DOI: 10.13382/j.jemi.B2205515

基于改进 Dueling DQN 的多园区网络动态路由算法*

李国燕 史东雨 张宗辉
(天津城建大学计算机与信息工程学院 天津 300392)

摘 要:针对高度“中心”连接的多园区网络中,负载不均衡造成传输时延长和网络拥塞问题,提出一种基于自适应多采样机制的决斗深度强化网络(adaptive multi-sampling Dueling deep Q-network, AMD-DQN)动态路由优化算法。首先,在网络模型中引入决斗网络(dueling DQN)的思想,同时对多层感知器组成结构进行中心化处理改进,防止高估计价值函数;然后,经验回放机制采用了自适应多采样机制,该机制融合了随机、就近和优先采样方式,根据负载情况进行自适应调整,并根据权值概率随机选取采样模式;最后,利用 AMD-DQN 网络结构结合强化学习信号和随机梯度下降来训练神经网络,选出每步最大价值动作,直至传输成功。实验结果表明,相比传统的 DQN 和 Dueling DQN 算法,AMD-DQN 算法平均时延为 128.046 ms,吞吐量达到 5.726 个/s,有效减少了数据包的传输时延,提高了吞吐量,同时从 5 个方向对拥塞程度进行评价,取得了较好的实验结果,进一步缓解了网络的拥塞。

关键词: 动态路由;深度强化学习;决斗网络;自适应多采样经验回放
中图分类号: TP181;TN91 **文献标识码**: A **国家标准学科分类代码**: 520.1040

Dynamic routing algorithm for multi campus network based on improved Dueling DQN

Li Guoyan Shi Dongyu Zhang Zonghui
(School of Computer and Information Engineering, Tianjin Chengjian University, Tianjin 300392, China)

Abstract: Aiming at the problems of transmission time delay and network congestion caused by load imbalance in highly “central” connected multi-campus networks, a dynamic routing optimization algorithm based on adaptive multi-sampling Dueling deep Q-Network (AMD-DQN) is proposed. Firstly, the idea of Dueling DQN is introduced into the network model, and the structure of the multilayer perceptron is improved by centralized processing to prevent high estimation of value function. Then, the experience playback mechanism adopts an adaptive multisampling mechanism, which combines random, nearest and priority sampling methods, adjusts adaptively according to the load situation, and randomly selects the sampling mode according to the weighted probability. Finally, the AMD-DQN network structure is combined with reinforcement learning signal and random gradient descent to train the neural network, and the maximum value action of each step is selected till the transmission is successful. The experimental results show that compared with the traditional DQN and Dueling DQN algorithms, the average delay of the AMD-DQN algorithm is 128.046 ms, and the throughput reaches 5.726/s, which effectively reduces the transmission delay of packets and improves the throughput. At the same time, the congestion degree is evaluated from five directions, and good experimental results are obtained, which further alleviates the congestion of the network.

Keywords: dynamic routing; deep reinforcement learning; fighting network; adaptive multisampling empirical playback

收稿日期: 2022-05-20 Received Date: 2022-05-20
*基金项目:天津市科技计划(19YFZCGX00130)项目资助

0 引　言

每座城市有多个园区,在设计多个园区之间的网络路由算法时要从整个网络的角度考虑,包括网络中数据的传输时延,以及遇到大量数据时的拥塞程度,其中具有高度“中心”连接的网络是多园区网络的常态,且中心更容易出现高度拥塞。高数据速率、大规模设备连接以及数据的快速增长给骨干网络带来了负担[1],导致网络拥塞风险的增加,对现有路由优化算法提出了严峻挑战[2-4]。路由算法通常分为两类,静态路由[5]和动态路由[6]。静态路由需要网络管理员手工配置的路由信息,一般适用于比较简单的网络;动态路由可以实现路由器自动建立路由表,并根据实际网络变化进行适量的调整,应用在复杂的网络环境。目前研究的主流是动态路由算法。

动态路由算法自20世纪80年代初期开始应用于网络,由最初的RIP算法[7]到满足大型网络需要的两种高级路由算法——OSPF算法[8]和IS-IS算法[9]。近年来,机器学习[10]在应对复杂的网络结构和网络流量特征时有十分良好的适应能力,为路由算法的优化提供了新的思路。其中,深度强化学习网络(deep reinforcement learning, DRL)更具有发展潜力[11-14]。

文献[15]和[16]分别提出了双深度强化网络和决斗深度强化网络,解决DQN算法的不稳定性问题。文献[17]提出了一种基于平均神经网络参数的DQN算法,解决算法中ε-贪婪策略探索效率不够高的问题。文献[18-20]通过引入经验回放机制打破DQN算法中样本之间的相关性,解决了训练过程时难收敛的问题。但是经验回放机制只是使用相同的频率对样本进行回放[21],没有对样本的重要性做进一步的判断,使得网络的训练过程很耗时。文献[22]提出了优先经验回放策略,利用样本误差(temporal difierence error,TD_er-ror)[23]作为优先级,提高优先级靠前样本[24]的选择频率。但每次对网络权重更新之前,都需要对经验池中的样本优先级进行计算,导致算法的额外开销和时间复杂度都很高。文献[25]采用了基于排序的优先经验回放,解决了DQN算法训练时间较长的问题。然而,以上研究只针对提升算法能力,没有针对特定的环境考虑,经验回放上都是单一的回放策略或在此基础上的优化。

综上所述,现有基于DQN算法的研究在高度中心化的环境下存在数据传输时延过长和网络拥塞程度过高的问题。针对此问题,面向多园区网络提出并改进一种基于Dueling DQN的动态路由优化算法。主要包括如下内容:在网络模型中引入决斗网络的思想,同时对多层感知器组成结构进行中心化处理;经验回放机制采用了自适应多采样机制,该机制融合了随机、就近和优先采样方式,根据负载情况进行自适应调整,并根据权值概率随机选取采样模式。

1 深度Q网络

强化学习[26]的常见模型是标准的马尔可夫决策过程(Markov decision process, MDP)[27],可以将其看作试探评价过程。智能体在当前环境中做出动作,环境受该动作的影响发生相应变化,同时给智能体返回一个强化信号(奖励或惩罚)。智能体根据当前环境状态和强化信号对下一个动作做出选择,选择的原则是让能产生正强化(奖)动作被选中的概率不断提升,如图1所示。选择的动作不仅对当前强化值产生影响,而且影响环境下一时刻的状态及最终的强化值。

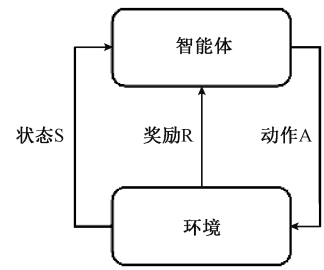

图1　强化学习流程

Fig. 1　Reinforcement learning process

在强化学习模型中包含如下3个要素:

1) 状态S:表示智能体所有可能状态的集合,即状态集。

2) 动作A:针对智能体的每个状态做出相应动作(Actions)的集合,即动作集。

3) 奖励R:表示各个状态之间的转换获得的对应回报,即奖励函数(reward function)。每个状态对应一个值,或者一个状态-动作对(state-action)对应。

其中智能体通过不断迭代更新的动作价值函数Q公式如下:

$$Q(S_t,A_t)\leftarrow Q(S_t,A_t)+\lambda(R_{t+1}+\gamma\max_a Q(S_{t+1},a)-Q(S_t,A_t)) \tag{1}$$

其中,$Q(S_t,A_t)$为在网络状态S_t的情况下智能体作出路径选择动作A_t的期望值;R_{t+1}为在网络状态S_t下做出动作A_t时的奖励值;$\max_a Q(S_{t+1},a)$表示在网络状态S_{t+1}下从各种路径选择中算出最大的期望值;γ为折扣因子值越小,未来奖励对当前奖励的影响就越小;λ为学习速率。

深度Q网络将$R_{t+1}+\gamma\max_a Q(S_{t+1},A_t)$作为目标$Q$值,并基于网络输出的$Q$值和目标$Q$值之间的偏差定义

损失函数 L：

$$L(w) = E[(R_{t+1} + \gamma \max_a Q(S_{t+1},a) - Q(S_t,A_t,w))^2] \quad (2)$$

其中，S_t,A_t 表示当前的状态和动作；S_{t+1} 表示下一时刻的网络状态；$Q(S_t,A_t,w)$ 表示神经网络在当前状态和动作，且神经网络参数为 w 时的输出值。可以采用随机梯度下降更新神经网络参数。

2　自适应多采样机制的决斗深度强化网络算法

2.1　决斗神经网络

DQN 算法通过多层感知器对智能体的状态-动作值进行估计，并根据所估计出的值进行下一步的策略决策。然而，随着神经网络参数的更新使目标状态-动作值发生变化，DQN 算法仅用一个多层感知器（神经网络）对状态-动作值进行估计，会过高估计动作的 Q 值，而且估计误差会随着动作次数的增加而增大。如果高估不一致，在某个状态下则会导致则次优动作的高估 Q 值将超过最优动作的 Q 值，并且永远不会找到最优的策略，从而导致算法训练不稳定。为了加强 DQN 算法的稳定性，Dueling DQN[28-29] 算法通过将 DQN 的神经网络输出层的状态-动作值改为独立的状态值和动作值，以突出每个动作相对于特定状态下所有动作平均值的优缺点，从而提高算法的稳定性。因此，本文引入决斗深度强化网络（Dueling deep Q-network，Dueling DQN）算法作为路由优化算法的基础。

Dueling DQN 将 Q 网络分成两部分，第 1 部分称为价值函数，仅与状态有关，不考虑其中的动作，记做 $V(S,w,\alpha)$；第 2 部分称为优势函数，同时与状态和动作相关，记为 $A(S,A,w,\beta)$，那么最终价值函数可以重新表示为：

$$Q(S,A,w,\alpha,\beta) = V(S,w,\alpha) + A(S,A,w,\beta) \quad (3)$$

其中，w 是公共部分的网络参数，α 是价值函数独有部分的网络参数，β 是优势函数独有部分的网络参数。

基于 Dueling DQN 表达的思想，本文引入决斗神经网络对网络模型进行优化。传统网络模型包括输入层、隐藏层、输出层。

为了更好地做出下一步动作价值的判断，找到最优策略，加强算法稳定性，在隐藏层和输出层之间添加状态价值函数和中心化处理模块如图 2 所示。

其中状态价值函数表示静态的环境本身具有的价值；动作价值函数表示在某状态下做出某个动作的额外价值。在实际应用中，一般要将动作价值函数设置为单独动作价值函数减去某状态下所有动作价值函数的平均值（中心化处理），这样做可以保证该状态下各动作的价值函数相对排序不变，而且可以缩小 Q 的取值范围，防止过高估计 Q 值，提高算法稳定性。最终 Q 网络的输出由状态价值函数和中心化处理的动作价值函数输出加和得到。

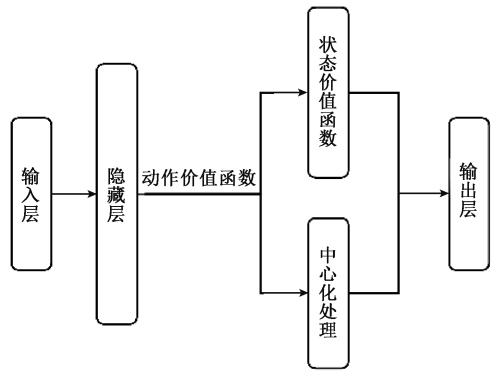

图 2　神经网络层结构

Fig. 2　Neural network layer structure

其中动作价值函数和状态价值函数分别对应 Dueling DQN 的优势函数和价值函数两部分，所以最终价值函数公式如下：

$$Q(S,A,w,\alpha,\beta) = V(S,w,\alpha) + \left(A(S,A,w,\beta) - \frac{1}{A}\sum_{a'\in A} A(S,a',w,\beta)\right) \quad (4)$$

2.2　自适应多采样经验回放机制

DQN 算法在经验池中采用随机经验回放[30] 来更新参数，以此打破样本之间的关联关系。但是，采用随机经验回放的方式，不仅效率低，而且大多数样本奖励很小，差异性不大。就近经验回放是采取最近的一些经验作为样本。以上两种方法只是等概率的随机抽取或按位置就近选取，不考虑样本的重要性。优先经验回放[31] 就是抽取经验样本时，优先抽取价值最高的一批样本，虽然为了防止产生过拟合，让低价值的样本也有较低的可能性抽到，但是程度不够，同时只采用优先经验回访采样时间会非常长。

针对网络路由中存在的拥塞程度问题，本文将融合随机经验、就近经验和最优经验回放机制，在执行经验回放前根据网络负载的变化，3 种采样方法的权值自适应发生变化，以此达到缓解网络路由拥塞的目的。自适应多采样经验回放结构如图 3 所示。

在本文的回放记忆单元处，当经验样本进来后首先判断“更新模式为随机采样”，“Y”则执行简单随机抽样，“N”则继续判断“更新模式为就近采样”，“Y”则执行就近采样，“N”则最后判断“更新模式为优先采样”是的话执优先经验回放。

通过自适应多采样经验回放机制通过优先经验回放采取优秀的样本，又可以通过随机和就近采样来缓解局

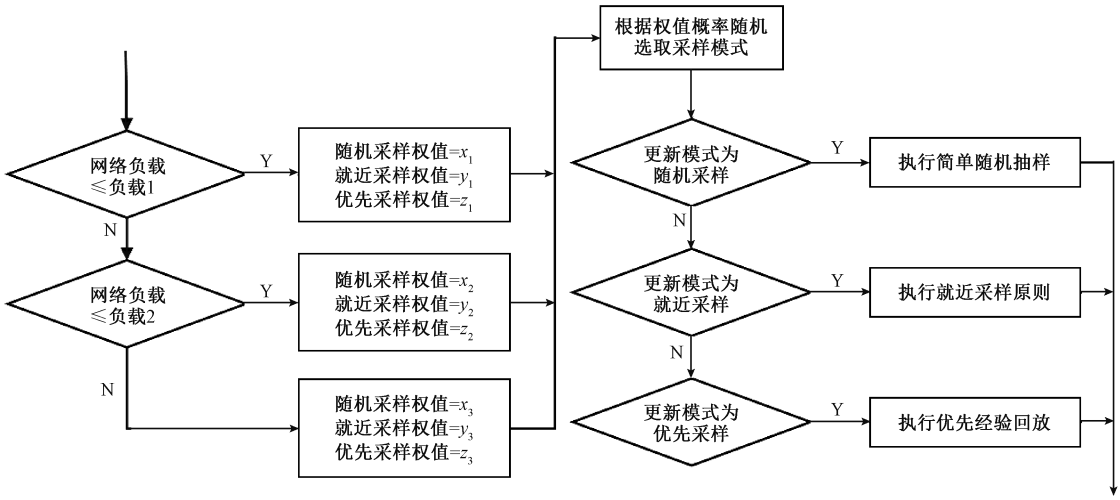

图 3　自适应多采样经验回放结构

Fig. 3　Adaptive multi-sample experience playback structure

部最优解。本文将上述采样机制应用于深度 Q 网络的结构如图 4 所示。

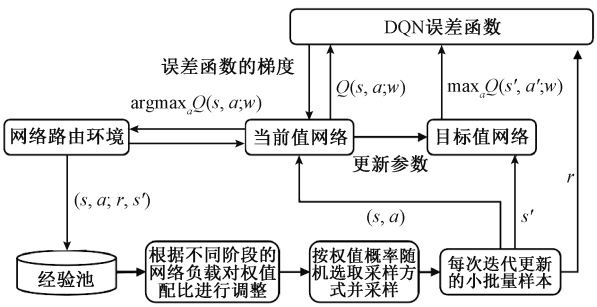

图 4　AMD-DQN 算法结构

Fig. 4　AMD-DQN algorithm structure

2.3　算法描述

将每个路由节点当作一个独立的智能体,当需要向下一节点传递数据时,路由节点根据当前网络状态下的值选择下一步的动作,以此进入下一状态,其中神经网络参数是值计算的重要参数,需要不断进行更新,算法完整流程如下:

算法　自适应多采样决斗动态路由算法

输入　网络拓扑 $G(V,E)$,奖励折扣因子 γ,经验回放池大小 N,样本批采样大小 n,学习率 λ,目标网络更新频率 L,神经网络集数 M,每个神经网络集数最大时间步 T

输出　竞争 Q 网络参数

1) 对于路由每次进行下一步路线选择的回合:

2) 初始化环境,得到初始状态 S_t。

3) 对于回合中的每一步:

4) 采用 ε-贪婪策略选择动作,随机选择一个动作,或者:

$$A_t = \text{argmax}_a Q(s_t, a; w, \alpha, \varepsilon)$$

5) 执行动作后环境产生相应的奖励和下一时刻的状态,计算时序误差:

$$\delta_t = R_t + \gamma \max_a Q(s_t, a_t; w, \alpha, \varepsilon) - Q(s_{t-1}, a_{t-1}; w, \alpha, \varepsilon)$$

6)将时序误差 δ_t 按从大到小排列,得到 rank(t)。

7) 根据此时的负载情况,利用自适应公式算出 3 种采样方法的权值配比,进行适当调整。

8) 按照随机、就近和优先采样机制各自的权值概率随机选取采样方法。

9) 在经验池中根据选取的采样方法进行采样。

10) 如果是优先经验回放需要按照: $P(i) = \dfrac{p_i}{\sum p_i}$ 和 $p_i = |\delta_t| + \epsilon$ 来计算样本的优先值。

11) 损失函数为: $E[(R_t + \gamma \max_{A_t} Q(S_t, A_t; w, \alpha, \varepsilon) - Q(S_{t-1}, A_{t-1}; w, \alpha, \varepsilon))^2]$,同时更新网络。

12) 每隔 L 步,将目标网络参数以竞争 Q 网络参数代替更新。

上述算法采用的是决斗网络结构,增加了价值函数和优势函数(隐藏层输出的初始结果),使算法的空间复杂度更高,当动作空间维数为 K,存储开销增长了 $O(1) + O(K)$,总存储开销为 $O(K)$。经验回放机制中随机、就近和优先采样,权值分别为 x、y 和 z,在容量为 N 的经验回放池中采样的时间复杂度分别为 $O(1)$、$O(1)$ 和 $O(N)$,因为 $x + y + z = 1$,所以总时间复杂度为: $\dfrac{z}{x + y + z} O(N) = zO(N)$。

3　仿真实验

3.1　实验场景及参数设计

为了在动态网络上测试 AMD-DQN 路由算法的性能，本文在 Win10 64 位系统下，利用 Python torch 深度学习框架，编程语言采用 Python3.8 进行仿真实验以及结果分析。

仿真实验中使用 Barabasi-Albert 模型生成多园区网络。创建一个中心节点，并向外添加节点分支，产生一个具有高度“中心”连接良好的图，如图 5 所示。这是 Internet 网络的代表，并且在其中心容易出现高度拥塞。网络环境为 50 个节点网络，边缘的平均度数为 3，节点在任何给定时间步可以容纳的数据包数 150 个，最大传输数据包 10 个，在每集模拟结束之前有 5 000 个数据包要开始和 5 000 个额外的数据包要交付，最大边缘权重为 10，移除边缘数最大为 10，最小为 0。

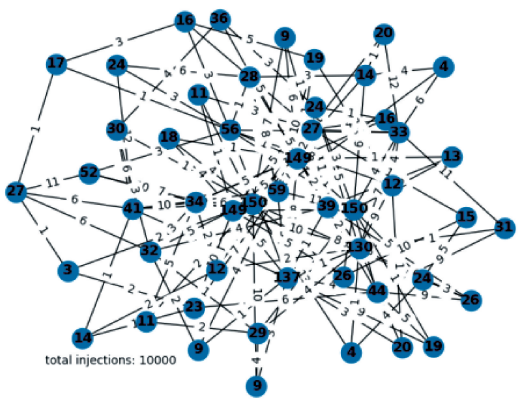

图 5　网络结构

Fig. 5　Network structure

在模拟的一集（仿真中每个负载节点测试的集数为 15，每集最大为 2 000 时间步，并去除异常数值后，数据均值化处理）中，随机选择边缘消失并在每个时间步恢复。此外，边缘权重在整个情节中以正弦方式波动。在每集开始时，在网络中生成大量数据包（网络负载），每个数据包都有一个随机的起始节点和目标节点。每次传递数据包时，都会在一定数量的时间步之后初始化一个新数据包当在在网络上生成并传送了一定数量的数据包，这一集就结束。

3.2　实验结果及分析

为验证本文所提出的自适应多采样决斗动态路由算法在网络负载较高的情况下所产生的优化效果，本文主要从网络拥塞程度和平均传输时延两部分进行实验验证。使用多采样经验回放的 DQN 称为 M-DQN，在这里本文使用 4 种不同权值 M-DQN 和 DQN 进行对比实验，如表 1 所示。

表 1　采样方法权重配比

Table 1　Sampling method weight ratio

算法	采样方法权重		
	随机采样	就近采样	优先采样
DQN	1	0	0
M-DQN1	0	0	1
M-DQN2	0.2	0.2	0.6
M-DQN3	0.4	0.3	0.3
M-DQN4	0.5	0.2	0.3

网络路由传输数据的过程中，从网络结构整体上看，相同网络负载下工作节点越少，空闲节点越多，拥塞程度越低；从路由节点上看，每个节点所具有的平均和最大数据包数量少，拥塞程度越低；从传输的数据上看，数据包的空闲时间越小，传输的就越快，拥塞程度越低。如果优先采样权值过高，在网络负载较低时由于数据样本较少优先采样过于耗时，会导致每个结点的数据包较多且空闲时间较长，网络拥塞程度提高；在网络负载较高时由于优先极高的样本较多，没有更好的样本进行替换，陷入局部最优解，网络传输的路径变长，会导致网络路由中工作节点变多，空闲节点变少，拥塞程度上升。

1）图 6 展示了网络负载增加的情况下网络中工作节点的占有比，百分比越小越好，从图 6 中可以看出 M-DQN2 跟 DQN 相似，M-DQN1、M-DQN3 和 M-DQN4 明显比 DQN 低，平均低出 3.314%、3.765% 和 5.153%，其中除了 1 500~2 500 负载阶段，M-DQN4 最优。

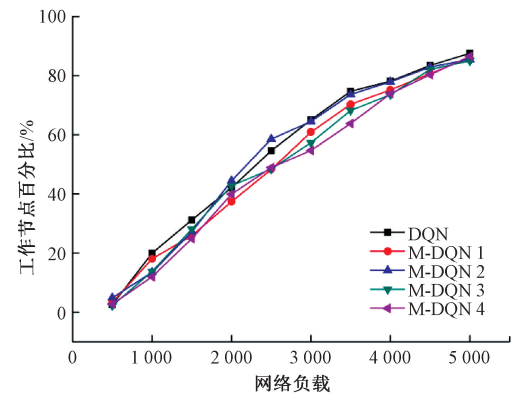

图 6　工作节点百分比

Fig. 6　Percentage of working nodes

2）图 7 展示了网络负载增加的情况下网络中空闲节点的占有比，百分比越大越好，从图中可以看出 M-DQN3 初始负载时比 DQN 高出约 5%，总体平均高出 2.038%，M-DQN1、M-DQN2 和 M-DQN4 初始负载时比 DQN 高出约 10%，但 M-DQN1 和 M-DQN2 总体平均高出 1.461% 和 1.096%，而 M-DQN4 总体平均高出 3.053%。

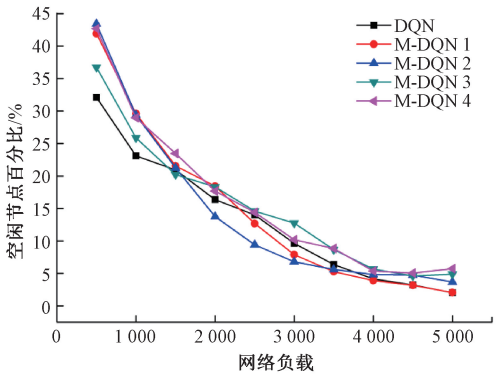

图7　空闲节点百分比

Fig. 7　Percentage of idle nodes

3)图8展示了网络负载增加的情况下网络中平均数据包的空闲时间,时间越小越好,从图中可以明显看出M-DQN3在整个负载情况下比DQN平均高出不少,M-DQN2在2 500负载前比DQN高出很多,M-DQ1与M-DQN4分别比DQN平均整体低出31.969和75.831 ms,同时也比其他权重配比的M-DQN低。

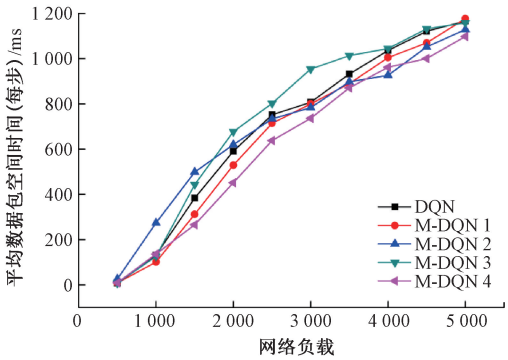

图8　平均数据包空闲时间

Fig. 8　Average packet idle time

4)图9展示了网络负载增加的情况下网络中每个节点平均数据包数量,平均数据包数量越少越好,从图中可以看出M-DQN3的数据都高于DQN,分别平均高出1.414个,M-DQN1、M-DQN2和M-DQN4优于DQN,分别平均比其低1.196、1.733和2.247个,所以M-DQN4最好。

5)图10展示了网络负载增加的情况下网络中每个节点平均数据包数量,在每个节点最大数据包数量和网络负载对比中,最大数据包数量越少越好,同时越晚达到最大量越好。从图10中可以看出M-DQN3在1 000负载时分别比DQN高35个,后面和DQN相似,平均低0.5个。M-DQN1在2500负载以前与DQN相似,平均整体低2.3个。M-DQN2和M-DQN4在1 000负载时分别比DQN高72和45个,但之后又平均低5.625和6.5个,其中M-DQN4低的更多且上涨趋势比M-DQN2缓慢。

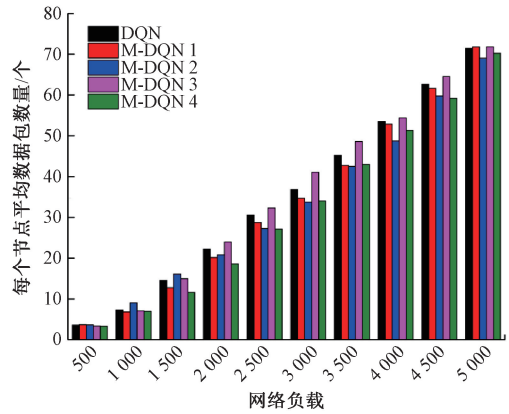

图9　每个节点平均数据包数量

Fig. 9　Average data packets per node

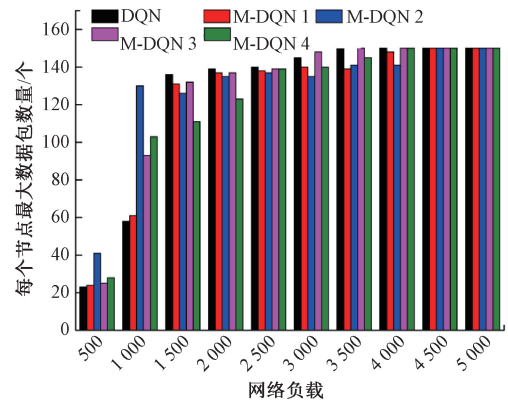

图10　每个节点最大数据包数量

Fig. 10　Maximum number of packets per node

为了进一步直观的进行判断,总结了4种权值配比的M-DQN与DQN差值的平均值,正数和负数分别代表高于和低于DQN的数值大小,具体如表2所示。

表2　5个方向上与DQN差值均值表

Table 2　Table of mean value of difference withDQN in five directions

方向	算法			
	M-DQN1	M-DQN2	M-DQN3	M-DQN4
工作节点百分比/%	-3.314	-0.641	-3.765	-5.153
空节点百分比/%	1.462	1.096	2.038	3.053
平均数据包空闲时间/ms	-31.969	1.313	43.146	-75.831
每个节点平均数据包数量/个	-1.196	-1.733	1.414	-2.247
每个节点最大数据包数量/个	-2.3	4.5	3.3	-0.2

通过上面的实验结果对比,选出其中的最优的M-DQN权值配比“0.5-0.2-0.3”,将经验回放机制更新以后,从图11中可以发现数据包的传输时延虽然和DQN

相比有明显的下降,但和其他比例算法的传输时延逊色许多,如表 3 所示。

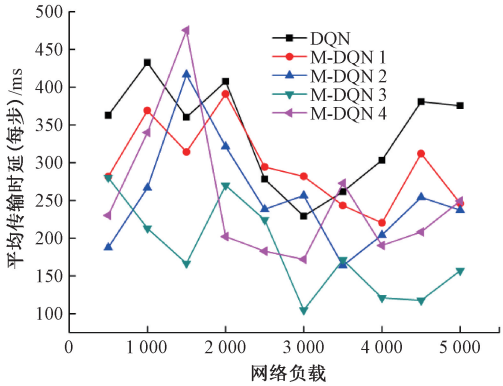

图 11　平均传输时延(a)

Fig. 11　Average transmission time (a)

表 3　4 种权值配比的 M-DQN 与 DQN 的传输时延对比

Table 3　Comparison of transmission time between M-DQN and DQN with four weight ratios

算法	平均传输时延/ms
DQN	369.999
M-DQN1	295.305
M-DQN2	207.741
M-DQN3	182.499
M-DQN4	252.312

本文对神经网络模型进行优化,改变了价值函数的组成结构,将加入改进后的 M-DQN 称为 MD-DQN。同时,加入基于优先经验回放的决斗神经网络(Dueling DQN with prioritized experience replay,Dueling DQN-PR)算法进行对比实验。由于 M-DQN 具有一定权重的优先采样机制,所以前期会有一段传输时延较高的情况;MD-DQN 在采用了多采样经验回放机制的基础上加入了决斗神经网络,改变了动作价值函数的组成结构,使路由的下一步选择更优,从而达到降低传输时延的作用;Dueling DQN-PR 整体采用的优先经验回访机制,就相当于权值配比为“0-0-1”的 MD-DQN;SMD-DQN 是对 MD-DQN 进行了分段式处理,在 0~1 500 负载和 4 000~5 000 负载之间用 MD-DQN(0.5-0.2-0.3),在 1 500~4 000 负载之间用 MD-DQN(0-0-1);AMD-DQN 是为了进一步降低 SMD-DQN 在拥塞程度 5 个小方面的效果,对权值变化加入了关于三角函数的自适应函数,具体如图 12 所示。

从实验结果图 13 中,可以明显看出 MD-DQN 的平均传输时延比 DQN 和 M-DQN 低很多,平均传输时延为 173.552 ms;从图 13 中可以看出在 1 500~4 000 负载之间,Dueling DQN-PR 的传输时延明显比 MD-DQN 低,且平均传输时延为 169.281 ms;SMD-DQN 吸收上面两种算法各自优秀的部分,按照不同阶段对权值进行固定比例调整,平均传输时延为 162.366 ms;AMD-DQN 在权值比例上采取的是自适应变化,平均传输时延为 128.046 ms,最优。表 4 为 6 种算法的传输时延对比。

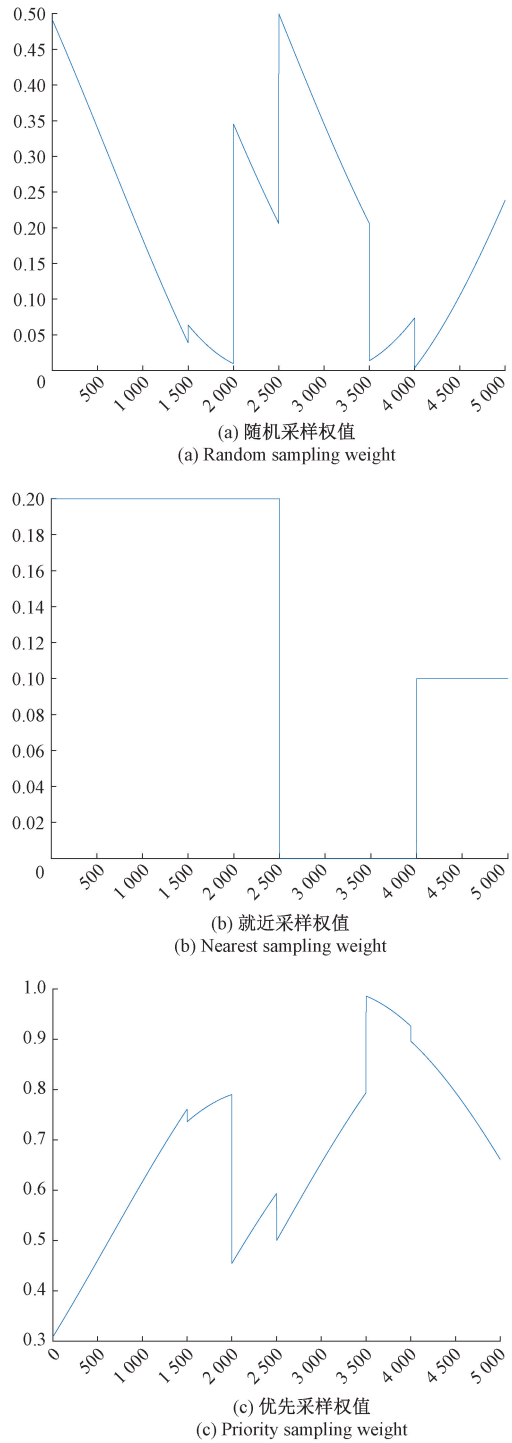

图 12　3 种采样方式权值变化

Fig. 12　Weight change of three sampling methods

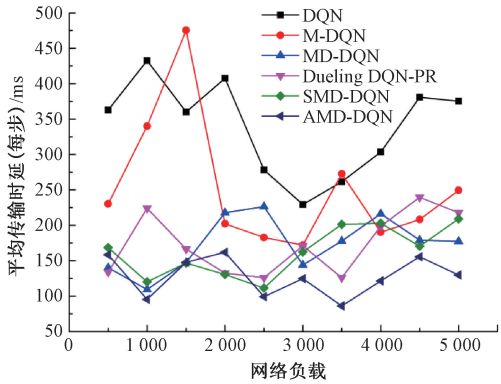

图 13　平均传输时延(b)

Fig. 13　Average transmission time (b)

表 4　6 种算法的传输时延对比

Table 4　Transmission time comparison of sixalgorithms

算法	平均传输时延/ms
DQN	339.146
M-DQN	252.312
MD-DQN	173.552
Dueling DQN-PR	169.281
SMD-DQN	162.366
AMD-DQN	128.046

表 5 是 AMD-DQN 在评价拥塞程度的 5 个方面和 DQN、M-DQN(0.5-0.2-0.3)、MD-DQN、Dueling DQN-PR 以及 SMD-DQN 的对比数据。首先要知道 AMD-DQN 是为了在保持 SMD-DQN 优秀的平均传输时延的基础上进一步优化其拥塞程度,除了空节点百分比均值,其他 4 个方向都比其有较大的优化。另一方面,在整个负载下是一个权值配比的算法中,MD-DQN 的整体拥塞度是较低的,权值为 0.5-0.2-0.3,Dueling DQN-PR 的拥塞程度较高,权值为 0-0-1,再加上前面对于权值配比的实验中可以发现当优先采样的权值不断升高后,拥塞程度会相应的升高,但在某些阶段优先采样的权值越高,越能做出好的选择,缩短平均传输时延。所以有了 AMD-DQN 的出现,通过不断让权值配比随着负载的不断上升而发生相应的变化,在保持较快的平均传输时延同时还有着较低的拥塞程度。

前面已经从传输时延和拥塞程度两个主要的方向进行了对比,最后将从吞吐量方向进行对比实验。

从图 14 和表 6 中,可以看到 6 种算法的吞吐量随着网络负载增加的变化,以及平均吞吐量的比较。在不同的网络负载下 AMD-DQN 算法的吞吐量都有着较好的表现,整体的吞吐量达到了 5.726 个/s,相比未进行改进的 DQN 高出 1.181 个/s,比对比算法 Dueling DQN-PR 高出 1.296 个/s。

表 5　5 个方向均值表

Table 5　Five direction mean table

方向	算法					
	DQN	M-DQN	MD-DQN	Dueling DQN-PR	SMD-DQN	AMD-DQN
工作节点百分比均值/%	53.918	48.765	49.561	51.166	51.919	49.230
空节点百分比均值/%	13.191	16.244	18.175	15.048	19.143	17.183
数据包空闲时间均值/ms	692.474	616.643	650.833	675.244	738.521	623.365
每个节点平均数据包数量均值/个	34.798	32.551	32.470	34.327	37.979	33.09
每个节点最大数据包数量均值/个	124.1	123.9	111.6	129.4	126.1	117.7

表 6　6 种算法的吞吐量对比

Table 6　Throughput comparison of six algorithms

算法	平均吞吐量/(个/s)
DQN	3.908
M-DQN	4.425
MD-DQN	4.468
Dueling DQN-PR	4.43
SMD-DQN	4.811
AMD-DQN	5.726

4　结　论

路由是网络的核心组成部分,随着路由数量增多,网络拓扑变得复杂,数据流量的爆炸式增长,传统路由策略在一些网络环境中已经不再适用了。而随着路由硬件设备和深度强化学习的快速发展,深度强化学习在网络路由中的潜力不断被人发觉。本文主要是在具有高度“中心”连接的多园区网络环境中,针对动态路由中数据传输时延和网络拥塞程度优化的问题,对基于自适应多采样经验回放的决斗深度 Q 网络算法进行研究。通过实验结果表明,基于自适应多采样机制的决斗深度强化网络算法有效地减少了数据包的传输时延和网络的拥塞程度,提高吞吐量。

参考文献

[1] LIU Z, LIU Y, MENG Q, et al. A tailored machine learning approach for urban transport network flow estimation [J]. Transportation Research Part C: Emerging Technologies, 2019, 108: 130-150.

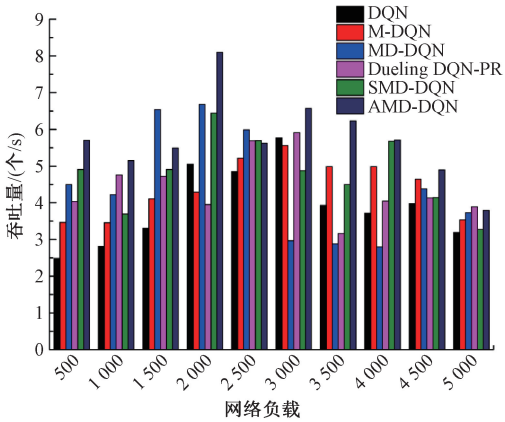

图 14　吞吐量

Fig. 14　throughput

[2] KHATARI M, ZAIDAN A A, ZAIDAN B B, et al. Multidimensional benchmarking framework for AQMs of network congestion control based on AHP and Group-TOPSIS [J]. International Journal of Information Technology & Decision Making, 2021, 20(5): 1409-1446.

[3] LAI J W, CHANG J, ANG L K, et al. Multi-level information fusion to alleviate network congestion [J]. Information Fusion, 2020, 63: 248-255.

[4] 欧阳一鸣,陈志谋,王奇,等. WiNoC 中基于 Edge-first 算法的流量平衡设计[J]. 电子测量与仪器学报, 2021,35(1):62-73.
OUYANG Y M, CHEN ZH M, WANG Q, et al. Traffic balance design based on edge first algorithm in winoc [J] Journal of Electronic Measurement and Instrumentation, 2021,35(1): 62-73.

[5] CAI S, SHU Y, WANG W. Dynamic routing networks[C]. Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021: 3588-3597.

[6] CHAMPATI J P, AL-ZUBAIDY H, GROSS J. Transient analysis for multihop wireless networks under static routing[J]. IEEE/ACM Trans. on Networking, 2020, 28(2): 722-735.

[7] BIRADAR A G. A comparative study on routing protocols: RIP, OSPF and EIGRP and their analysis using GNS-3 [C]. 2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE). IEEE, 2020: 1-5.

[8] ASHAR M, WIBAWA A P. IPv6 vs IPv4 performance simulation and analysis using dynamic routing OSPF[C]. 2021 4th International Conference of Computer and Informatics Engineering (IC2IE). IEEE, 2021: 452-456.

[9] MULIANDRI E, TRISNAWAN P H, AMRON K. Analisis perbandingan kinerja routing protokol IS-IS dengan routing protokol eigrp dalam dynamic routing[J]. Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer e-ISSN, 2019, 2548: 964X.

[10] 王桂芝,吕光宏,贾吾财,等. 机器学习在 SDN 路由优化中的应用研究综述[J]. 计算机研究与发展,2020, 57(4):688-698.
WANG G ZH, LYU G H, JIA W C, et al. A review on the application of machine learning in SDN routing optimization [J]. Journal of Computer Research and Development, 2020, 57(4): 688-698.

[11] 伍元胜. 面向动态拓扑网络的深度强化学习路由技术[J]. 电讯技术,2021,61(6):659-665.
WU Y SH. Deep reinforcement learning routing for dynamic topology networks [J]. Telecommunication Engineering, 2021, 61(6): 659-665.

[12] 刘全, 翟建伟, 章宗长, 等. 深度强化学习综述[J]. 计算机学报, 2018, 41(1): 1-27.
LIU Q, ZHAI J W, ZHANG Z CH, et al. A survey on deep reinforcement learning [J]. Chinese Journal of Computers, 2018, 41(1): 1-27.

[13] LEI C. Deep Reinforcement Learning [M]. Deep Learning and Practice with MindSpore. Springer, Singapore, 2021: 217-243.

[14] CHEN P, LU W. Deep reinforcement learning based moving object grasping[J]. Information Sciences, 2021, 565: 62-76.

[15] VAN HASSELT H, GUEZ A, SILVER D. Deep reinforcement learning with double q-learning [C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2016, 30(1): 2094-2100.

[16] WANG Z, SCHAUL T, HESSEL M, et al. Dueling network architectures for deep reinforcement learning [C]. International Conference on Machine Learning. PMLR, 2016: 1995-2003.

[17] 黄志勇,吴昊霖,王壮,等. 基于平均神经网络参数的 DQN 算法[J]. 计算机科学,2021,48(4):223-228.
HUANG ZH Y, WU H L, WANG ZH, et al. DQN algorithm based on averaged neutral network parameters[J]. Computer Science, 2021, 48(4): 223-228.

[18] SCHAUL T, QUAN J, ANTONOGLOU I, et al. Prioritized experience replay[J]. arXiv preprint arXiv: 1511.05952, 2015.

[19] 朱斐, 吴文, 刘全, 等. 一种最大置信上界经验采样的深度 Q 网络方法[J]. 计算机研究与发展, 2018, 55(8): 1694-1705.
ZHU F, WU W, LIU Q, et al. A deep Q-network

method based on upper confidence bound experience sampling [J]. Journal of Computer Research and Development, 2018, 55(8): 1694-1705.

[20] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning [J]. Nature, 2015, 518(7540): 529-533.

[21] YANG Y, JUNTAO L, LINGLING P. Multi-robot path planning based on a deep reinforcement learning DQN algorithm [J]. CAAI Transactions on Intelligence Technology, 2020, 5(3): 177-183.

[22] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning[J]. arXiv preprint arXiv:1312.5602, 2013.

[23] 王俊茜,郑文先,徐勇. 基于测试样本误差重构的协同表示分类方法[J]. 计算机科学, 2020, 47(6): 104-113.
WANG J X, ZHENG W X, XU Y. Next generation network technology, intelligent information processing and other collaborative tables based on test sample error reconstruction show classification method[J]. Computer Science, 2020,47(6):104-113.

[24] KE F, ZHAO D, SUN G, et al. A priority experience replay sampling method based on upper confidence bound [C]. Proceedings of the 2019 3rd International Conference on Deep Learning Technologies, 2019: 38-41.

[25] 周瑶瑶,李烨. 基于排序优先经验回放的竞争深度Q网络学习[J]. 计算机应用研究, 2020, 37(2): 486-488.
ZHOU Y Y, LI Y. Dueling deep Q network learning with rank-based prioritized experience replay[J]. Application Research of Computers, 2020, 37(2): 486-488.

[26] 车向北,康文倩,欧阳宇宏,等. 基于强化学习的SDN路由优化算法[J]. 计算机工程与应用,2021,57(12): 93-98.
CHE X B, KANG W Q, OUYANG Y H, et al. SDN routing optimization algorithm based on reinforcement learning[J]. Computer Engineering and Applications, 2021, 57(12): 93-98.

[27] TARBOURIECH J, LAZARIC A. Active exploration in markov decision processes [C]. The 22nd International Conference on Artificial Intelligence and Statistics. PMLR, 2019: 974-982.

[28] BAN T W. An autonomous transmission scheme using dueling DQN for D2D communication networks [J]. IEEE Transactions on Vehicular Technology, 2020, 69(12): 16348-16352.

[29] SEWAK M. Deep Q Network (DQN), Double DQN, and Dueling DQN [M]. Deep Reinforcement Learning. Springer, Singapore, 2019: 95-108.

[30] GUAN Y, LIU B, ZHOU J, et al. A new subsampling deep Q network method [C]. 2020 International Conference on Computer Network, Electronic and Automation (ICCNEA). IEEE, 2020: 26-31.

[31] GAN J, LI X, LIU W, et al. Prioritized experience replay method based on experience reward [C]. 2021 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE). IEEE, 2021: 214-219.

作者简介

李国燕,2006年于河北师范大学获得学士学位,2009年于河北工业大学获得硕士学位,2013年于河北工业大学获得博士学位,现为天津城建大学副教授,主要研究方向为下一代网络技术、智能信息处理等。
E-mail: ligy@tcu.edu.cn

Li Guoyan received her B. Sc. degree from Hebei Normal University in 2006, M. Sc. degree from Hebei University of Technology in 2009, and Ph. D. degree from Hebei University of Technology in 2013, respectively. Now she is an associate professor in Tianjin Chengjian University. Her main research interests include next generation network technology and intelligent information processing, etc.

史东雨,2019年于河北师范大学汇华学院获得学士学位,现为天津城建大学计算机与信息工程学院硕士研究生,主要研究方向为面向智慧园区的网络路由优化设计等。
E-mail:2786732068@qq.com

Shi Dongyu received his B. Sc. degree from Huihua College of Hebei Normal University in 2019. Now he is a M. Sc. candidate in School of Computer and Information Engineering, Tianjin Urban Construction University. His main research interests include network routing optimization design for smart parks, etc.

张宗辉(通信作者),2011年于衡水学院获获得学士学位,2014年于河北工业大学获得硕士学位,现为天津城建大学助理实验师,主要研究方向为数据库等。
E-mail: 1173888404@qq.com

Zhang Zonghui (Corresponding author) received his B. Sc. degree from Hengshui University in 2011, M. Sc. degree from Hebei University of Technology in 2014. Now he is an assistant experimentalist in Tianjin Chengjian University. His main research interests include database, etc.