·人工智能与模式识别·　　文章编号：1000-3428(2023)06-0099-08　　文献标志码：A　　中图分类号：TP181

Q学习演化博弈中决策机制对网络合作水平的影响

张尊栋[1,2]，王岩楠[1]，周慧娟[1]，张艺帆[3]

（1.北方工业大学 城市道路交通智能控制技术北京市重点实验室，北京 100144；2.华盛顿大学 智能城市交通系统实验室，美国 西雅图 98195；3.北京交通大学 轨道交通控制与安全国家重点实验室，北京 100044）

摘　要：针对博弈决策过程中个体无法获取邻居收益的问题，基于Q学习自我经验学习的特性，提出Q学习演化博弈模型。考虑到不同Q学习决策机制会对网络合作水平产生不同的影响，采用ε-greedy决策机制、Boltzmann决策机制和Max-plus决策机制，针对不同的网络类型、不同的博弈模型参数和不同的强化学习参数进行对比实验，量化分析决策机制对网络合作水平的影响。实验结果表明：与传统的演化博弈模型相比，Q学习演化博弈模型能够普遍提高网络的合作水平，并且不同的Q学习决策机制会对网络合作水平产生不同的影响，使用ε-greedy决策机制的模型合作水平比另两种模型高约35%和37%；较低的学习率、较高的折扣率以及适中的收益均匀性能够促进网络中个体间的合作，使用ε-greedy决策机制的模型合作水平比在较高学习率和较低折扣率下的合作水平分别高约40%和45%；在较高的探索率下，引入考虑个体全局属性的Max-plus决策机制的网络平均收益比引入另两种决策机制的Q学习模型高约22%和17%。

关键词：Q学习；决策机制；网络演化博弈；合作水平；折扣率

开放科学（资源服务）标志码（OSID）：

中文引用格式：张尊栋，王岩楠，周慧娟，等. Q学习演化博弈中决策机制对网络合作水平的影响[J]. 计算机工程，2023，49(6)：99-106，114.

英文引用格式：ZHANG Z D，WANG Y N，ZHOU H J，et al. The influence of decision mechanisms on network cooperation level in Q-learning evolutionary game[J]. Computer Engineering，2023，49(6)：99-106，114.

The Influence of Decision Mechanisms on Network Cooperation Level in Q-learning Evolutionary Game

ZHANG Zundong[1,2]，WANG Yannan[1]，ZHOU Huijuan[1]，ZHANG Yifan[3]

（1.Beijing Key Laboratory of Urban Intelligent Traffic Control Technology，North China University of Technology，Beijing 100144，China；2.Intelligent Urban Transportation Systems Laboratory，University of Washington，Seattle 98195，USA；3.State Key Laboratory of Rail Traffic Control and Safety，Beijing Jiaotong University，Beijing 100044，China）

【Abstract】Aiming at addressing the problem that individuals face an inability to obtain benefits from their neighbors in the process of game decision making，this study examines the characteristics of self-experiential learning of Q-learning，thereby proposing a Q-learning evolutionary game model. Considering that different Q-learning decision mechanisms have different effects on the cooperation level of the network，the influence of the decision mechanism on the network cooperation level is quantitatively analyzed using three Q-learning decision mechanisms：ε-greedy，Boltzmann，and Max-plus by conducting comparative experiments on different network types，game model parameters，and reinforcement learning parameters. Experiments show that compared with the traditional evolutionary game models，the Q-learning evolutionary game model can generally improve the cooperation level of the network，with different Q-learning decision mechanisms having different effects on the cooperation level of the network. The cooperation level of the model using the ε-greedy decision mechanism is approximately 35% and 37% higher than that of the models using the Boltzmann and Max-plus decision mechanisms，respectively. Lower learning rates，higher discount rates，and moderate benefit uniformity promote cooperation between individuals in the network，such that for the ε-greedy decision mechanism，the cooperation level of the model using lower learning and higher discount rates is about 40% and 45% higher than that of the models using higher learning and lower discount rates，respectively. At the higher exploration level，introducing the Max-plus decision mechanism to consider global attributes of individuals improves the cooperation level by about 22% and 17% compared to using the ε-greedy and Boltzmann decision mechanisms，respectively.

基金项目：“十三五”国家重点研发计划（2018YFB1601000）。

作者简介：张尊栋（1979—），男，讲师、博士，主研方向为智能交通；王岩楠，硕士研究生；周慧娟，副教授、博士；张艺帆，博士研究生。

收稿日期：2022-04-14　　**修回日期：**2022-07-04　　**E-mail：**zdzhang@ncut.edu.cn

【Key words】Q-learning;decision mechanism;network evolutionary game;cooperation level;discount rate
DOI:10.19678/j.issn.1000-3428.0064463

0 概述

演化博弈理论为研究自私个体间合作行为的起源及演化提供了强大的理论框架[1-2]。囚徒困境(PDG)作为最典型的演化博弈模型被广泛用于研究群体间的合作行为,很多机制能够促进囚徒困境博弈中的合作[3],例如异质性[4]、重复作用[5]、空间效应[6-7]、利他性[8-9]等。近期,收益的分布对合作行为的影响引起了学者的关注[10-11]:PERC[10]提出了若干重新调节收益的方法,发现收益非均匀性太强的分布阻碍合作;刘永奎[12]发现在传统的演化博弈中适当地调节收益矩阵的不均匀性能够提高网络的合作水平。此外,策略更新规则在合作演化过程中也起到关键作用。

对于博弈策略的更新方式,研究者们提出了不同的规则,如无条件模仿[13-14]、复制因子动力学[15-16]、费米规则[17]、吸气驱动策略更新规则等。DU等[18]和CHIO等[19]提出了基于PSO算法的个体策略更新规则,研究了囚徒困境[20]和公共物品博弈[21]下的合作行为。在大多数现有的研究中,个体以一定概率学习邻居的策略来改变自身的策略。然而考虑到信息的私密性,在演化博弈过程中个体有时无法获取邻居的收益状况,只能通过从自己的经验中学习来做出决定。因此,引入强化学习方法进行策略选择已成为网络演化博弈研究的趋势。

强化学习(RL)是强大的机器学习方法之一,广泛用于解决对各种环境中的状态、行为、奖励和决策难以建模的问题[22-23]。常用的RL算法包括时间差异、动态编程[24-25]、Sarsa、Q学习[26]等。Q学习是一种具有代表性的强化学习算法,其不需要对所处的环境进行建模,并且具有能在智能体与环境的相互作用中在线使用的优良特性,非常适用于不完全信息博弈。因此,更多的研究开始将强化学习作为策略选择机制引入复杂网络演化博弈中:NING等[27]提出了一种具有可移动个体的新型空间演化雪堆博弈模型,在此模型中,个体每回合仅与最近的邻居互动,并基于强化学习过程做出决策;ZHANG等[28]介绍一种典型的强化学习算法——演化博弈Q学习,个体基于自学习追求最佳决策而不是基于传统演化博弈中的生死模仿过程;ZHANG等[29]将Q学习算法引入到演化博弈中,提出了演化博弈中的多智能体强化学习模型,并基于Q学习给出了算法流程;徐琳等[30]在多智能体Q学习中引入协作学习,在个体学习后对通信Q值进行融合再学习,从而提高算法性能和收敛速度;韩晨等[31]将Q学习与Stackelberg博弈模型结合,提出了一种基于分层Q学习的联合抗干扰学习算法,该算法具有较强的抗干扰能力。

实际上,个体进行策略更新的强化学习决策机制并不是唯一的,不同的强化学习决策机制会对网络合作水平产生不同的影响。因此,本文采用3种不同的Q学习决策机制:单纯以ε概率选择最优策略的ε-greedy决策机制,决策随机性随时间变化的Boltzmann决策机制,考虑个体全局属性的Max-plus决策机制,通过在规则网络、小世界网络、随机网络和无标度网络上进行PDG-Q学习来研究其对合作演变的影响。除此之外,还在强化学习演化博弈中引入参数θ来调节个体收益的均匀性。本文主要研究以下几个问题:引入Q学习是否能够提高网络的合作水平;不同的Q学习决策机制如何影响不同复杂网络的合作水平;在Q学习演化博弈中,个体收益的均匀性以及Q学习模型的参数(学习率和折扣率)对网络合作水平产生的影响;考虑个体全局属性的Q学习决策机制是否能够提高网络的平均收益。

针对上述问题,本文描述引入强化学习之后的囚徒困境演化博弈模型,并对比分析3种强化学习决策机制,即ε-greedy决策机制、Boltzmann决策机制、Max-plus决策机制在4种复杂网络模型中的有效性,探究多个参数对网络合作水平的影响。

1 Q学习演化博弈模型

1.1 演化博弈模型

本文构造了4个网络模型:无标度网络,小世界网络,随机网络,规则网格网络。

在建立网络之后,每一个网络的点都被一个个体占据。每一个个体只有2种状态:背叛(D)或者合作(C)。同样,每个个体都是一个纯策略者并且只有背叛(D)或者合作(C)2种策略,如式(1)所示:

$$\boldsymbol{z}_x=\begin{pmatrix}0\\1\end{pmatrix} \tag{1}$$

根据之前的研究,本文使用囚徒困境博弈的简化模型。本文选择R=1,P=S=0,T=b(1≤b≤2),这里b代表对于合作者来说背叛的诱惑。因此,本文可以通过调整单个参数b来重新调整这个囚徒困境博弈。收益矩阵如式(2)所示:

$$\boldsymbol{A}=\begin{bmatrix}1&0\\b&0\end{bmatrix} \tag{2}$$

在一个迭代过程中,每一个个体都和一阶邻居进行一次囚徒困境博弈,收益的累计取决于收益矩阵的因素。一个确定的个体总的收益是在一个迭代中遍历所有相互作用之后的总和,可以表示为:

$$R_x=\sum_{y\in\Omega_x}\boldsymbol{z}_x^{\mathrm{T}}\boldsymbol{A}\boldsymbol{z}_y \tag{3}$$

其中:Ω_x代表b最近的所有邻居。

在经典的囚徒困境演化博弈模型中,个体的策略选择机制是玩家通过向在前一轮中获得最高收益的邻居学习来改变其策略。然而在实际博弈中考虑到信息的私密性个体无法提前获取对方收益的情况下,只能通过从自己的经验中学习来做出策略选择。

1.2　Q学习规则

Q学习是一种能处理动态环境中的不确定信息，使智能体选择最优策略的算法，并能在智能体与环境的相互作用下在线使用，这种优良的特性非常适合用在不完全信息博弈。因此，本文选择把Q学习算法引入演化博弈模型中，将其作为演化博弈进程中个体的策略选择机制。

在演化博弈最初的阶段，每个个体选择合作或者背叛的概率是相同的。在演化过程中，每个个体通过自身的Q表来选择下一轮的策略，在每一个迭代中同步地更新它的Q表和状态。

Q表是状态（行）-动作（列）组合成的矩阵，其定义和更新过程如式（4）所示：

$$\boldsymbol{Q}_\tau=\begin{bmatrix} Q_{\mathrm{CC}}(\tau) & Q_{\mathrm{CD}}(\tau) \\ Q_{\mathrm{DC}}(\tau) & Q_{\mathrm{DD}}(\tau) \end{bmatrix} \tag{4}$$

得到上述矩阵后，如果节点i的状态s和动作a在第τ轮，则元素Q_a^s的更新过程如式（5）所示，其中$w(\tau)$代表个体本轮的奖励，由式（6）给出。

$$Q_a^s(\tau+1)=(1-\alpha)Q_a^s(\tau)+\alpha\left[w(\tau)+\gamma\max Q_{a'}^{s'}(\tau)\right] \tag{5}$$

$$w(\tau)=(R_x/n)^\theta \tag{6}$$

其中：R_x表示个体与邻居博弈的收益总和；n表示邻居个数；$\theta\in(0,1]$是用来调节个体收益均匀性的参数；$\alpha\in(0,1]$是学习率；$\gamma\in[0,1)$是确定未来奖励重要性的折现因子；$\max Q_{a'}^{s'}(\tau)$是在下一个状态s'的行中的最大元素，是在s'处的最佳将来值的估计。

当个体i在第τ轮处于状态s时，它将根据Q表选择最优动作，如式（7）所示，其中$\underset{a'}{\operatorname{argmax}}\,Q_{a'}^s(\tau)$表示状态$s$行中具有最大$Q$值的动作，在每个回合结束时将$Q_a^s(\tau)$替换为$Q_a^s(\tau+1)$。

$$a(\tau)=\underset{a'}{\operatorname{argmax}}\,Q_{a'}^s(\tau) \tag{7}$$

2　3种Q学习决策机制的对比

在演化博弈中，当面对多重Nash均衡问题时，Q学习算法常使用不同的动作决策机制进行对比，常见的有ε-greedy决策机制和Boltzmann决策机制。本文实验选择了3种不同的Q学习的动作决策机制：ε-greedy决策机制，Boltzmann决策机制，Max-plus决策机制。ε-greedy决策机制作为实验的基础决策机制，虽然能够有效地描述个体学习过程中的探索状态，但是总是以相同的概率ε对所有的动作进行探索，过高估计了较差的动作策略被选择的概率；Boltzmann决策机制改善了ε-greedy决策机制的缺陷，该机制根据各动作的奖励值确定其被选择的概率，随着对环境信息的了解程度逐渐加深，随机探索的概率逐渐降低，进行最优选择的概率逐渐上升；Max-plus决策机制与前2个决策机制不同，其优点在于考虑了网络中个体的全局属性，是计算无向图模型中最大后验概率的常用方法，类似于贝叶斯网络中的置信度传播算法，节点不断向邻居节点发送局部奖励函数，且只需要对接收的信息进行求和处理，而不需要列举邻居节点的所有动作组合，使得算法易于求解。下文对使用3种决策机制的Q学习算法进行对比分析。

2.1　ε-greedy决策机制

ε-greedy决策机制是以ε的概率在每一时段选择使自己得益最大的策略。

将使用ε-greedy作为决策机制的Q学习算法记为QL，算法描述如下：

算法1　QL算法

输入　网络参数，强化学习参数

输出　网络合作水平，强化学习奖励值

1）初始化所有个体的Q表为**0**，并随机初始化每个个体的状态s。

2）对于每个回合，每个个体i在演化过程中与其一阶邻居进行博弈，并以$1-\varepsilon$的概率在当前状态s的行中选择具有最大$Q_a^s(\tau)$的动作a，或以ε的概率随机地选择一个动作a。个体i的每个邻居仅在其当前状态s的行中采取$Q_{a'}^s(\tau)$最大值的动作a作为响应。在回合结束时，个体i会根据其行为及其对手的行为获得奖励。

3）在强化学习过程中，个体i会根据式（5）更新Q_a^s值，并且状态也根据式（7）更新为$s(\tau+1)=a(\tau)$，但是个体i的邻居在当前回合不改变自身状态，$s(\tau+1)=s(\tau)$。

4）重复步骤2）和步骤3），直到系统状态达到稳定状态或演变为所需要的持续时间。

2.2　Boltzmann决策机制

ε-greedy决策机制是以ε的概率在每一时段选择使自己得益最大的策略，因而容易陷入局部最优。Boltzmann概率分布法是以P的概率在每一时段选择使自己得益最大的策略，P的计算方式如下：

$$P=\frac{\mathrm{e}^{Q_a^s/\lambda}}{\sum_{a\in A}\mathrm{e}^{Q_a^s/\lambda}} \tag{8}$$

$$\lambda=5\times0.999^t \tag{9}$$

其中：λ表示演化博弈时段（或重复博弈的次数）的函数；t是回合数，与环境相关，刻画了智能体决策的随机性。当t增大时，智能体决策的随机性也随之增大；当t减小时，决策的随机性变小。由此可见，Boltzmann概率分布法与Q学习算法结合起来具有自适应学习的能力。

将使用Boltzmann作为决策机制的Q学习算法记为BQL，算法描述如下：

算法2　BQL算法

输入　网络参数，强化学习参数

输出　网络合作水平，强化学习奖励值

1）初始化所有个体的Q表为**0**，并随机初始化每个个体的状态s。

2）对于每个回合，每个个体i在演化过程中与其一阶邻居进行博弈，根据式（8）计算在当前状态s下选择动作a的概率P，并根据概率P选择动作，个体i的每个邻居仅在其当前状态s的行中采取$Q_{a'}^s(\tau)$最大值的动作a作为响应。在回合结束时，个体i会根据

其行为及其对手的行为获得奖励。

3)在强化学习过程中,个体i会根据式(5)更新Q_a^s值,并且状态也根据式(7)更新为$s(\tau+1)=a(\tau)$,但是个体i的邻居在当前回合不改变自身状态,$s(\tau+1)=s(\tau)$。

4)重复步骤2)和步骤3),直到系统状态达到稳定状态或演变为所需的持续时间。

2.3 Max-plus决策机制

Max-plus决策机制与前2种决策机制的区别在于选择策略时考虑了个体的全局属性,以求出每个个体每轮属于自身的全局最优动作。

Max-plus决策机制通过反复地在网络图中相连节点i和j间发送局部最优消息u_i^j来近似得到最优的联合动作,因此,将Max-plus用于多Agent最优动作决策问题,则Agent_i发给邻居Agent_j的消息描述为一个局部奖励函数,其定义如式(10)所示:

$$u_i^j(a_j)=\max_{a_i}\left\{f_i^j(a_i,a_j)+\sum_{k\in\Gamma(i)/j}u_{k_i}(a_i)-c_i^j\right\} \tag{10}$$

其中:u_{k_i}表示邻居发送给Agent_i的消息;$\Gamma(i)/j$表示除了j之外的Agent_i的所有邻居的集合;f_{ij}表示联合动作的值函数,如式(11)所示;c_i^j是一个归一化值,Max-plus将其定义为所有传出的消息的平均值,如式(12)所示。

$$f_i^j=Q(s_i+a_i)+Q(s_j+a_j) \tag{11}$$

$$c_i^j=\frac{1}{|A_j|}\sum_{a_j}u_i^j(a_i^j) \tag{12}$$

在每一个时间步,定义Agent_i的每个动作的函数值如式(13)所示:

$$g(a_i)=\sum_{j\in k\in\Gamma(i)}u_j^i(a_i) \tag{13}$$

Agent_i最优动作决策如式(14)所示:

$$a^*=\operatorname*{argmax}_{a_i}\ g(a_i) \tag{14}$$

$\operatorname*{argmax}_{a'}Q_{a'}^s(\tau)$表示状态$s$行中具有最大$Q$值的动作,在每个回合结束时将$Q_a^s(\tau)$替换为$Q_a^s(\tau+1)$。

将使用Max-plus决策机制的Q学习算法记为MQL,算法描述如下:

算法3 MQL算法

输入 网络参数,强化学习参数

输出 网络合作水平,强化学习奖励值

1)初始化所有个体的Q表为**0**,并随机初始化每个个体的状态s。

2)对于每个回合,每个个体i在演化过程中与其一阶邻居进行博弈,并以$1-\varepsilon$的概率根据式(12)选择最优动作a^*,或以ε的概率随机地选择一个动作a。个体i的每个邻居仅在其当前状态根据式(12)选择最优动作a^*作为响应。在回合结束时,个体i会根据其行为及其对手的行为获得奖励。

3)在强化学习过程中,个体i会根据式(5)更新Q_a^s值,并且状态也根据式(14)更新为$s(\tau+1)=a^*$,但是个体i的邻居在当前回合不改变自身状态,$s(\tau+1)=s(\tau)$。

4)重复步骤2)和步骤3),直到系统状态达到稳定状态或演变为所需的持续时间。

3 仿真结果及分析

为了表征系统的宏观行为,本实验采用PDG作为博弈模型,引入合作频率p_c表示网络中合作者数量与所有节点数量的比值。显然,合作频率在0和1之间。当合作者的比例为0时,所有的个体都是背叛者,1表示整个网络都是合作者。本文考虑了有限节点N=100的4个网络模型:无标度网络(记为BA网络),小世界网络(记为WS网络),随机网络(记为ER网络),规则网格网络(记为RG网络)。RG网络为最简单的网络,网络中任意2个节点之间的联系遵循既定的规则。幂律分布广泛存在于物理学、生物学、社会学、经济学等众多领域中,对于许多现实世界中各节点连接数服从幂律分布的复杂网络,如互联网、社会网络等,这些网络均属于BA网络。绝大多数节点之间并不相邻,但任一给定节点的邻居却很可能彼此相邻,并且大多数任意节点都可以用较少的步或跳跃访问到其他节点,具有这种一些彼此并不相识的人可以通过一条很短的熟人链被联系在一起性质的网络为WS网络,其广泛存在于社交网络中。ER随机网络模型是个机会均等的网络模型,在计算机科学、统计物理、生命科学、通信工程等领域都有广泛应用。上述4个网络模型为经典网络模型,并广泛存在于众多领域,因此本文以这4个网络作为实验平台来验证所提模型的有效性。以合作者和叛逃者的比例相等开始,对4个网络进行了模拟,并选取BA网络为代表,在其他参数改变的情况下做进一步研究。

在每个参数组合仿真迭代的过程中,动态系统经过1 000个仿真步长逐渐达到稳定的状态,网络合作水平是50次仿真稳定结果的平均值。通过实验结果发现系统在0到100步之间变化明显,过渡状态后,系统达到稳定状态,并记录稳定状态中的合作水平。

3.1 实验结果

BA网络不同ε下合作水平如图1所示,其中,θ=1,α=0.5,γ=0.7,b=1.5。总体上来说,引入强化学习后,网络合作水平将提高,网络合作率在短时间内达到稳定状态。稳定状态下的合作水平很高,并且在稳定值附近波动。

当ε=0.02时,网络合作水平达到稳定状态的时间小于ε=0.4时,这是因为当ε较小时,策略选择的随机性较小,所以网络合作水平达到稳定状态的时间将较短。当ε较大时,策略选择的随机性将增加,达到稳定时间将更长一些。当ε=0.02时,使用QL算法

的博弈网络(记为QL-PDG)的合作水平最高,高于使用BQL算法的博弈网络(记为BQL-PDG)、使用MQL算法的博弈网络(记为MQL-PDG)和传统博弈网络(记为PDG)的合作水平。当ε=0.4时,使用3种决策机制的Q学习算法的合作水平几乎相同,与传统的博弈网络相比,合作水平显着提高。

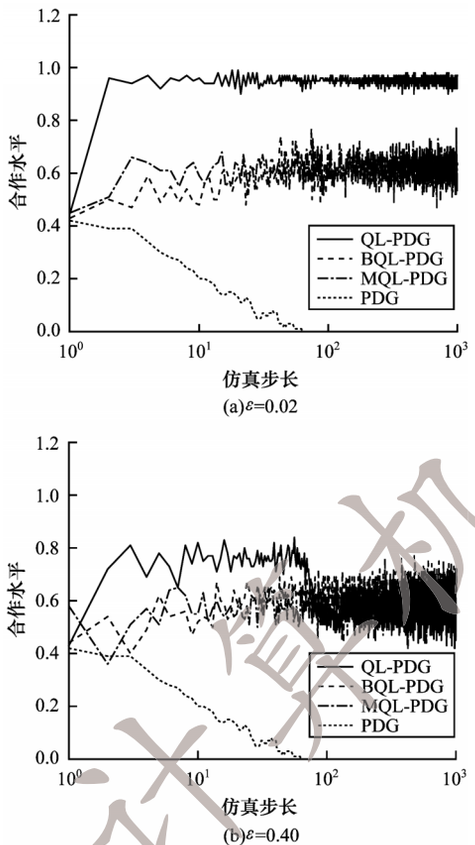

图1　BA网络不同ε下合作水平

Fig.1　The cooperation level of BA network under different ε

通过仿真结果可以发现,引入强化学习之后的网络合作水平得到提高,并且达到稳定状态之后仍会维持在比较高的水平。下文使用3种强化学习算法分别讨论和分析不同的网络模型、博弈参数的变化、强化学习参数的变化以及考虑个体全局属性的影响这4个方面对网络合作水平的影响。

3.2　具体分析

3.2.1　引入强化学习对不同网络合作水平的影响

实验使用QL-PDG与BQL-PDG、MQL-PDG以及PDG进行仿真比较,结果如图2所示,其中,θ=1,α=0.5,γ=0.7,ε=0.02。在BA、ER、SW、RG上引入不同强化学习算法之后实验结果类似。无论对于哪种网络拓扑结构,整体的合作水平都会提高,而且当ε=0.02时,3种网络对应的合作水平较高,其中QL-PDG合作水平最高,其次是BQL-PDG和MQL-PDG,最后是PDG。因此可以得出结论:在引入Q学习下加强合作的现象对于相互作用网络的拓扑结构是普遍的。以下将以BA网络为例,从3个具体的影响因素分析引入3种强化学习决策算法对整体合作水平的影响。

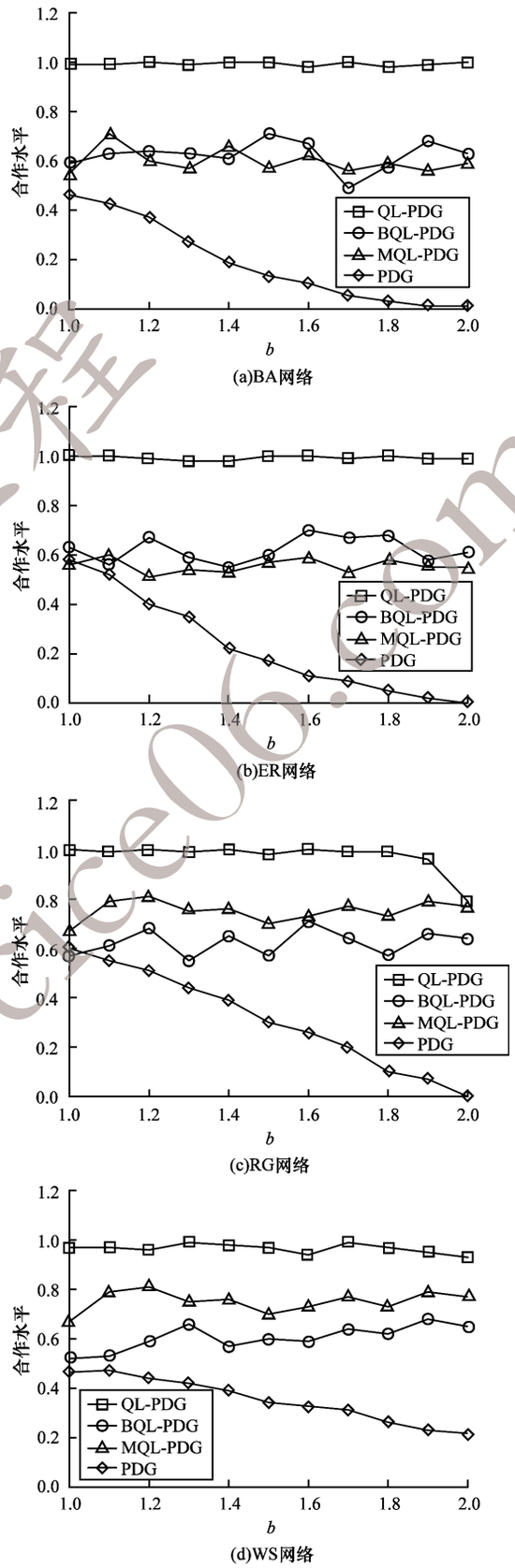

图2　不同复杂网络合作水平的比较

Fig.2　Comparison of cooperation level of different complex networks

3.2.2 网络合作水平随背叛诱惑的变化

b代表背叛诱惑。对于不同的b,实验使用3种强化学习算法的博弈网络与PDG进行比较。BA网络不同b取值情况下的网络合作水平如图3所示,其中,θ=1,α=0.5,γ=0.7,ε=0.4。随着b的增加,只有BQL-PDG的合作水平不会降低,呈现出随b的增加合作水平上下波动的变化,QL-PDG和MQL-PDG的合作水平略有下降,而PDG的合作水平随b的增加而急剧下降为0。从结果可以看出,引入强化学习之后系统对于背叛诱惑的抵御能力更强,稳定性也更强。

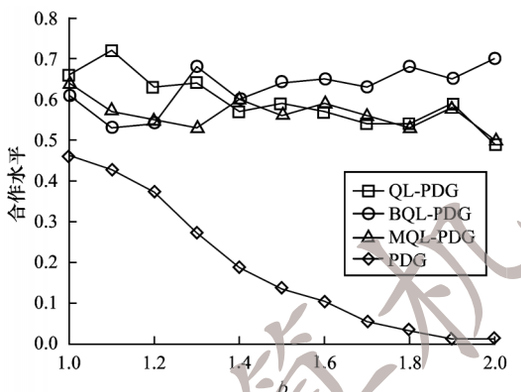

图3 BA网络不同b下的合作水平

Fig.3 The cooperation level of BA network under different b

文献[32]研究了动态变化的交互范围对系统合作演化的影响,结果证明调整自身交互范围可以显著促进合作,并能在非常高的背叛诱惑下维持高水平的合作。本文通过使用强化学习算法控制节点的博弈过程,即使在背叛诱惑持续增大的情况下,仍能维持较为稳定的合作水平。

3.2.3 网络合作水平随背叛诱惑的变化

实验使用了3种强化学习算法来观测不同的α对网络合作水平的影响,如图4所示,其中,θ=1,γ=0.7,ε=0.02,α表示个体的学习率。可以看出,3种强化学习算法对于不同的α对应的网络合作水平的变化趋势总体是相同的,学习率越低,网络合作水平越高。BQL-PDG和MQL-PDG对于不同的α,合作程度显示出细微的变化,并且随着b的变化,合作程度变化不大;对于QL-PDG,对于不同的α,当$\alpha\geqslant$0.5且当b>1.4时,合作水平随着b的变化而呈现急剧下降趋势。

γ表示折扣率(个体为未来奖励的贪婪程度)。不同γ对合作水平的影响如图5所示,其中,θ=1,α=0.5,ε=0.02。可以看出,对于3种强化学习决策算法,折扣率越高,网络合作水平越高。QL-PDG的变化趋势最明显,MQL-PDG的变化趋势表现最弱。

θ表示收益的异质性。θ对合作水平的影响如图6所示,其中,α=0.5,γ=0.7,ε=0.02。可以看出,无论对于哪种强化学习决策算法,θ为中间水平时,网络的合作水平最高。BQL-PDG和MQL-PDG对于不同的θ,合作程度显示出细微的变化,随着$b\in$[1.0,2.0]的变化,中等水平的θ对应的合作水平最高。对于QL-PDG本文做了更细微的研究,在$b\in$[1.0,1.1]时,实验中观察了不同的θ随着b的变化仍然呈现出中等水平的θ对应的合作水平最高的趋势。因此,在强化学习中适当地调节收益异质性可以提高网络合作水平,过高或者过低的异质性都会使网络的合作水平降低。

文献[32]研究了在RG网络和ER网络上,空间公共物品博弈中历史收益诱导的异质性对群体合作的影响,实验结果表明,异质性越强,群体合作水平越高。与其不同,本文研究的强化学习决策算法中群体合作水平并不与合作水平成正比,中等水平的异质性对应的合作水平最高。

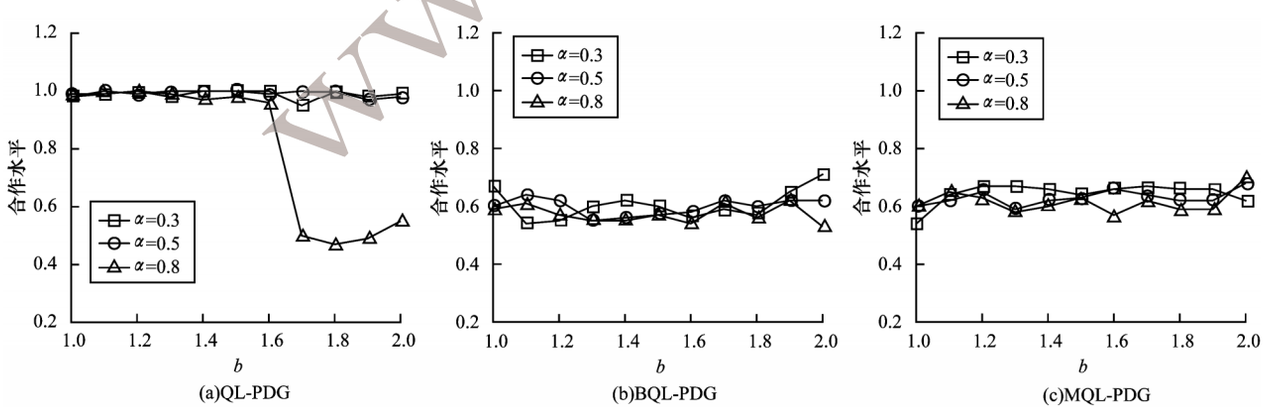

图4 不同α对合作水平的影响

Fig.4 The influence of different α on cooperation level

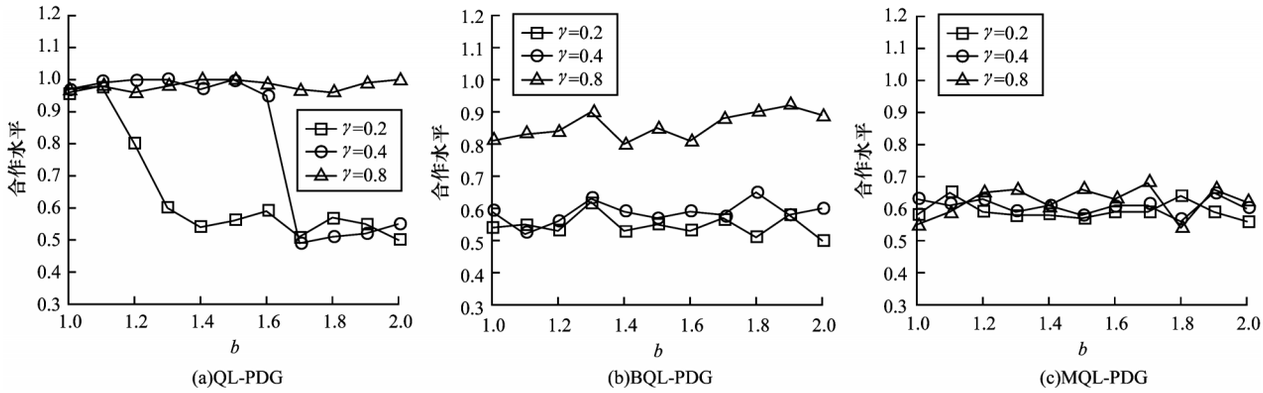

图5　不同γ对合作水平的影响

Fig.5　The influence of different γ on cooperation level

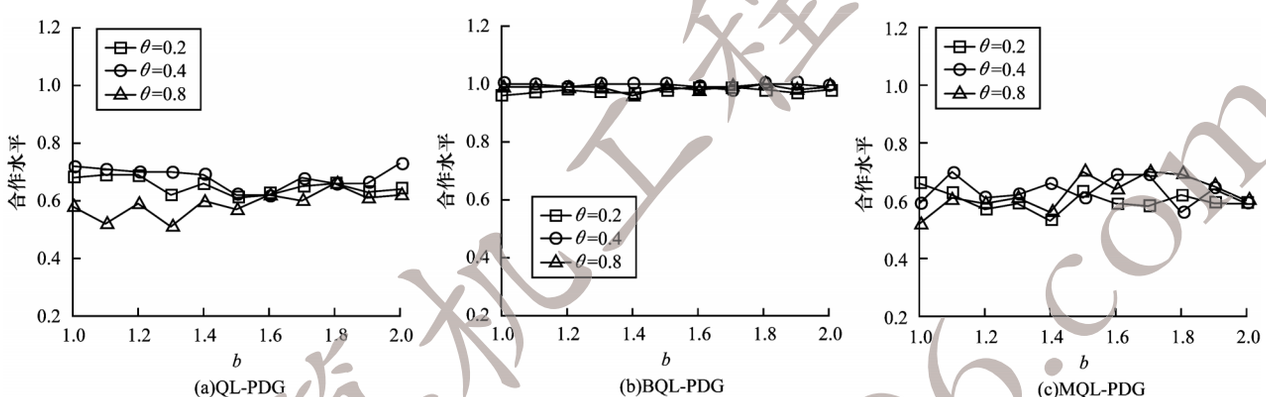

图6　不同θ对合作水平的影响

Fig.6　The influence of different θ on cooperation level

3.2.4　个体的全局属性对网络平均收益的影响

BA网络3种算法的平均奖励如图7所示,其中,θ=1,α=0,γ=0.7,b=1.5。当ε较小时,QL-PDG的平均收益最高,其次是MQL-PDG,BQL-PDG的平均收益最低;当ε较大时,MQL-PDG就会超过QL-PDG成为平均收益最高的。这是因为,当ε较小时个体的探索率较小,个体就会最直接自私地选择对于自身而言最优的策略,所以此时QL-PDG的平均收益占据最高位;而当个体的探索率增大时,个体进行决策的随机性就会增大,此时个体考虑自身的全局属性做出的选择的最优策略就是全局最优策略,所以此时MQL-PDG的平均收益是最高的。由此可以得出结论:当个体的探索率较大时,使用考虑个体属性的Max-plus决策机制的Q学习算法能够提高网络的平均收益。

文献[33]研究了在普适社交网络的环境中,所有个体利用强化学习方法中的Q学习算法对囚徒困境和古诺模型中的社交行为进行学习的效果,实验结果证明,学习的个体奖励值在多次学习后达到收敛,且有效地解决了囚徒困境问题。本文实验中使用3种决策机制的Q学习算法均能实现均衡,本文在此基础上研究了不同探索率下的奖励变化情况。

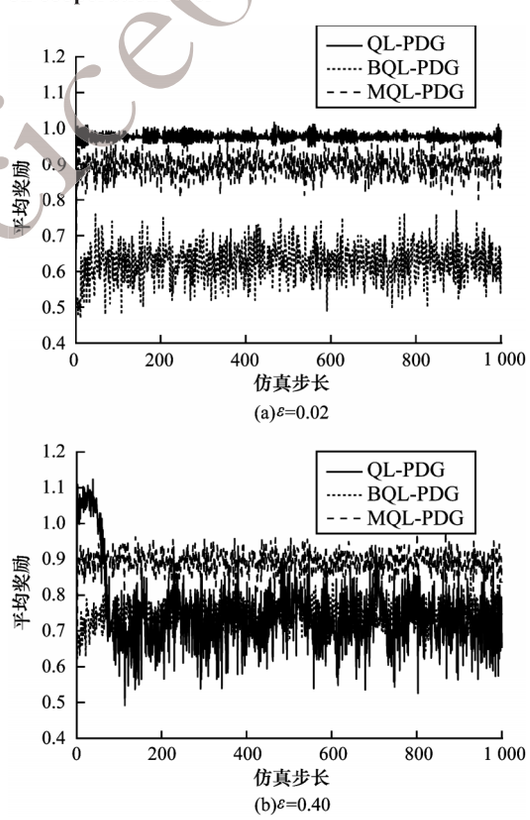

图7　BA网络3种算法的平均奖励

Fig.7　The average reward of three algorithms on BA network

4 结束语

本文引入3种Q学习决策机制作为个体的策略选择机制，引入用于调整个体收益的异质性的参数θ，针对不同的网络类型、不同的博弈模型参数和不同的强化学习参数进行对比实验。实验结果表明：在复杂网络演化博弈中引入强化学习作为决策机制，能够提高网络的合作水平；对于不同的网络拓扑结构，引入强化学习合作水平得到提高的结果也是普遍的；当Q学习应用于个体决策时，较低的学习率和较高的折扣率将促进网络个体间的合作行为；在调节个体收益的异质性时，中等水平的个体收益的异质性会提高网络的合作水平，太低或太高都会降低网络合作水平；此外，当个体以较高的探索率进行策略选择时，将考虑个体全局属性的Max-plus作为决策机制的网络获得的平均收益最高。

本文仅在4种经典网络上验证了所提模型的有效性，后续将继续研究模型在新的和更加复杂的网络上的应用和表现。本文针对强化学习算法的应用对网络合作水平的影响进行研究，后续可以研究其他强化学习算法和参数对网络合作水平的影响。此外，还可以将深度学习算法与强化学习算法结合应用到演化博弈中，开发更加智能的决策模式。

参考文献

[1] VIOSSAT Y. Correlated equilibria, evolutionary games and population dynamics[D]. Paris, France: Ecole Polytechnique, 2005.

[2] 张青青，汤红波，游伟，等. 基于演化博弈的NFV拟态防御架构动态调度策略[J]. 计算机工程，2022，48(4)：30-38，49.
ZHANG Q Q, TANG H B, YOU W, et al. Dynamic scheduling strategy of NFV mimic defense architecture based on evolutionary game[J]. Computer Engineering, 2022, 48(4): 30-38, 49. (in Chinese)

[3] NOWAK M. Five rules for the evolution of cooperation[J]. Science, 2006, 314(5805): 1560-1563.

[4] SANTOS F C, PACHECO J M, LENAERTS T. Evolutionary dynamics of social dilemmas in structured heterogeneous populations[J]. Proceedings of the National Academy of Sciences, 2006, 103(9): 3490-3494.

[5] WU Z X, XU X J, HUANG Z G, et al. Evolutionary prisoner's dilemma game with dynamic preferential selection[J]. Physical Review E, 2006, 74(2): 021107.

[6] GÓMEZ-GARDEÑES J, CAMPILLO M, FLORÍA L M, et al. Dynamical organization of cooperation in complex topologies[J]. Physical Review Letters, 2007, 98(10): 108103.

[7] KESSLER D A, SANDER L M. Fluctuations and dispersal rates in population dynamics[J]. Physical Review E, 2009, 80(4): 041907.

[8] HAUERT C, DE MONTE S, HOFBAUER J, et al. Volunteering as red queen mechanism for cooperation in public goods games[J]. Science, 2002, 296(5570): 1129-1132.

[9] LANGER P, NOWAK M A, HAUERT C. Spatial invasion of cooperation[J]. Journal of Theoretical Biology, 2008, 250(4): 634-641.

[10] PERC M. Transition from Gaussian to Levy distributions of stochastic payoff variations in the spatial prisoner's dilemma game[J]. Physical Review E, 2007, 75(2): 022101.

[11] CHEN X, WANG L. Promotion of cooperation induced by appropriate payoff aspirations in a small-world networked game[J]. Physical Review E, 2008, 77(1): 017103.

[12] 刘永奎. 复杂网络及网络上的演化博弈动力学研究[D]. 西安：西安电子科技大学，2010.
LIU Y K. Research on complex networks and evolutionary game dynamics on networks[D]. Xi'an: Xidian University, 2010. (in Chinese)

[13] SZABÓ G, FÁTH G. Evolutionary games on graphs[J]. Physics Reports, 2007, 446(4/5/6): 97-216.

[14] ROCA C P, CUESTA J A, SÁNCHEZ A. Imperfect imitation can enhance cooperation[J]. Europhysics Letters, 2009, 87(4): 48005.

[15] CRESSMAN R, DASH A T. Evolutionarily stable strategies with two types of player I. Two-species haploid or randomly mating diploid[J]. Journal of Applied Probability, 1985, 22(1): 1-14.

[16] TAYLOR P D, JONKER L B. Evolutionary stable strategies and game dynamics[J]. Mathematical Biosciences, 1978, 40(1/2): 145-156.

[17] SZOLNOKI A, PERC M. Promoting cooperation in social dilemmas via simple coevolutionary rules[J]. The European Physical Journal B, 2009, 67(3): 337-344.

[18] DU J M, WU B, ALTROCK P M, et al. Aspiration dynamics of multi-player games in finite populations[J]. Journal of the Royal Society, Interface, 2014, 11(94): 20140077.

[19] DI CHIO C, DI CHIO P, GIACOBINI M. An evolutionary game-theoretical approach to particle swarm optimisation[C]//Proceedings of 2008 Conference on Applications of Evolutionary Computing. New York, USA: ACM Press, 2008: 575-584.

[20] ZHANG J L, ZHANG C Y, CHU T G, et al. Resolution of the stochastic strategy spatial prisoner's dilemma by means of particle swarm optimization[J]. PLoS One, 2011, 6(7): e21787.

[21] CHEN Y S, YANG H X, GUO W Z, et al. Promotion of cooperation based on swarm intelligence in spatial public goods games[J]. Applied Mathematics and Computation, 2018, 320: 614-620.

[22] Reinforcement learning: an introduction[J]. IEEE Transactions on Neural Networks, 2005, 16(1): 285-286.

[23] SILVER D, HUBERT T, SCHRITTWIESER J, et al. Mastering chess and Shogi by self-play with a general reinforcement learning algorithm[EB/OL]. [2022-03-01]. https://arxiv.org/abs/1712.01815.

[24] SUTTON R S. Learning to predict by the methods of temporal differences[J]. Machine Learning, 1988, 3(1): 9-44.

[25] WATKINS C J C H, DAYAN P. Technical note: Q-learning[J]. Machine Learning, 1992, 8(3): 279-292.

（下转第114页）

(上接第106页)

[26] VAN HASSELT H, GUEZ A, HESSEL M, et al. Learning values across many orders of magnitude[EB/OL]. [2022-03-01]. https://arxiv. org/abs/1602. 07714.
[27] NING J, MA S. Evolution of cooperation in the snowdrift game among mobile players with random-pairing and reinforcement learning[J]. Physica A: Statistical Mechanics and Its Applications, 2013, 392(22): 5700-5710.
[28] ZHANG S J, WANG X Q, CHENG X F, et al. Single crystal growth, structural characterization, thermal and optical properties of a novel organometallic nonlinear optical crystal: $MnHg(SCN)_4(C_2H_5NO)_2$[J]. Physica B: Condensed Matter, 2010, 405(4): 1071-1080.
[29] ZHANG H F, WU Z X, WANG B H. Universal effect of dynamical reinforcement learning mechanism in spatial evolutionary games[J]. Journal of Statistical Mechanics: Theory and Experiment, 2012, 2012(6): P06005.
[30] 徐琳,赵知劲. 基于分布式协作Q学习的信道与功率分配算法[J]. 计算机工程,2019,45(6):160-164,174.
XU L, ZHAO Z J. Channel and power allocation algorithm based on distributed cooperative Q learning[J]. Computer Engineering, 2019, 45(6): 160-164, 174. (in Chinese)
[31] 韩晨,牛英滔. 基于分层Q学习的联合抗干扰算法[J]. 计算机工程,2019,45(5):279-284.
HAN C, NIU Y T. Joint anti-jamming algorithm based on hierarchical Q learning[J]. Computer Engineering, 2019, 45(5): 279-284. (in Chinese)
[32] ZHANG L, XIE Y, HUANG C, et al. Heterogeneous investments induced by historical payoffs promote cooperation in spatial public goods games[J]. Chaos Solitons & Fractals, 2020, 133: 109675.
[33] ZHANG Y, SONG B, ZHANG P. Social behavior study under pervasive social networking based on decentralized deep reinforcement learning[J]. Journal of Network and Computer Applications, 2017, 86: 72-81.

编辑 金胡考