计算机应用
Journal of Computer Applications
ISSN 1001-9081,CN 51-1307/TP

《计算机应用》网络首发论文

题目：基于全局状态预测与公平经验重放的交通信号控制算法
作者：缪孜珺，罗飞，丁炜超，董文波
收稿日期：2024-01-19
网络首发日期：2024-04-09
引用格式：缪孜珺，罗飞，丁炜超，董文波．基于全局状态预测与公平经验重放的交通信号控制算法[J/OL]．计算机应用．
https://link.cnki.net/urlid/51.1307.TP.20240407.1337.006

Journal of Computer Applications ISSN 1001-9081
计算机应用 CODEN JYIIDU http://www.joca.cn

DOI:10.11772/j.issn.1001-9081.2024010066

基于全局状态预测与公平经验重放的交通信号控制算法

缪孜珺，罗飞*，丁炜超，董文波

(华东理工大学 信息科学与工程学院，上海 200237)

(*通信作者电子邮箱 luof@ecust.edu.cn)

摘 要: 为了应对交通拥堵，设计高效的交通信号控制算法能够显著增强现有交通网络下的车辆通行效率。尽管深度强化学习算法在单路口交通信号控制问题上已展现出卓越的性能，然而其在多路口环境下的应用仍然面临着重大的挑战——即因多智能体强化学习算法产生时间和空间部分可观测性而引发算法出现非平稳性问题，这会导致算法无法保证稳定地收敛。为此，提出一种基于全局状态预测与公平经验重放的多路口交通信号控制算法(IS-DQN)。一方面，基于不同车道的车流历史信息预测多交通路口的全局状态，扩展 IS-DQN 的状态空间，以避免算法产生空间部分可观测性而带来非平稳性问题。另一方面，为了应对传统经验重放策略的时间部分可观测性，IS-DQN 采用了蓄水池算法以保证经验重放池的公正性，进而避免其中的非平稳性问题。在复杂的多路口环境下应用 IS-DQN 进行三种不同的交通压力仿真实验，实验结果表明：在不同交通流情况下，尤其是在中低交通流量下，相对独立深度强化学习算法，IS-DQN 算法能够得到更低的车辆平均行驶时间，并表现出了更优的收敛性能与收敛稳定性。

关键词: 深度强化学习；交通信号控制；时序预测；蓄水池算法；长短期记忆网络

中图分类号:TP181 文献标志码： A

Traffic signal control algorithm based on overall state prediction and fair experience replay

MIAO Zijun, LUO Fei*, DING Weichao, DONG Wenbo

(School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China)

Abstract: In order to cope with traffic congestion, designing efficient traffic signal control algorithms can significantly enhance the traffic efficiency of vehicle. Although deep reinforcement learning algorithms have shown excellent performance in single intersection traffic signal control problems, their application in multi-intersection environments still faces significant challenges - the non-stationarity problem caused by the spatiotemporal partial observability of multi-agent reinforcement learning cannot guarantee stable convergence. To this end, a multi-intersection traffic signal control algorithm based on overall state prediction and fair experience replay (IS-DQN) was proposed. On the one hand, to avoid the problem of non-stationarity caused by spatial observability in algorithm, the state space of IS-DQN was expanded by predicting the overall state of multiple traffic intersections based on historical traffic flow information from different lanes. On the other hand, in order to cope with the time partial observability brought about by traditional experience replay strategies, IS-DQN adopted a reservoir sampling algorithm to ensure the fairness of experience replay pool and avoid non-stationary problems it brings. Three different traffic pressure experiments were conducted using IS-DQN in complex multi-intersection environments. Simulation experiments under three different traffic pressure were conducted in complex multi-intersection environments. Results shows that under different traffic pressure conditions, especially in low and medium traffic pressure, IS-DQN algorithm showed lower average vehicle travel time, better convergence performance and stability compared to independent deep reinforcement learning algorithms.

Keywords: deep reinforcement learning; traffic signal control; time series prediction; reservoir sampling algorithm; Long Short-Term Memory (LSTM) network

收稿日期：2024-01-19； **修回日期**：2024-03-15； **录用日期**：2024-03-25。

基金项目：国家自然科学基金面上项目(62276097)；上海市自然科学基金资助项目（22ZR1416500，23ZR1414900）；上海市基础研究特区计划（22TQ1400100-16）。

作者简介： 缪孜珺(1999—)，男，浙江宁波人，硕士研究生，主要研究方向：强化学习； 罗飞(1978—)，男，湖北武汉人，副教授，博士，CCF 会员，主要研究方向：认知计算、强化学习； 丁炜超(1989—)，男，山东青岛人，副教授，博士，CCF 会员，主要研究方向：云计算、群智计算、联邦学习； 董文波（1992—），男，河南新乡人，讲师，博士研究生，CCF 会员，主要研究方向：机器学习、人工智能。

0 引言

随着市民汽车保有量的增加和城市化进程的加剧，交通拥堵问题日益突出，严重影响了市民的生活质量和城市的可持续发展。因此，优化交通信号控制方法，实现智能化、精细化的交通管理，进而提高道路通行效率，对解决城市交通拥堵、减少公路运输尾气排放具有重要的意义[1]。

交通信号控制方法从固定配时控制(Fixed-time Control)、感应式控制(Vehicle Actuated)发展到实时自适应控制(Real-time Adaptive Traffic Control)[2]。其中，固定配时控制方法有固定的信号变换周期。感应式控制方法通过使用道路检测器的实时测量结果来优化信号计时[3]，因其实现了较优的交通流量控制效果，该方法已被纳入一些商业交通信号控制系统。然而，该方法受到其模型预先定义参数的限制，在波动的交通需求中，以强化学习算法为代表的自适应控制系统是一种更有效的解决方案[2]。最新的研究表明：将深度强化学习算法--深度 Q 学习(Deep Q-Network, DQN)算法应用于交通信号控制问题，在解决交通拥堵方面，DQN 的性能显著优于感应式控制方法[4-8]。

尽管单智能体的强化学习算法在单路口的交通信号控制问题上已经有所建树，但如果想要在现实中的复杂多路口交通环境下应用，则需要使用多智能体强化学习算法(MARL)。MARL 基于已经熟悉的单智能体算法来进行学习，这主要分为两种思路：完全中心化方法和去中心化方法[9]。完全中心化方法将多个智能体的决策视为一个超级智能体的决策；然而，随着交通网络的复杂度增加，联合 Q 函数策略的复杂度也呈指数级增长。去中心化方法则不考虑其他智能体的变化，每个智能体都有独立的 Q 函数，并且只根据自己观察到的环境进行独立学习；在大规模城市交通网络下，去中心化的独立强化学习算法表现出更好的适应性，这也是本文关注的重点。

然而，去中心化的独立强化学习算法在交通信号控制问题上的应用会引入一个非平稳性问题：独立智能体观测到的环境包含其他正在学习的智能体，因此其所假设的收敛目标会不断变化，从而导致算法无法平稳收敛。为了解决这个问题，一部分研究人员通过图卷积网络或注意力机制学习智能体之间的依赖性[10-12]，但他们忽略了状态信息的时空相关性；另一部分则侧重于聚合来自其他智能体的状态信息[13-15]，但不幸的是在实际应用中畅通的通信环境难以实现。Fang 等[16]则利用注意力机制捕捉交叉口的时空依赖性，改进了算法的收敛性与可解释性。该算法采用集中式训练，在大规模交通网络中难以应用，但复杂多交叉口之间的时空依赖性值得本文进行进一步探讨。

因此，针对独立强化学习在多路口交通信号控制中所带来的非平稳性问题，同时从算法的状态信息获取与经验重放两方面对 DQN 进行改进，从而提出一种面向多路口交通信号控制的基于全局状态预测与公平经验重放的独立稳态深度强化学习算法(IS-DQN)，以提升算法的性能与稳定性。本文的主要贡献如下：

一方面，提出了一种基于自注意力机制和长短期记忆神经网络(LSTM)的全局交通状态预测网络，利用该网络预测多交通路口的全局状态，从而为 IS-DQN 设计了基于全局状态预测的增广状态空间，以解决多智能体强化学习算法因其空间部分可观测性而导致的非平稳性问题。

另一方面，提出了一种基于蓄水池算法的经验池重放机制，避免了多路口交通控制算法因其时间部分可观测性而导致的非平稳问题，从而提升算法的鲁棒性与收敛稳定性。

在真实路网的低中高三种不同交通流量条件下进行了实验。实验结果表明：IS-DQN 算法相较于独立深度强化学习算法，车辆平均行驶时间减少了 8.15%、9.59%、0.85%都，并表现出了更优的收敛稳定性，有效提升了车辆通行效率。

1 交通信号控制问题中的非平稳性

多路口交通信号控制问题一般会被描述为一个马尔可夫决策过程(Markov Decision Process, MDP)，定义为元组 $<\boldsymbol{S},\boldsymbol{A},\boldsymbol{P},\boldsymbol{R},\gamma>$。在多交通路口环境下(Environment)将每个路口设定为一个智能体(IS-DQN agent)。在该条件下，$\boldsymbol{S}$ 代表各个路口所观测到的交通流状态(State)集合；$\boldsymbol{A}$ 代表各个路口所作出的动作(Action)集合；$\boldsymbol{P}$ 代表不同状态动作下的状态转移概率；$\boldsymbol{R}$ 代表奖励函数(Reward)。当环境处于时间步长 t 下时，每个智能体会通过构建的策略 π 与观测所得的当前交通状态 s_t 对当前时间步的信号灯变换动作进行选择与执行，然后系统转换到下一个状态 s_{t+1}。在时间步 $t+1$，智能体再次观测环境根据定义的奖励函数获取奖励 r，并根据该奖励对策略 π 进行更新。图 1 描绘了交通信号控制问题的 MDP 过程。

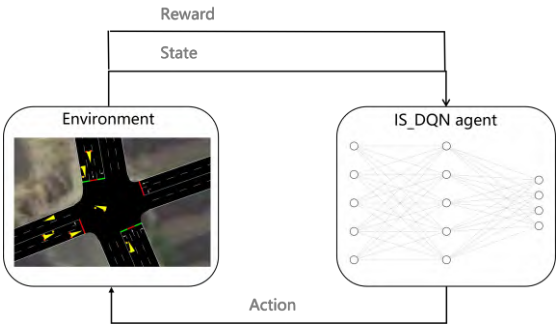

图 1 交通信号控制问题 MDP 过程示意图

Fig. 1 Picture of MDP process for traffic signal control problem

为了进一步分析多智能体环境中的非平稳性，首先假设一个完全可观察的多智能体环境。在 m 个智能体中选取一个智能体 n (其中 $n \in N \equiv \{1...m\}$)，使用下标 $-n$ 来表示所有其

他的智能体，如 $\boldsymbol{A}=[a_n,a_{-n}]$。在每个时间步中，智能体 n 根据其策略 π_n 选取动作 $a_n\in\boldsymbol{A}$，对于下一状态 s' 的转移概率为 $P(s'\mid s_t,\boldsymbol{A})$。在这一假设下，智能体 n 的最优动作价值函数 Q_n^* 对于所有其他智能体可能的动作 a_{-n} 都是已知的，考虑到所有其他智能体的联合策略 π_{-n}。单智能体的迭代公式如式(1)所示：

$$Q(s_t,a_t)\leftarrow(1-\alpha)Q(s_t,a_t)+\alpha\left[r_{t+1}+\gamma\max_a Q(s',a)\right] \tag{1}$$

可以写出在多智能体强化学习系统中，智能体 n 的 Bellman 最优方程如式(2)所示：

$$Q_n^*(s_t,a_n\mid\pi_{-n})=\sum_{A_{-n}}\pi_{-n}(a_{-n}\mid s_t)\left[r+\gamma\sum_{s'}P(s'\mid s_t,a_n,a_{-n})\max_{a'_n}Q_n^*(s',a'_n)\right] \tag{2}$$

从公式 2 中可看出，随着其他智能体在学习中改变策略非平稳因子 $\pi_{-n}(a_{-n}\mid s)=\Pi_{i\in-n}\pi_i(a_i\mid s)$ 也会随之发生改变。这意味着，算法的收敛过程会随着该非平稳因子的变化而产生波动。为了获取其他智能体的状态变化对全局策略进行观察是一个选择，但在实践中这样的假设可能过于严格。智能体对其周围环境的观测可能有限，与其他智能体的沟通可能不可行或不可靠[17]，其他智能体所观察获取的状态可能不可用[18]。在这种情况下，智能体必须只使用自己观察可用的信息进行推理。因此可以得出结论：在类似多路口交通信号控制问题的多智能体强化学习算法中，非平稳性的主要来源是智能体对于当前真实环境的时空部分可观测性与对于其他智能体所采取动作或者策略的不可知性[19]。

针对单路口交通信号控制，为了避免环境的部分可观测性，深度强化学习算法的一个重要解决方案是采用经验重放机制。经验重放的工作机制是在训练神经网络时，从经验重放池中均匀地批量采样经验，而不是仅仅使用最新的经验。这种方法的优势主要有两个方面。首先，通过随机采样，可以避免网络过度拟合最近的经验，从而提高了样本的多样性和利用效率。这种随机性有助于网络学习到更通用的模式，而不是特定于最近经验的模式。其次，DQN 智能体不是直接在整体的观察样本上进行学习，而是在小批量样本上进行学习。这种小批量学习的方式，相较于全体样本学习，大幅提高了训练的效率和速度，使训练过程更高效和灵活。

然而，针对多路口交通信号控制算法，直接采用经验重放机制则可能会进一步导致深度强化学习算法对于环境认知的非平稳性：每个路口智能体中的重放经验无法反映当前学习的动态性。Schaul 等[20]的研究表明在选择样本进行训练时，经验选择可以提高算法训练效果。然而，这些研究并没有考虑如何作出更持久的决策，即选择哪些样本要保留，哪些样本要丢弃。也就是说经验重放是巩固学习不同状态下最优策略的必要组成部分。通过保留过去的经验，并将之融入学习过程，可以有效地克服按时间顺序进行训练时对早期记忆的遗忘。因此，IS-DQN 期望一个与全局分布相匹配的经验分布从而避免算法的部分可观测性。

2 算法设计

2.1 基于全局状态预测的增广状态空间

在多路口交通信号控制问题中，不同路口的行动-状态关系以一种复杂的方式相互关联。如果想要通过一些计算的方式推算全局状态避免部分可观测性，则需要对每个路口的观察状态与选择策略进行推断，而这很难实现。然而，在多路口交通环境中，路口之间的状态并非毫无关联。对于单个路口，可以通过分析当前和历史状态下不同方向车道的交通流量长度，对相邻路口的状态进行一定程度的推断。

研究表明，相比直接预测其他智能体的策略，将其他智能体的混合策略估计包含在状态中，让 Q 函数对整个混合策略进行评估，是一个更优的选择[21]。因此，基于路口对全局状态的推断构建了基于自注意力机制和 LSTM 的交通状态预测网络，并设计了新的增广状态空间与混合策略从而避免智能体部分可观测性导致的非平稳性。

在每个时间步下，IS-DQN 算法将当前与历史状态输入预测网络最后获得预测全局状态 S，将预测全局状态与当前状态扩展为增广状态后输入 DQN 网络。

预测网络以 LSTM 单元即双层的 LSTM 隐含层为核心，添加注意力机制，并设计输入输出向量格式。隐含层状态的传递给该网络带来了优良的时序预测性能，而添加的注意力机制可以动态捕获复杂动脉网络上的时空依赖关系[22]。整体预测网络结构如下图 2 所示：

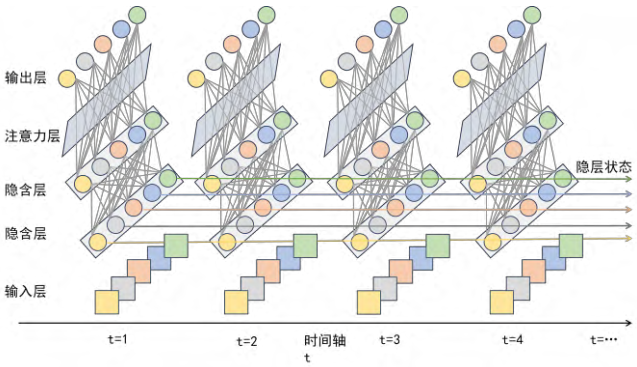

图 2 IS-DQN 中预测网络结构

Fig. 2 Prediction Network Structure in IS-DQN

基于全局状态预测的增广状态空间计算流程如下所示：

预测网络的输入以路口从环境所获取的状态信息为基准，并规范化为当前时间步下的相位索引 p 与各车道等待车辆数量 $\boldsymbol{Cnt}$。输入向量公式如式(3)所示：

$$Input(t)=<p(t),\left[\boldsymbol{Cnt}^1(t),...,\boldsymbol{Cnt}^k(t)\right]> \tag{3}$$

LSTM 单元通过其特殊的门控机制学习提取输入向量中的时序关联信息。在时间步 t 下，隐含层中的 LSTM 单元输

入时间步 t-1 下的输出信息 h_{t-1}、细胞状态 C_{t-1} 与输入向量 $Input(t)$。LSTM 单元运算公式如式(4)-(8)所示：

$$i_t = \sigma\left(W_{\mathrm{i}} \cdot \left[h_{t-1}, Input(t)\right] + b_{\mathrm{i}}\right) \tag{4}$$

$$o_t = \sigma\left(W_{\mathrm{o}} \cdot \left[h_{t-1}, Input(t)\right] + b_{\mathrm{o}}\right) \tag{5}$$

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, Input(t)\right] + b_f\right) \tag{6}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot tanh\left(W_{\mathrm{c}} \cdot \left[h_{t-1}, Input(t)\right] + b_{\mathrm{c}}\right) \tag{7}$$

$$h_t = o_t \cdot tanh\left(C_t\right) \tag{8}$$

其中，$\sigma(\ )$为 Sigmoid 激活函数，$tanh(\ )$为双曲正切函数，W 为训练权重，b 为偏置项参数。下标 i、o、f、c 为 LSTM 单元中的输入门、输出门、遗忘门与细胞状态，h_t 为当前时间步下 LSTM 单元的输出。

注意力层计算输入数据与上下文的相似度从而让模型捕捉输入数据的关键信息。注意力层的输入为隐含层中 LSTM 单元的输出 h_t，使用基于感知机的相似度函数 $f(.)$ 计算获得单个输入与整体输入的匹配权重 a_t 如式(9)所示：

$$a_t = \frac{exp(f(b_t, h_t))}{\sum_n exp(f(b_t, h_i))} \tag{9}$$

然后将注意力权重与每一个输入进行点乘累加得到注意力层的输出 $Attention(t)$，如式(10)所示：

$$Attention(t) = \sum_T a_t h_t \tag{10}$$

输出层通过 Softmax 函数将注意力层的输出归一化为输出向量 $Output(t)$，即预测全局状态 S，如式(11)所示：

$$S = Output(t) = Soft\,max\left(Attention(t)\right) \tag{11}$$

考虑到实验过程中对全局状态直接进行预测的效果并不佳，因此综合实验结果考虑选择了与路口状态关联程度较高的不同车道方向下拥堵程度作为预测全局状态进行输出。在每个时间步 t 下，假设共有 k 个车道，该预测模型的输出向量即预测全局状态如式(12)所示：

$$\tilde{S} = Output(t) = <Qp_1(t), Qp_2(t), ..., Qp_k(t)> \tag{12}$$

最后，将当前状态 s_t 与全局状态 S 组合获得基于全局状态预测的增广状态空间 $S_T = <s_t, S>$ 输入 DQN 网络。

2.2 基于蓄水池抽样算法的经验选择

正如第一节中所分析的那样，IS-DQN 的经验选择算法期望能够保留下不同类型的经验样本，避免带来时间上的部分可观测性，从而使智能体能够对不同状态空间下的经验进行公平地学习。目前最常用也是最简单的方法是先进先出算法(FIFO)：在经验重放池填满后，让新的样本不断取代最旧样本。

然而，这种算法不可避免地会引入时间相关的偏见，并可能导致样本多样性不足。此外，如果使用大容量的经验重放池来增加样本多样性，将对存储空间提出更高要求，并降低算法训练的效率。因此，提出使用蓄水池抽样算法进行经验样本池更新，从而在经验顺序到达的情况下保证样本的随机性。基于蓄水池抽样的经验选择流程如算法 1 所示：

算法 **1**　　基于蓄水池抽样的经验选择算法

输入：样本池 Sp，重播样本数 k

输出：经验重放池 Rp

```
for i in Sp(n):
    if i<k:
        Rp[i] = Sp[i]      # 前 k 个直接保留
    else:
        random_int = random (0, i)  # 获取随机数
        if random_int < k:
            Rp[random_int] = Sp[i]
        end if
    end if
end for
```

以上流程可解释为：

- 第 $i \in \{1,...,k\}$ 个样本按照先后次序直接放入到重播样本池中；
- 第 $j \in \{k+1,...,n\}$ 个样本，每次先以概率 $p = \frac{k}{j}$ 选择是否让该样本留下，若能够留下则以等概率从重放样本池中选择一条进行替换。

基于以上算法流程，提出并证明以下定理：

定理：所有顺序抵达的经验样本均以等概率即 $\frac{\mathrm{k}}{\mathrm{n}}$ 的概率被保存进入经验重放池。

证明：第 $i \in \{1,...,k\}$ 个样本最终被保存在经验重放池的概率 $P(\tilde{i})$ 为被该经验样本被保存的概率 $p(i)$ 乘以不被第 $j \in \{k+1,...,n\}$ 个样本替换的概率。概率公式如式(13)所示：

$$P(\tilde{i}) = p(i) \cdot p(\bar{i} \mid k+1) \cdot p(\bar{i} \mid k+2) ... p(\bar{i} \mid n) \tag{13}$$

对于第 $i \in \{1,...,k\}$ 个样本被选中的概率为 1。而第 k+1 个样本被选中的概率为 $\frac{k}{k+1}$，则第 i 个样本被第 k+1 个元素替换的概率如式(14)所示：

$$p(i \mid k+1) = \frac{k}{k+1} * \frac{1}{k} = \frac{1}{k+1} \tag{14}$$

以此类推则可以得到，在训练结束即索引位置为 n 时，第 i 个样本被保存在经验重放池的概率如式(15)所示：

$$P(\tilde{i}) = 1 \times \frac{\mathrm{k}}{\mathrm{k}+1} \times \frac{\mathrm{k}+1}{\mathrm{k}+2} \times \frac{\mathrm{k}+2}{\mathrm{k}+3} \times ... \times \frac{\mathrm{n}-1}{\mathrm{n}} = \frac{\mathrm{k}}{\mathrm{n}} \tag{15}$$

对于第 $j\in\{k+1,...,n\}$ 个样本，最终被保存在经验重放池的概率 $P(\tilde{j})$ 为该样本被选中保存的概率 $p(j)$ 乘以不被之后样本替换的概率。概率公式如式(16)所示：

$$P(\tilde{j})=p(j)\cdot p(\bar{j}\mid j+1)\cdot p(\bar{j}\mid j+2)...p(\bar{j}\mid n) \tag{16}$$

第 $j\in\{k+1,...,n\}$ 个样本被选中的概率为 $\frac{k}{j}$。而第 j+1 个样本被选中的概率为 $\frac{k}{j+1}$，则第 j 个样本被第 j+1 个样本替换的概率如以式(17)所示：

$$p(\bar{j}\mid j+1)=\frac{k}{j+1}*\frac{1}{k}=\frac{1}{j+1} \tag{17}$$

以此类推则可以得到，在训练结束即索引位置为 n 时，第 j 个样本被保存在经验重放池的概率如式(18)所示：

$$P(\tilde{j})=\frac{k}{j}\times\frac{j}{j+1}\times\frac{j+1}{j+2}\times\frac{j+2}{j+3}\times...\times\frac{n-1}{n}=\frac{k}{n} \tag{18}$$

因此可以证明所有经验样本均以等概率被保存，即经验重放池的样本是与样本池中近似匹配的小种群，进而避免了算法的部分可观测性。

2.3 IS-DQN 算法

基于以上拓展状态空间与经验选择算法，本文提出 IS-DQN 算法。该算法整体流程如算法 2 所示：

算法 **2** **IS-DQN** 算法
输入：多路口测试仿真环境
输出：各回合测试仿真结果

1 初始化 Q 网络并设置 $\hat{Q}=Q$
2 **for** 每个训练回合：
3 | **for** 每个时间步：
4 | | 将环境状态输入预测网络并计算获得 $S_T=<s_t,S>$
5 | | 以 ε 的概率选择随机动作 a_t 或 $a_t=arg\,max_a\,Q(S',a;\theta)$
6 | | 执行动作 a，环境观测获得奖励 r，得到下一个状态 s_{t+1}
7 | | 存储经验元组 $\langle s,u_a,r,\varphi,s_h,S'\rangle$ 进入样本池 Sp
8 | | 从全局样本池中更新重播样本池
9 | | 更新 $y=\begin{cases}r & \text{if done}\\ r+\gamma\,max_{a'}\hat{Q}(s_t,a_t;\bar{\theta}) & \text{otherwise}\end{cases}$
10 | | 根据重放池更新 DQN 网络，根据样本池更新预测网络
11 | | 每 C 步使 $\hat{Q}=Q$
12 | **end for**
13 **end for**

整个 IS-DQN 算法的流程与采用经验池的 DQN 算法基本一致。在每个时间步中，步骤 4 至 7 为采样阶段，主要过程为从环境获取信息与并通过预测网络预测全局状态并组合为增广状态空间，并将采样结果存储进入经验重放池 D；步骤 8 至 11 为训练阶段，每隔 C 个时间步，算法从存储的经验中通过蓄水池抽样算法获取一个小的抽样样本给予 DQN 网络进行训练。步骤 4 和步骤 8 为前述分别为 2.1 的增广状态空间获取和 2.2 的经验选择算法，但由于预测与训练的需要，IS-DQN 算法整体交互流程如图 3 所示。

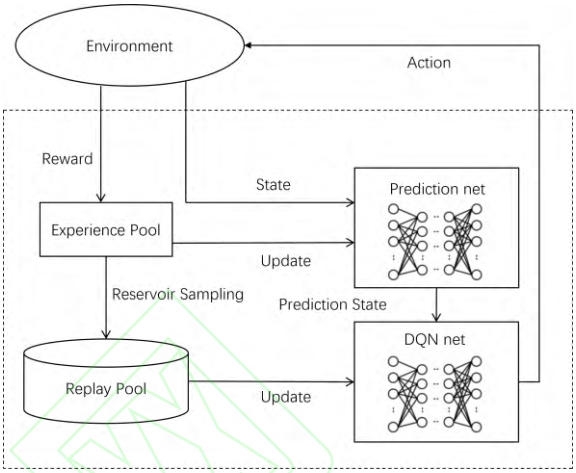

图 3 IS-DQN 整体交互流程

Fig. 3 Overall interactive flowchart of IS-DQN

如上图所示，状态空间会先经过预测网络处理并计算获得基于全局状态预测的增广状态空间后再输入 DQN 网络。两个网络所需要的训练样本并不一致，因此预测网络与 DQN 网络会分别在样本池与经验重放池中进行更新。需要注意的是，在第一轮训练结束后才会对预测网络进行初始化，因此步骤 4 中的预测过程在第一轮中会跳过。最后，经过 150 回合与环境的交互训练可以得到该算法在环境中每一轮的仿真结果。

3 实验与结果分析

3.1 实验环境

本研究使用开源交通模拟平台 CityFlow 开展实验。CityFlow 是专为大规模交通信号控制设计的平台，可以支持基于合成和真实数据的道路网络和交通流的灵活定义，并且 CityFlow 优秀的计算速度适用于大规模信号灯路口的模拟[23]。在实验过程中，首先根据提供的数据集信息配置了 CityFlow 平台的环境，包括路口布局、路网结构和交通流量等。然后，将实际交通数据输入配置好的模拟环境中，模拟车辆在该环境中行驶。接下来，根据交通状况，IS-DQN 向 CityFlow 平台发送信号控制指令，包括调整信号灯的变化周期和模式等。CityFlow 平台接收并执行这些指令后，将新的交通状态反馈给 IS-DQN 算法。通过这种方式，在一个高度控制和可再现的环境中，可以对 IS-DQN 进行全面和深入的测试和优化。

本实验选取了一个复杂多路口交通环境，在 3600s 的仿真时长下，分别使用随机种子生成 2100、2700、3300 辆不同

的车辆以随机路径进入路网，分别对应为低、中、高三种车流量条件。对于每种流量条件，用随机种子生成 5 组车流数据，最后取这 5 组车流数据下 10 次实验结果的平均值进行比较。该多路口交通环境仿真图如图 4 所示，六个路口的编号在图 4 中依次标出。所有交叉口的道路网络设置与现实中一致。文献[24]发现模型性能对不同交通信号控制方法与环境之间的时间步长这一参数并不敏感，因此每个时间步长设置为 10 秒。

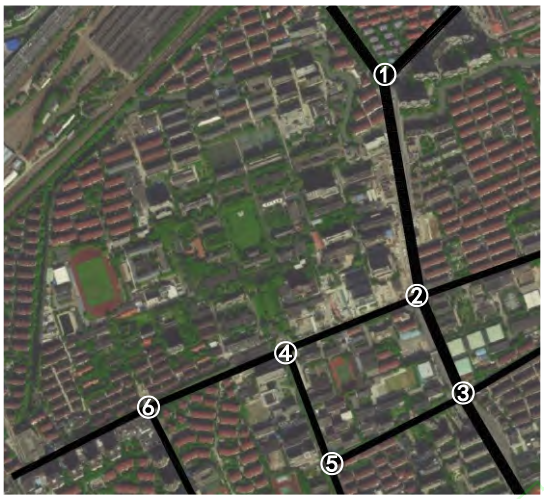

图 4 多路口测试环境仿真图

Fig. 4 Simulation diagram of multi intersection testing environment

3.2 对比算法

为了评估 IS-DQN 的有效性与效率，除了与 DQN 算法进行比较之外，还将它们与以下已被广泛应用的经典交通控制方法进行比较。

- 固定配时算法(FIXTIME)：固定配时算法采用预先确定的相位周期与顺序，是目前使用最广泛的交通信号控制方法。
- 自组织交通灯控制(SOTL)[25]：自组织交通灯控制算法是一种基于动态调节车辆等待数量阈值的自适应交通信号控制方法。由于该算法具备调节能力，因此在众多城市中得到了广泛应用。
- 最大压力算法(MaxPressure)[26]:MaxPressure 算法引入了压力的概念——上游排队车辆与下游排队车辆的数量差异。MaxPressure 计算每个阶段的压力，并进行比较，最终激活压力最大的阶段。该方法能够有效地调控交通拥堵，具备经过数学证明的吞吐量特性。

此外，本文还设置了相应的消融实验算法：

- DQN-PS：仅添加了基于全局状态预测的增广状态空间的 DQN 算法。
- DQN-ER：使用基于蓄水池抽样的经验选择算法的 DQN 算法。

除以上算法外，在实验过程中也添加了基于递归神经网络与 DQN 算法的 DRQN 算法作为对比。DRQN 算法在 DQN 网络中添加 LSTM 层来保留并利用先前状态信息，从而能训练出具有长期记忆能力的智能体。但实验结果表明其无法保证收敛，因此不在实验结果中列举。

3.3 参数设置

为了提高学习性能，算法中的参数都通过实验调整为了最优。经典交通信号控制算法的超参数设置如表 1 所示：

表 1 经典算法的超参数设置

Tab. 1 Hyperparameter setting of classical timing algorithms

算法名称	超参数	值	含义
FIXTIME	C	80	相位周期
MAXPRESSURE	$\varphi^{\min}$	5	最小绿灯时间
SOTL	$\varphi^{\min}$	2	最小绿灯时间
SOTL	θ	4	车辆数阈值
SOTL	μ	28	绿灯车辆数阈值

在所有的训练和测试过程中，IS-DQN 算法与 DQN 算法中的 DQN 网络都采用了学习率为 0.0005 的 Adam 优化器；最小经验池大小为 600；最大经验池大小为 50000；折扣因子为 0.8。在预测网络中，采用学习率为 0.001 的 Adam 优化器，隐藏层数为 2，隐藏层节点个数为 128，采用二值交叉熵作为损失函数。IS-DQN 算法共迭代 150 轮。

3.4 实验结果

根据对现有交通信号控制的研究，实验选择了汽车行驶时间作为代表性的指标。在交通领域，这也是判断性能最常用的衡量标准。该指标以秒为单位，同时考虑了汽车行驶抵达路口所需时间和汽车在路口等待的时间。因此，更优秀的算法在相同车流量数据集下能够得到更低的平均行驶时间，代表该算法能够有效提升车辆通行效率。此外，累积奖励是强化学习的基本性能指标，实验也将其纳入了对比。在本实验中，奖励值设定为车辆停止时间的负数。

首先对不同算法在不同交通状况下优化后的平均行驶时间进行比较，实验结果如表 2 所示，其中深度强化学习算法选取 150 轮迭代中的最后 10 轮平均值作为结果进行比较。

从表中可以看出在所有交通流情况下，DQN 和 IS-DQN 算法相较于传统方法都有着显著的优势，这在低交通流量下尤其明显。而改进后的算法相较于 DQN 算法在大多数情况下都能保证更优的收敛结果。在最终结果上，IS-DQN 相较于 DQN 在低中高的流量情况下各路口的平均行驶时间之和上分别有 8.15%、9.59%、0.85%的优势。

表 2 不同算法在不同交通压力下优化后的平均行驶时间

Tab. 2 The average travel time optimized by different algorithms under different traffic pressure

	IS-DQN	DQN	DQN-PS	DQN-ER	Fixtime	MaxPressure	SOTL
低流量路口 1	**92.520**	103.638	93.750	98.904	285.387	143.212	148.432
低流量路口 2	**85.072**	93.101	85.771	89.633	279.508	126.871	128.802
低流量路口 3	**73.975**	83.960	74.444	82.593	272.065	109.183	109.260
低流量路口 4	**94.085**	99.426	94.440	98.913	400.112	152.070	162.543
低流量路口 5	95.291	99.374	**94.877**	96.614	349.340	140.418	149.679
低流量路口 6	**74.085**	81.204	74.266	80.416	345.519	118.796	146.500
中流量路口 1	**112.058**	121.691	114.467	117.215	244.768	156.211	163.154
中流量路口 2	**106.165**	116.094	108.171	109.196	216.307	141.574	153.677
中流量路口 3	**85.988**	94.483	88.368	96.213	225.756	117.185	165.453
中流量路口 4	**124.960**	141.970	126.198	135.349	289.689	171.429	204.084
中流量路口 5	**108.599**	113.521	110.960	115.294	286.784	143.639	178.838
中流量路口 6	**91.078**	107.781	92.318	100.988	258.487	135.075	206.624
高流量路口 1	129.778	128.808	**127.764**	131.759	276.707	166.963	146.500
高流量路口 2	125.551	**123.464**	124.294	131.589	208.357	149.933	162.850
高流量路口 3	**107.029**	109.390	114.653	114.301	201.837	124.199	173.341
高流量路口 4	164.431	**161.762**	161.873	166.006	336.926	188.821	200.651
高流量路口 5	118.135	**117.691**	121.091	121.621	388.053	148.797	163.518
高流量路口 6	130.814	141.184	**130.136**	133.617	242.165	145.862	197.414

高交通流量下，所有路口都处于且长时间处于拥堵状态，这反而降低了环境之间的非平稳性，这导致了 IS-DQN 与 DQN-PS 算法失去了中低交通流量下的显著优势。尽管高交通流量下 DQN 算法在多数路口上有一些优势，但在路口 6 上相较于 IS-DQN 却有一个明显地劣势，体现出 IS-DQN 算法所预测全局状态能够更兼顾不同路口拥堵的公平性。此外，由于固定配时算法在三种流量条件下都处于拥堵状态，因此其性能表现与车辆随机路径具有更高的相关性。

为了更进一步分析算法的有效性，实验对各个算法在不同交通流的 150 轮迭代过程进行了分析。其中的传统交通信号控制算法如 fixtime 等都没有迭代过程，以常函数进行表示，整体迭代流程如图 5~7 所示。

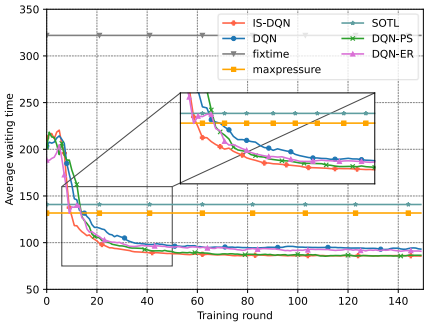

图 5 各算法在低交通流量下收敛过程

Fig. 5 The convergence process of each algorithm under low traffic flow

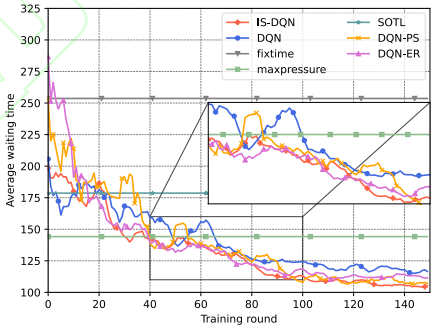

图 6 各算法在中交通流量下收敛过程

Fig. 6 The convergence process of each algorithm under medium traffic flow

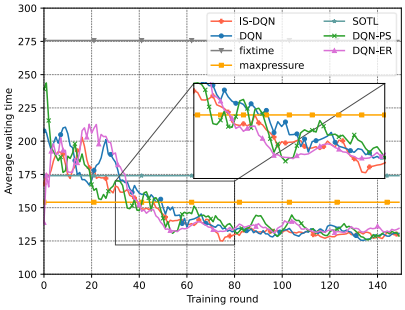

图 7 各算法在高交通流量下收敛过程

Fig. 7 Convergence process of each algorithm under high traffic flow

从图 5~7 中可以看出，DQN 与 IS-DQN 算法在各个交通流情况下至多 50 轮迭代就能够拥有超越传统算法的性能。而正如所预料的那样，IS-DQN 算法相较于 DQN 算法在多数

情况下都表现出了良好的性能，尤其是在低交通流情况下。数据显示改进后的 IS-DQN 算法在整体收敛过程中低中高三种交通流情况下 150 个收敛回合的平均行驶时间分别减少了 9.17%、6.58%、2.63%。

在表 1 中低交通流量路口中，应用了增广状态空间的算法 IS-DQN 与 DQN-PS 有着明显更优的收敛结果。在图 5~7 放大的收敛区域中可以明显看出 IS-DQN 与 DQN-ER 算法能够更快更稳定地达到收敛。因此，综合图 5~7 与表 1 的结果可以得出，通过基于状态预测的增广状态空间能够给算法带来更好的收敛结果，而使用基于蓄水池抽样的经验选择算法能够给算法带来更稳定快速的收敛过程。

累计奖励是强化学习算法重要的性能判断指标之一。在六个交通路口上，最具代表性的中等交通流量情况下，DQN 算法和改进后的 IS-DQN 算法的累计奖励如图 8 所示。显然，改进后的 IS-DQN 算法在累计奖励方面无论是数值还是稳定性的表现都更出色。结合图 5~7 的收敛过程，可以得出结论：在中低流量情况下，改进后的 IS-DQN 算法相比 DQN 算法能够处理交通信号控制问题的非平稳性，从而有效学习交通环境状态、提升车辆通行效率，获得了更好的收敛性能和收敛结果。

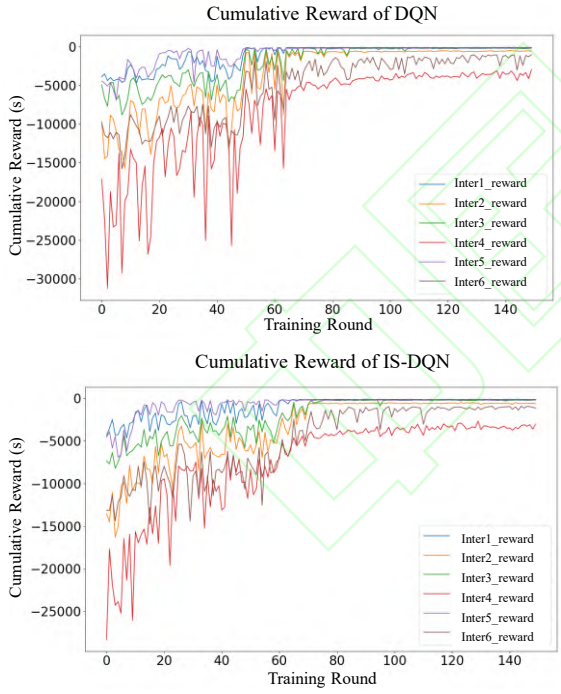

图 8 DQN 与 IS-DQN 算法在中交通流量下的累计奖励

Fig. 8 Cumulative Reward of DQN and IS-DQN by episode of medium traffic flow

4 结语

本文提出了使用当前成熟的交通预测算法设计新的增广状态的方法来解决多智能体强化学习中的非平稳性问题。同时，为了避免经验重放中的偏见，IS-DQN 算法采用了蓄水池算法确保经验重放池的公正性。通过在复杂多路口环境下进行三种不同交通压力的实验，证明了本文所设计的增广状态空间与样本抽样算法能够有效地避免部分可观测性从而提升路口车辆通行效率。尤其是在中低车流量下，改进后的 IS-DQN 算法更能保证算法的收敛结果。尽管在动态交通环境下的性能有待进一步验证，但这已表明 IS-DQN 算法在应对多路口环境下的交通信号控制问题方面的现实应用具有很大的潜力。未来的研究将进一步探索结合天气节假日等条件的交通预测方法在现实交通系统中的应用，建立更全面的拟真环境，并对其性能进行更深入的评估和优化。

参考文献

[1] KÓVÁRI B, KOLAT M, BÉCSI T, et al. Competitive multi-agent reinforcement learning for traffic signal control [C]// Proceedings of the 2022 IEEE 20th Jubilee International Symposium on Intelligent Systems and Informatics (SISY). IEEE, 2022: 361-366.

[2] NOAEEM M, NAIK A, GOODMAN L, et al. Reinforcement learning in urban network traffic signal control: a systematic literature review[J]. Expert Systems with Applications, 2022, 199: 116830.

[3] MA D, XIAO J, SONG X, et al. A back-pressure-based model with fixed phase sequences for traffic signal optimization under oversaturated networks[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(9): 5577-5588.

[4] DUCROCQ R, FARHI N. Deep reinforcement Q-learning for intelligent traffic signal control with partial detection[J]. International Journal of Intelligent Transportation Systems Research, 2023, 21(1): 192-206.

[5] HAN G, LIU X, WANG H, et al. An attention reinforcement learning-based strategy for large-scale adaptive traffic signal control system[J]. Journal of Transportation Engineering, Part A: Systems, 2024, 150(3): 04024001.

[6] YAZDANI M, SARVI M, BAGLOEE S A, et al. Intelligent vehicle pedestrian light (IVPL): a deep reinforcement learning approach for traffic signal control[J]. Transportation research part C: emerging technologies, 2023, 149: 103991.

[7] ZHU R, LI L, WU S, et al. Multi-agent broad reinforcement learning for intelligent traffic light control[J]. Information Sciences, 2023, 619: 509-525.

[8] KOLAT M, KŐVÁRI B, BÉCSI T, et al. Multi-agent reinforcement learning for traffic signal control:a cooperative approach[J]. Sustainability, 2023, 15(4): 3479.

[9] ZHANG K, YANG Z, BAŞAR T. Multi-agent reinforcement learning: A selective overview of theories and algorithms[J]. Handbook of reinforcement learning and control, 2021: 321-384.

[10] YANG S. Hierarchical graph multi-agent reinforcement learning for traffic signal control[J]. Information Sciences, 2023, 634: 55-72.

[11] YANG S, YANG B. An inductive heterogeneous graph attention-based multi-agent deep graph infomax algorithm for adaptive traffic signal control[J]. Information fusion, 2022, 88: 249-262.

[12] ZHAO Z, WANG K, WANG Y, et al. Enhancing traffic signal control with composite deep intelligence[J]. Expert Systems with Applications, 2024, 244: 123020.

[13] GUO J, CHENG L, WANG S. Cotv: cooperative control for traffic light signals and connected autonomous vehicles using deep reinforcement learning[J]. IEEE Transactions on Intelligent Transportation Systems, 2023.

[14] REN F, DONG W, ZHAO X, et al. Two-layer coordinated reinforcement learning for traffic signal control in traffic network[J]. Expert Systems with Applications, 2024, 235: 121111.

[15] BOKADE R, JIN X, AMATO C. Multi-agent reinforcement learning based on representational communication for large-scale traffic signal control[J]. IEEE Access, 2023.
[16] FANG J, YOU Y, XU M, et al. Multi-objective traffic signal control using network-wide agent coordinated reinforcement learning[J]. Expert Systems with Applications, 2023: 120535.
[17] STONE P, KAMINKA G, KRAUS S, et al. Ad hoc autonomous agent teams: collaboration without pre-coordination [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2010, 24(1): 1504-1509.
[18] GMYTRASIEWICZ P J, DOSHI P. A framework for sequential planning in multi-agent settings[J]. Journal of Artificial Intelligence Research, 2005, 24: 49-79.
[19] HERNANDEZ-LEAL P, KAISERS M, BAARSLAG T, et al. A survey of learning in multiagent environments: dealing with non-stationarity[J]. arXiv preprint arXiv:1707.09183, 2017.
[20] SCHAUL T, QUAN J, ANTONOGLOU I, et al. Prioritized experience replay[J]. arXiv preprint arXiv:1511.05952, 2015.
[21] TESAURO G. Extending Q-learning to general adaptive multi-agent systems[J]. Advances in neural information processing systems, 2003, 16.
[22] ABBASIMEHR H, PAKI R. Improving time series forecasting using LSTM and attention models[J]. Journal of Ambient Intelligence and Humanized Computing, 2022: 1-19.
[23] TANG Z, NAPHADE M, LIU M Y, et al. Cityflow: a city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 8797-8806.
[24] ZHENG G, XIONG Y, ZANG X, et al. Learning phase competition for traffic signal control [C]// Proceedings of the 28th ACM international conference on information and knowledge management. 2019: 1963-1972.
[25] COOLS S B, GERSHENSON C, D'HOOGHE B. Self-organizing traffic lights: a realistic simulation[J]. Advances in applied self-organizing systems, 2013: 45-55.
[26] LEVIN M W. Max-Pressure traffic signal timing: a summary of methodological and experimental results[J]. Journal of Transportation Engineering, Part A: Systems, 2023, 149(4): 03123001.

This work is partially supported by General Program of National Natural Science Foundation of China (No.62276097); Natural Science Foundation of Shanghai (No. 22ZR1416500、No.23ZR1414900); Shanghai Pilot Program for Basic Research (22TQ1400100-16).

MIAO Zijun, born in 1999, M. S. candidate. His research interests include deep learning, few-shot learning.

LUO Fei, born in 1978, Ph. D. His research interests include cognitive computing and reinforcement learning.

DING Weichao, born in 1989, Ph. D. His research interests include cloud computing, swarm intelligence computing, federated learning.

DONG Wenbo, born in 1978, Ph. D. His research interests include machine learning and artificial intelligence.