批数量，训练样本表示为$\{s_t,a_t,s_{t+1},r_t\}$。采用时序差分 (Temporal Difference, TD)法计算误差，y_j表示状态$s'=s_{t+1}$和动作$a'=a_{t+1}$在第j个样本的目标网络的值函数，如式(4)，r_j表示第j个样本的立即奖励

$$y_j = r_j + \gamma \cdot \max_{a'} Q^j(s', a') \tag{4}$$

中心化训练时，式(3)中误差反向传播以更新状态动作值函数网络的参数。式(5)是对一个批次样本的TD误差计算梯度，并根据式(6)进行参数更新，α表示学习率

$$\nabla_{\theta^i} L(\theta^i) = \frac{1}{M}\sum_{j=1}^{M}\left[-2(y_j - Q^j(s_j,a_j)) \cdot \frac{\partial Q^j(s_j,a_j)}{\partial \theta^i}\right] \tag{5}$$

$$\theta^i \leftarrow \theta^i - \alpha \cdot \nabla_{\theta^i} L(\theta^i) \tag{6}$$

在去中心化执行阶段，每个智能体使用学习到的策略π^{θ^i}，以聚类内全部智能体对交通状况的观测和上一步的隐状态作为输入，输出动作，如式(7)，其中A表示智能体动作空间

$$\begin{aligned} a_t^i &= \pi^{\theta^i}(\boldsymbol{o}_t^1, \cdots, \boldsymbol{o}_t^N, h_{t-1}^1, \cdots, h_{t-1}^i) \\ &= \arg\max_{a_t^i \in A} Q_t^i(s_t, a_t^i) \end{aligned} \tag{7}$$

如图1(b)所示，将智能体看作探索参数空间的粒子，每轮训练结束聚类内每个智能体的训练网络和目标网络的参数分别朝着全局和个体最优参数移动，以提高训练速度

$$v_{\theta^i} = v_{\theta^i} + c_1 \cdot (\text{pbest}_{\theta^i} - \theta^i) + c_2 \cdot (\text{gbest}_\theta - \theta^i) \tag{8}$$

$$\theta^i \leftarrow \theta^i + v_{\theta^i} \tag{9}$$

$$\text{gbest}_\theta = \arg\max_{\theta^i} Q^i \tag{10}$$

按式(8)和式(9)移动参数，pbest_{θ_i}和gbest_θ分别是单个智能体和全局的最优值函数网络参数。全局参数是获得最高评价值的值函数网络的参数，如式(10)。

整个算法的伪代码如算法1，注意在第15行和第20行、第16行和第21行，虽然对聚类和智能体进行遍历，但其实与环境的交互和模型训练，在不同的聚类以及聚类内的智能体之间，都是并行的。

4 实验环境与参数设计

将芜湖弋江区部分路网导入到SUMO，如图2(a)，将路口映射至图2(b)。

将交通信号灯控制建模成马尔可夫决策过程(Markov Decision Process, MDP)，对于MDP的6元组$<S, A, R, S', P, \gamma>$，其中S表示为智能体对路口交通状况的观测，具体如图3所示，观察的车道长度是400 m，设置车辆的长度是5 m，将进入路口的4个车道的车辆构成的向量相加得到最终的向量，1表示对应的位置有车，0表示无车；A表示动作空间，包括4个动作，{南北直行，南北左转，东西直行，东西左转}。绿灯的时长设置为10 s，红绿灯切换间隙执行4 s的黄灯。奖励函数R设置为路口全部车辆在两个相邻时间步的平均等待时长之差，因此奖励是负数。P表示状态转移概率，是由环境决定的，因此是未知的。γ表示立即奖励的折扣因子，设置为0.9。

算法1 MADRQN算法伪代码

初始化：MDRQN网络及目标网络，经验回放池D，训练轮数N_{e}=100，每轮训练步数T=300，智能体数量N=15，智能体位置向量$\boldsymbol{p}^i, i=1,2,\cdots,N$，聚类数量$N_{\mathrm{c}}$(未定，初始化为0)，聚类内智能体数量$N_{\mathrm{A}}$(未定，初始化为0)，聚类间隔步数$C$=1 000，聚类迭代次数$K$=2 500，节点移动系数$\varepsilon_{\mathrm{b}}=0.1$和$\varepsilon_{\mathrm{n}}=0.1$，边的最大年龄AM = 30，误差衰减系数$\beta=0.9$，参数移动系数$c_1=0.9$和$c_2=0.3$奖励折扣因子$\gamma=0.9$，学习率$\alpha=5\mathrm{e}-4$

(1) **for** ep=1 to N_{e} **do**

(2)　获取时间步t=0的路口观测$\boldsymbol{o}_0^i, i=1,2,\cdots,N$

(3)　初始化立即奖励$r_0^i=0$，动作$a_0^i=-1$，动作选择的贪婪参数$\varepsilon=\left(1-\dfrac{\mathrm{ep}}{M}\right)^2$

(4)　**while** $t<T$ **do**

#　　以下是每隔C个时间步对智能体聚类

(5)　**if** $t\%C$=0 **Then**

(6)　　从全部智能体的混合特征集合$\{(\boldsymbol{p}^i,\boldsymbol{o}_\tau^i)\}$中随机选择两个特征向量$\boldsymbol{v}_a$和$\boldsymbol{v}_b$，分别将其映射为GNG网络的初始节点$a$和$b$

(7)　　**for** k=0 to K **do**

(8)　　从集合$\{(\boldsymbol{p}^i,\boldsymbol{o}_\tau^i)\}$中选择一个新的特征向量$\boldsymbol{x}$，计算$\boldsymbol{x}$与网络节点对应向量的距离$||\boldsymbol{v}_j-\boldsymbol{x}||^2, j=1,2,\cdots,N_k$，$N_k$表示网络当前节点数量。然后找到距离最近和次近节点$s_1$和$s_2$，对应特征向量为$\boldsymbol{v}_{s_1}$和$\boldsymbol{v}_{s_2}$

(9)　　朝着x的方向，分别移动节点s_1和s_1的邻居节点$S_{N_{s_1}}$

$$s_1 \leftarrow s_1+\varepsilon_{\mathrm{b}}\cdot||\boldsymbol{x}-\boldsymbol{v}_{s_1}||_2^2$$

$$S_{N_{s_1}} \leftarrow S_{N_{s_1}}+\varepsilon_n\cdot||\boldsymbol{x}-\boldsymbol{v}_{N_{s1}}||_2^2$$

(10)　　若s_1与s_2没有边，则连接s_1和s_2，将所有包含s_1的边的年龄加1

(11)　　遍历网络全部的边，将年龄大于AM的边删除，再删除孤立的节点

(12)　　找出累计误差$\sum_{k=0}^{K}\left\|\boldsymbol{x}_k-\boldsymbol{v}_{s_j}\right\|_2^2$最大的节点$q$和次最大的节点$p$，在二者的中点插入新的节点$r$，分别连接$r$与$q$、$r$与$p$，删除$p$与$q$的连接

(13)　　将节点p与q的误差乘以衰减系数β，将q的误差作为r的误差14

end for

#　以下与环境交互，收集经验数据

(14)　**for** $i_{\mathrm{c}}=1$ to N_{c} **do** #不同的聚类

(15)　　**for** $i_{\mathrm{a}}=1$ to N_{A} **do** #聚类内不同的智能体

(16)　　获取观测$\boldsymbol{o}_t^{i_{\mathrm{a}}}$，根据策略执行动作$a_t^{i_{\mathrm{a}}}$，接收奖励$r_t^{i_{\mathrm{a}}}$，获取新观测$o_{t+1}^{i_{\mathrm{a}}}$

(17)　　将数据$\left(\boldsymbol{o}_t^1,\cdots,\boldsymbol{o}_t^{N_{\mathrm{A}}},a_t^1,\cdots,a_t^{N_{\mathrm{A}}},r_t^1,\cdots,r_t^{N_{\mathrm{A}}},\boldsymbol{o}_{t+1}^1,\cdots,\boldsymbol{o}_{t+1}^{N_{\mathrm{A}}}\right)$存入经验池D

(18)　**end while**

#　以下为训练部分

(19)　**for** $i_{\mathrm{c}}=1$ to N_{c} **do** #不同的聚类

(20)　　**for** $i_{\mathrm{a}}=1$ to N_{A} **do** #聚类内不同的智能体

(21)　　**for batch in**由D构造的DataLoader **do**

(22)　　共享观测$s_t=\left\{\boldsymbol{o}_t^1,\boldsymbol{o}_t^2,\cdots,\boldsymbol{o}_t^{N_{\mathrm{A}}}\right\}, s_{t+1}=\left\{\boldsymbol{o}_{t+1}^1,\boldsymbol{o}_{t+1}^2,\cdots,\boldsymbol{o}_{t+1}^{N_{\mathrm{A}}}\right\}$动作、奖励分别为$a_t^{i_{\mathrm{a}}}$和$r_t^{i_{\mathrm{a}}}$

(23)　　根据式(4)—式(6)，计算损失和更新智能体的网络参数

(24)　　**end for**

(25)　　**end for**

(26)　　根据式(8)—式(10)将网络参数向聚类内最优和个体历史最优移动

(27)　**end for**

(28)　清空经验池D

(29) **end for**

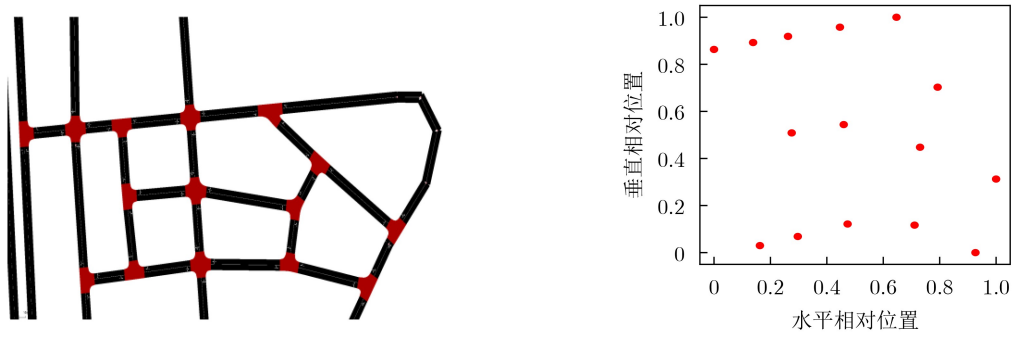
实验设置整个路网的车流量为1 000 辆/s，在实验中随机分配给所有路口，为了展示聚类算法的效果，算法参数设置如表1。

5　实验结果与分析

对于算法1中的智能体聚类、基于观测共享的中心化训练、参数移动等方法，开展实验并对结果进行分析。

5.1　根据位置和观测对智能体聚类

在表1参数设置下，截取算法1中随机训练过程中的一轮聚类结果，如图4所示。其中图4(a)是中间过程的GNG网络拓扑，图4(b)是最终的聚类结



(a) 路网 (b) 路口映射到网络中的节点

图 2 路网拓扑图及路口映射图

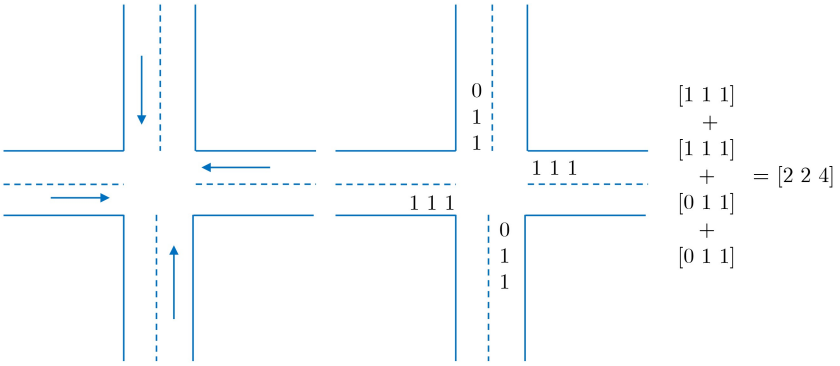


图 3 路口状态定义

表 1 实验及算法参数设置

参数	数值(范围)
观测范围	各路口进车道400 m
智能体数量N	15
车辆长度	5 m
最小车辆间距	1～2 m
最大速度	40～50 km/h
车辆加速度	0.5～1 m/s^2
车辆减速度	–5～–3.5 m/s^2
训练轮数M	100
每轮最大步数T	5 400
学习率α	5e–4
折扣因子γ	0.9
贪婪参数ε	初始为1递减至0.001
回放池大小$\|D\|$	5 000

果，将智能体聚成了4类，分别用红色点、绿色点、蓝色点、深蓝色点表示。由此可得，聚类结果主要由路口位置决定，同时受到车流量的影响，当距离过远即便交通流相似也不会被聚为一类。由于交通状况是不断变化的，聚类结果也将变化。

5.2 信息共享的MADRQN独立训练

在对全部路口的交通信号灯控制智能体聚类的基础上，采用观测与动作共享的MADRQN算法对智能体的值函数网络进行中心化训练，并在每一轮训练结束，将每个智能体的训练网络和目标网络的参数分别朝着全局和个体最优参数移动。以下分别从整体的算法效果与其它多智能体深度强化学习算法的效果对比、聚类与不聚类的实验结果、聚类与不聚类的训练时间3个方面进行分析，验证本文的假设。

5.2.1 MADRQN算法整体的控制效果与分析

图5是本文提出的信息共享与参数分享的MADRQN算法与其它典型的多智能体深度强化学习算法，包括QMIX、值分解网络(Value Decomposition, Network, VDN)、深度循环Q网络(Deep Recurrent Q-learning Network, DRQN)，对15个路口的交通信号灯控制时计算的每轮平均奖励的对比。横坐标是训练的轮数，共100轮，纵坐标是计算的每一轮的平均奖励，即对一轮中全部步数的负立即奖励求和，得到累积负奖励，然后再对每个算法运行3遍(每一遍都是100轮，每轮步数$T_i=300$)，对3遍的累积负奖励求平均，得到平均每轮奖励aer，如式(11)所示，目的一是为了观察算法的稳定性，二是可以计算标准差(如图5中曲线的上下着色部分)。

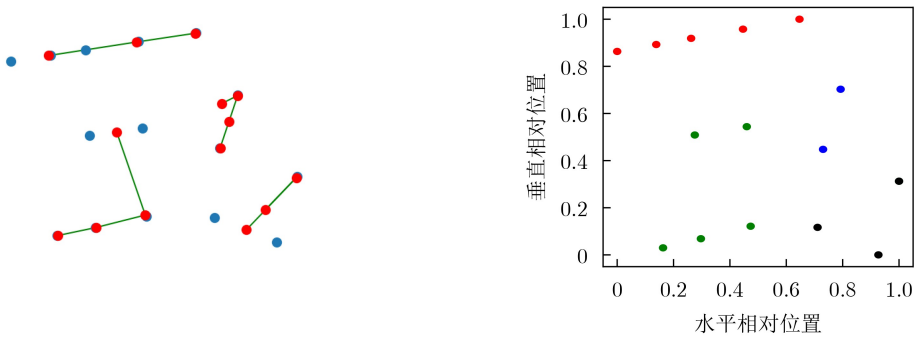
$$\text{aer}=\frac{1}{3}\sum_{i=1}^{3}\sum_{t=1}^{T_i}r_t,\ r_t<0 \tag{11}$$

从图5(a)和图5(b)可以看出，MADRQN略优于VDN，原因是使用GRU提取了时间维度上的相关特征，对Q值的预测更加准确。相比DRQN进行独立训练，MADRQN的效果更好，原因是通过聚类将具有相似位置和观测的智能体进行观测共享，解决了环境部分可观测并缓解了环境非静态的问题，以及通过中心化训练，提高了智能体之间的协同水平。相比于QMIX以全局状态进行参数化超网络的值函数分解，由于MADRQN采用了聚类并在聚类内部进行观测共享，智能体的环境更相似而且信息共享更加准确，因此，MADRQN的效果优于QMIX。

5.2.2 智能体聚类对控制效果的影响与分析

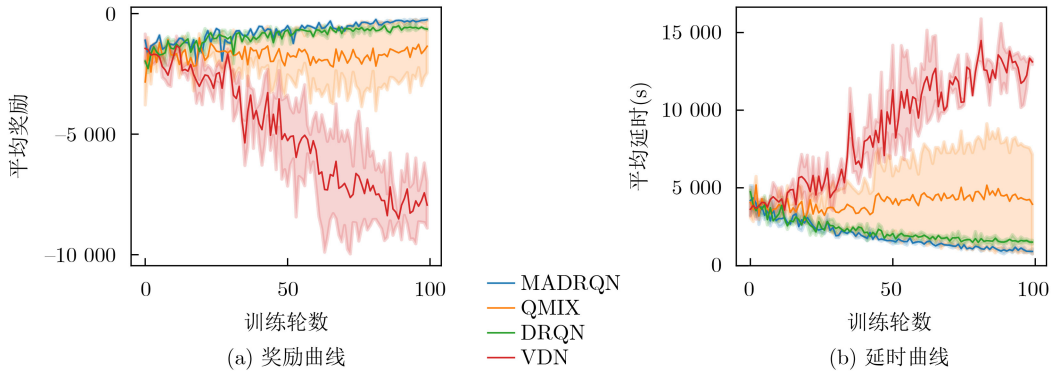
本文一个重要假设是在聚类内部共享的信息数据量少，而且信息的相关性和作用更大，更有利于提高中心化训练效果。

为验证该假设，将MADRQN算法分别应用于全部15个路口的智能体以及GNG算法得到的4个聚类的智能体，计算方法与4.2.1中相同，如图6所示，蓝色曲线(有GNG)是做聚类后分别中心化训练的效果，而橙色曲线(无GNG)是对全部智能体直接用MADRQN算法的效果，如图6(a)、图6(b)可见，不做智能体聚类时的每轮累积奖励更小、平均每轮累



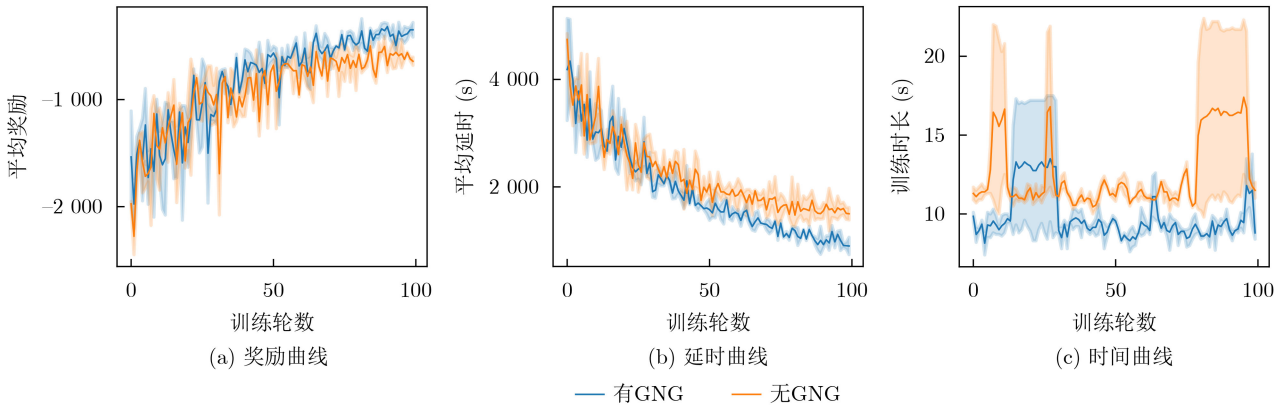
(a) 聚类过程网络拓扑　　(b) 聚类结果

图 4 根据表2参数设置的智能体聚类结果



(a) 奖励曲线　　(b) 延时曲线

图 5 算法对比图



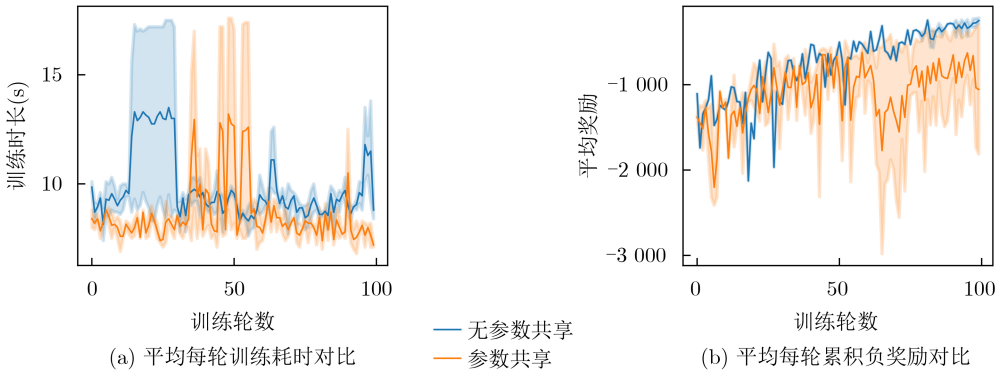
(a) 奖励曲线　　(b) 延时曲线　　(c) 时间曲线

图 6 聚类算法对比图

积延迟更大。因为不做聚类时在所有智能体之间进行信息共享与参数分享，状态空间大导致训练缓慢。此外，距离较远路口分享的参数与共享的观测不合适导致了整体算法性能下降。图6(c)反应了多个聚类并行训练的训练时长，明显低于不做聚类的训练用时。由此，验证了本文对智能体聚类作用的假设。

5.2.3 最优参数分享的影响与分析

如图7，橙色曲线(参数共享)表示在每轮训练结束时将Q值最大的值函数网络的参数分享给其它智能体。蓝色曲线(无参数共享)表示不做参数分享。由图7(a)可见，参数分享能够降低训练时间。但由图7(b)可得参数共享的方式可能会牺牲一定的性能。



(a) 平均每轮训练耗时对比 (b) 平均每轮累积负奖励对比

图 7 最优参数实验对比图

6 结束语

本文对智能体根据位置和观测等信息聚类，提高了信息共享的效率和可行性，解决了环境部分可观测和非静态的问题，提出的最优参数分享对于加速训练有明显效果，本文方法对于缓解城市拥堵也存在一定意义。本文不足之处在于，对智能体聚类所用特征需要进一步研究，聚类的依据应该还包括智能体对自身任务的理解，如何表示任务目标并将任务、观测、位置等进行融合，以表示智能体的认知，是值得研究的课题。

参 考 文 献

[1] PANDIT K, GHOSAL D, ZHANG H M, *et al*. Adaptive traffic signal control with vehicular ad hoc networks[J]. *IEEE Transactions on Vehicular Technology*, 2013, 62(4): 1459–1471. doi: 10.1109/TVT.2013.2241460.

[2] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, *et al*. Deep residual learning for image recognition[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770–778. doi: 10.1109/CVPR.2016.90.

[3] 邵明莉, 曹鹗, 胡铭, 等. 面向优先车辆感知的交通灯优化控制方法[J]. 软件学报, 2021, 32(8): 2425–2438. doi: 10.13328/j.cnki.jos.006191.

SHAO Mingli, CAO E, HU Ming, *et al*. Traffic light optimization control method for priority vehicle awareness[J]. *Journal of Software*, 2021, 32(8): 2425–2438. doi: 10.13328/j.cnki.jos.006191.

[4] HADDAD T A, HEDJAZI D, and AOUAG S. A new deep reinforcement learning-based adaptive traffic light control approach for isolated intersection[C]. The 5th International Symposium on Informatics and its Applications, M'sila, Algeria, 2022: 1–6. doi: 10.1109/ISIA55826.2022.9993598.

[5] GENDERS W and RAZAVI S. Using a deep reinforcement learning agent for traffic signal control[J]. arXiv preprint arXiv: 1611.01142, 2016.

[6] TIGGA A, HOTA L, PATEL S, *et al*. A deep Q-learning-based adaptive traffic light control system for urban safety[C]. The 4th International Conference on Advances in Computing, Communication Control and Networking, Greater Noida, India, 2022: 2430–2435. doi: 10.1109/ICAC3N56670.2022.10074123.

[7] 邹翔宇, 黄崇文, 徐勇军, 等. 基于深度学习的通信系统中安全能效的控制[J]. 电子与信息学报, 2022, 44(7): 2245–2252. doi: 10.11999/JEIT211611.

ZOU Xiangyu, HUANG Chongwen, XU Yongjun, *et al*. Secure energy efficiency in communication systems based on deep learning[J]. *Journal of Electronics & Information Technology*, 2022, 44(7): 2245–2252. doi: 10.11999/JEIT211611.

[8] 唐伦, 李质萱, 蒲昊, 等. 基于多智能体深度强化学习的无人机动态预部署策略[J]. 电子与信息学报, 2023, 45(6): 2007–2015. doi: 10.11999/JEIT220513.

TANG Lun, LI Zhixuan, PU Hao, *et al*. A dynamic pre-deployment strategy of UAVs based on multi-agent deep reinforcement learning[J]. *Journal of Electronics & Information Technology*, 2023, 45(6): 2007–2015. doi: 10.11999/JEIT220513.

[9] KANG Leilei, HUANG Hao, LU Weike, *et al*. A dueling deep Q-network method for low-carbon traffic signal

control[J]. *Applied Soft Computing*, 2023, 141: 110304. doi: 10.1016/j.asoc.2023.110304.

[10] TUNC I and SOYLEMEZ M T. Fuzzy logic and deep Q learning based control for traffic lights[J]. *Alexandria Engineering Journal*, 2023, 67: 343–359. doi: 10.1016/j.aej.2022.12.028.

[11] BÁLINT K, TAMÁS T, and TAMÁS B. Deep reinforcement learning based approach for traffic signal control[J]. *Transportation Research Procedia*, 2022, 62: 278–285. doi: 10.1016/j.trpro.2022.02.035.

[12] RASHID T, SAMVELYAN M, DE WITT C S, *et al*. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement Learning[C]. The 35th International Conference on Machine Learning, Stockholm, Sweden, 2018: 6846–6859.

[13] SON K, KIM D, KANG W J, *et al*. QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning[C]. The 36th International Conference on Machine Learning, Long Beach, USA, 2019: 5887–5896.

[14] LOWE R, WU Y I, TAMAR A, *et al*. Multi-agent actor-critic for mixed cooperative-competitive environments[C]. The 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017: 6382–6393.

[15] FOERSTER J, FARQUHAR G, AFOURAS T, *et al*. Counterfactual multi-agent policy gradients[C]. The 32nd AAAI Conference on Artificial Intelligence, New Orleans, USA, 2018. doi: 10.1609/aaai.v32i1.11794.

[16] SU Haoran, ZHONG Y D, DEY B, *et al*. EMVLight: A decentralized reinforcement learning framework for efficient passage of emergency vehicles[C]. The 36th AAAI Conference on Artificial Intelligence, 2021: 4593–4601. doi: 10.48550/arXiv.2109.05429.

[17] YANG Shantian, YANG Bo, ZENG Zheng, *et al*. Causal inference multi-agent reinforcement learning for traffic signal control[J]. *Information Fusion*, 2023, 94: 243–256. doi: 10.1016/j.inffus.2023.02.009.

[18] WANG Zixin, ZHU Hanyu, HE Mingcheng, *et al*. GAN and multi-agent DRL based decentralized traffic light signal control[J]. *IEEE Transactions on Vehicular Technology*, 2022, 71(2): 1333–1348. doi: 10.1109/TVT.2021.3134329.

[19] 丛珊. 基于多智能体强化学习的交通信号灯协同控制算法的研究[D]. [硕士论文], 南京信息工程大学, 2022. doi: 10.27248/d.cnki.gnjqc.2022.000386.

CONG Shan. Multi-agent deep reinforcement learning based traffic light cooperative control[D]. [Master dissertation], Nanjing University of Information Science & Technology, 2022. doi: 10.27248/d.cnki.gnjqc.2022.000386.

[20] ZHU Ruijie, LI Lulu, WU Shuning, *et al*. Multi-agent broad reinforcement learning for intelligent traffic light control[J]. *Information Sciences*, 2023, 619: 509–525. doi: 10.1016/j.ins.2022.11.062.

[21] FRITZKE B. A growing neural gas network learns topologies[C]. The 7th International Conference on Neural Information Processing Systems, Denver, USA, 1994: 625–632.

杜同春：男，讲师，硕士生导师，研究方向为深度强化学习、智慧交通等.

王　波：男，硕士生，研究方向为深度强化学习、智慧交通.

程浩然：男，硕士生，研究方向为深度强化学习、智慧交通.

罗　乐：男，讲师，研究方向为高性能并行计算.

曾能民：男，副教授，研究方向为大数据与智能决策、智能预测、供应链风险管理.

责任编辑：余　蓉