Operations Research and Fuzziology 运筹与模糊学, 2024, 14(1), 131-139
Published Online February 2024 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2024.141012

Hans汉斯

时间尺度多样性结合强化学习促进囚徒困境博弈中的合作

李卓君[1]，王书博[2]，杨　梅[1]，程贞敏[1]

[1]贵州大学数学与统计学院，贵州　贵阳
[2]华润电力投资有限公司中西分公司，河南　郑州

收稿日期：2023年11月14日；录用日期：2023年12月4日；发布日期：2024年2月18日

摘　要

演化博弈论为解决社会困境提供了关键框架，并且不一定局限于统一时间尺度，同时强化学习已被证明是研究博弈论中策略更新动态和智能体学习过程的有效方法。因此，本文研究了时间尺度机制结合自我关注Q学习算法对空间囚徒困境博弈中合作的影响。具体来说，博弈交互和策略更新具有不同的时间尺度，时间尺度多样性影响策略的概率更新公式，并且将自我关注Q学习算法当作策略更新规则。数值结果表明，在这样的框架下，能够显著地促进合作。最后，分析了影响Q学习的参数以及在不同的初始设置下验证了机制的鲁棒性。

关键词

社会困境，合作，Q学习，时间尺度

Time Scale Diversity Combined with Reinforcement Learning to Promote Cooperation in Prisoner's Dilemma Game

Zhuojun Li[1], Shubo Wang[2], Mei Yang[1], Zhenmin Cheng[1]

[1]School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou
[2]Zhongxi Branch, China Resources Power Investment Co., Ltd., Zhengzhou Henan

Received: Nov. 14th, 2023; accepted: Dec. 4th, 2023; published: Feb. 18th, 2024

Abstract

Evolutionary game theory provides a key framework for solving social dilemmas, and it is not

文章引用: 李卓君, 王书博, 杨梅, 程贞敏. 时间尺度多样性结合强化学习促进囚徒困境博弈中的合作[J]. 运筹与模糊学, 2024, 14(1): 131-139. DOI: 10.12677/orf.2024.141012

necessarily limited to a unified time scale. At the same time, reinforcement learning has been proven to be an effective method to study the strategy update dynamics and agent learning process in game theory. Therefore, this paper studies the influence of time scale mechanism combined with self-focused *Q*-learning algorithm on cooperation in spatial prisoner's dilemma game. Specifically, game interaction and strategy update have different time scales. The diversity of time scales affects the probability update formula of the strategy, and the self-focused *Q*-learning algorithm is used as the strategy update rule. The numerical results show that under such a framework, cooperation can be significantly promoted. Finally, the parameters affecting *Q*-learning are analyzed and the robustness of the mechanism is verified under different initial settings.

Keywords

Social Dilemma, Cooperation, *Q*-Learning, Time Scale

1. 引言

合作行为广泛存在于自然系统和人类社会中，并且对种群进化和社会繁荣起到重要促进作用。但由于自私个体总是追求自身利益最大化，这与牺牲自身利益提高集体收益相矛盾，故而形成了社会困境。因此，研究自私个体之间合作行为的出现和维持成为一个有趣且具有挑战性的问题，引起了学者们的极大关注[1] [2] [3] [4] [5]。囚徒困境博弈作为解释合作行为的一般隐喻，已经得到了很好的研究[6]。在这个基本模型中，两个参与者应该同时选择合作(C)或背叛(D)策略。他们的收益确定如下：如果双方合作，每个玩家将获得奖励 R，如果双方背叛，则获得惩罚 P。如果合作者遇到背叛者，背叛者获得诱惑收益 T，合作者获得收益 S。并且需满足 $T>R>P>S$ 和 $2R>T+S$，以确保博弈的性质[7]，这意味着无论对手的选择如何，背叛都是最好的策略。其结果是背叛行为在所有参与者中蔓延，称为社会困境。

解决这一困境的方法有很多种，主要可以分为两类。第一类是引入空间结构，让合作者可以形成紧凑的集群，以防止背叛者的入侵[8]。第二类是引入一些促进合作的机制[9] [10] [11]，如针锋相对(TFT，记忆)，奖惩[12] [13] [14]，自愿参与[15] [16]，声誉[17]，愿望[18]和环境[19]。不仅如此，通常假设所有智能体在每一轮交互后都会同时更新他们的策略，然而，现实中博弈交互时间尺度与策略选择时间尺度并不总是相同的，研究表明，如果考虑时间尺度多样性，合作的演变可能会发生变化[19] [20] [21]，此研究尽管取得了一些进展，但这些研究通常是在费米策略更新规则的框架下进行的。也就是说，每个玩家选择一个邻居后以相同的概率或偏好学习其策略，这可能会忽略环境的影响。最近，强化学习(即 Q 学习算法)得到了很好的研究，并被纳入演化博弈中，以了解合作行为的出现。与直觉相反，强化学习未能促进囚徒困境博弈中的合作。这是因为这些研究主要是在混合良好的人群中进行的，没有考虑任何其他机制。

受到所有这些创新的启发，一个有趣的问题浮出水面：如果我们在方格网络上同时结合时间尺度多样性和强化学习，合作水平是否会得到提升？在这项工作中，我们考虑了囚徒困境博弈中每一轮交互结束后受时间尺度影响更新概率公式决定要不要更新策略，策略更新规则是自关注 Q 学习。数值模拟结果表明，时间度多样性可以促进合作。因此，关于时间尺度多样性与强化学习的影响的研究对于进一步理解人类合作变得有意义。

2. 预备知识

博弈论是研究人与人互动的数学框架[22] [23]。在两人两策略博弈中，两个智能体应该同时选择合作(*C*)或背叛(*D*)策略。他们的收益确定如下：如果双方合作，每个智能体将获得奖励 *R*，如果双方背叛，则获得惩罚 *P*。如果合作者遇到背叛者，背叛者获得诱惑收益 *T*，合作者获得收益 *S*。两人两策略博弈可以由矩阵给出，如下所示：

$$M=\begin{pmatrix} R,R & S,T \\ T,S & P,P \end{pmatrix},$$

其中第一个元素表示行智能体的收益，第二个元素表示列智能体的收益。这四个参数的大小决定了博弈的类型和纳什均衡的位置。纳什均衡是指一套策略，没有人可以通过单方面改变自己的策略来获得更高的回报。如果参数满足 $T>R>P>S$，博弈类型为囚徒困境博弈，纳什均衡为 (D,D)；如果参数满足 $T>R>S>P$，博弈类型为雪堆博弈，纳什均衡为 (C,D) 和 (D,C)；如果参数满足 $R>T>P>S$，博弈类型为猎鹿博弈，其纳什均衡为 (C,C) 和 (D,D)；如果参数满足 $R>T$ 和 $S>P$，博弈类型为协调博弈且纳什均衡为 (C,C)。

Q-learning 是强化学习的基本算法之一[24] [25]，用于表征多智能体系统中由联合行动触发的决策过程。*S* 为状态集，$A_1\times\cdots\times A_n$ 表示具有 *n* 个代理的系统中的动作集。学习过程 $S\times A_1\times\cdots\times A_n\to R$ 意味着状态信息由所有智能体共享，并受联合行动的影响。*Q*-learning 定义了一个 *Q* 表，其中每个元素(状态–动作对)代表在状态 *S* 中执行行动 *a* 的累计奖励。给定一个代理 *i*，它在第 *t* 轮的状态 *S* 下执行动作 $a_{ij}\in A_i$。设 *F* 表示 *i* 通过与环境交互获得的收益。*Q*-learning 是一种自适应值迭代方法，每个智能体在基于当前时间步的 *Q* 值 $Q_{s,a_j}(t)$ 和采取动作 a_j 后的状态 s' 下的 *Q* 值 $Q_{s',a'}(t)$ 估计其下一步的状态–动作值 $Q_{s,a_j}(t+1)$

$$Q_{s,a_j}(t+1)=(1-\alpha)Q_{s,a}(t)+\alpha\left[F+\gamma\max\left\{Q_{s',a'}(t)\right\}\right],$$

其中 α 代表学习率，γ 表示折扣因子。$\max\left\{Q_{s',a'}(t)\right\}$ 为下一状态 s' 行下 *Q* 表的最大值，这是对状态 s' 下最佳未来值的估计。该函数表示在状态 *s* 下采取行动 a_j 有多好。

3. 基于自关注 *Q*-learning 的演化博弈模型

考虑一个空间囚徒困境博弈(PDG)，玩家占据一个 $L\times L$ 具有周期性边界的方形格子。在不失一般性的情况下，研究了弱 PDG [7]，其中收益矩阵

$$M=\begin{pmatrix} 1 & 0 \\ b & 0 \end{pmatrix}, \tag{1}$$

其中 *b* 表示背叛诱惑。

每一步中，每个玩家只能与他们的四个邻居(冯诺依曼邻居)互动以获得累积的收益。累积收益

$$F_i=\sum_{j\in N_i} I_i^T M I_j, \tag{2}$$

其中 I_i 和 I_j 为玩家 *i* 和对手 *j* 选择的策略，*M* 为收益矩阵，N_i 为玩家 *i* 的邻居。

博弈交互结束后，根据更新概率公式[19]决定要不要更新

$$P_i=\frac{1}{1+\eta\cdot F_i}, \tag{3}$$

其中用 $\eta\geq 0$ 来刻画时间尺度，当 $\eta=0$ 时，$P_i=1$，即每次博弈交互后一定会更新策略，代表没有时间尺度多样性的常规情形；当 $\eta>0$，时间尺度多样性随着的增加而增加。

至于策略更新过程，玩家采用自适应 Q 学习算法，这与 η 涉及与环境交互的传统 Q 学习方法不同。在这样一个新颖的框架下，玩家 i 根据他们的经验用最大的 Q 值更新策略，而不考虑邻居的策略(这在以前的 Q 学习方法中是必需的)。Q 值由 Q 表定义，用于记录不同动作在不同状态下的相对效用。在下文中，状态集 S 和动作集 A 是相同的，即 $\{C,D\}$ 带有状态(行)和操作(列)的 Q 表提供如下：

$$Q(t)=\begin{bmatrix} Q_{C,C}(t) & Q_{C,D}(t) \\ Q_{D,C}(t) & Q_{D,D}(t) \end{bmatrix}, \tag{4}$$

其中 $Q_{s,a}(t)$ 表示状态为 s 且动作 a 在时间步长 t 处玩家的 Q 值。为简单起见，但不失通用性，s 表示玩家当前的状态，a 表示可能采取的行动。当玩家与邻居交互完后，Q 表会根据以下等式[26]进行更新：

$$Q_{s,a}(t+1)=(1-\alpha)Q_{s,a}(t)+\alpha\left[F(t)+\gamma\max_{a'}\left\{Q_{s',a'}(t)\right\}\right], \tag{5}$$

其中 $\alpha\in(0,1]$ 为学习率，$F(t)$ 为当前时刻收益，并且 $\gamma\in[0,1)$ 表示决定更新策略的未来奖励比例的折扣系数，$\max_{a'}\left\{Q_{s',a'}(t)\right\}$ 为下一状态 s' 行下 Q 表的最大值。为了避免 Q 值收敛到局部最优值，在每一轮更新过程中使用 ε-贪婪探索法则。玩家以 ε 的概率随机行动或者以 $1-\varepsilon$ 概率选取 Q 表中的最大值行动。这样高收益的行动将得到加强。

总的来说，整个算法流程如下：一、最初，所有玩家被分配到一个大小为 $Z=L\times L$ 网络的格子上，随机分配一个初始状态，即玩家以相同的概率选择合作或者背叛。二、由于玩家最初不知道博弈或者环境，因此 Q 表初始化为零。三、每个回合中，随机选择一个状态 s 和动作 a 的玩家 i，玩家 i 与周围四个邻居交互并根据公式(2)获得收益，根据公式(5)更新 Q 值。四、获得收益后，根据公式(3)决定要不要更新策略，若要，根据所选动作 a，将玩家 i 的状态从当前 s 更新为 s'；否则这一轮交互后依旧保留当前状态 s。五、重复程序三和程序四直至 T 步停止。

本文借助MATLAB进行仿真，演化博弈根据蒙特卡洛模拟过程向前迭代，方形格子大小为200 × 200。在实验过程中观察到，经过 3500 次迭代即可达到稳定状态，合作水平是从稳定状态的平均值得到的，因此，取 5000 次迭代最后 500 步稳态值的平均值。同时，为了保证较高的精度，进行了 20 次独立实验。

4. 结果和分析

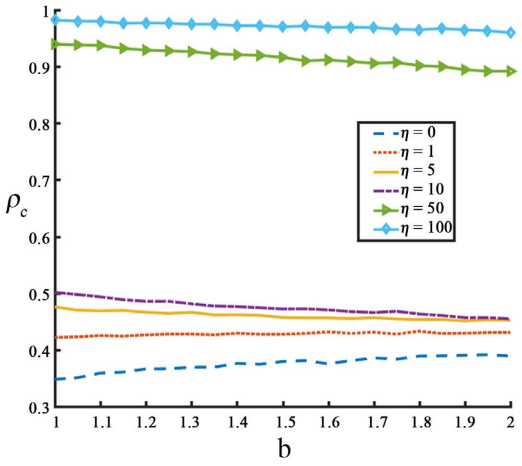

Figure 1. The effect of betrayal temptation b on the proportion of cooperators ρ at different η values. The remaining parameters $\alpha=0.8,\gamma=0.8,\varepsilon=0.02$

图 1. 不同 η 值时背叛诱惑 b 对合作者比例 ρ 的影响。其余参数 $\alpha=0.8,\gamma=0.8,\varepsilon=0.02$

我们先讨论强化学习解决传统 PDG ($\eta=0$)的情形。从图 1 中我们可以看到，在没有时间尺度多样性时，自适应 Q 学习可以让合作者比例维持在一个可观的数量上，而不像在传统的费米更新规则中消失。因此，Q 学习可以保持部分自私个体选择合作，但依然无法动摇背叛者的统治。接着我们讨论时间尺度多样性在 Q 学习下对合作者比例的影响。当考虑时间尺度机制时，与传统情形相比，合作者的占比随着 η 的增加而提高。当 $\eta=100$ 时，完全由合作者主导。简言之，当同时考虑时间尺度机制与 Q 学习，尽管背叛诱惑大，依然能极大的促进玩家选择合作行为。

接下来从宏观角度观察当 η 取不同情况下，合作率是如何随时间演变的。如图 2 所示，大概到 4000 步时所有情况都能达到稳态，并且没有时间尺度多样性时($\eta=0$)合作水平是最低的，背叛者数量更多，随着 η 逐渐增高，合作水平逐渐增强，当 $\eta\geq 50$ 之后合作者优势愈发明显，甚至能统治全局。由此可见，当时间尺度参数达到一定阈值时，合作行为能主导全局。

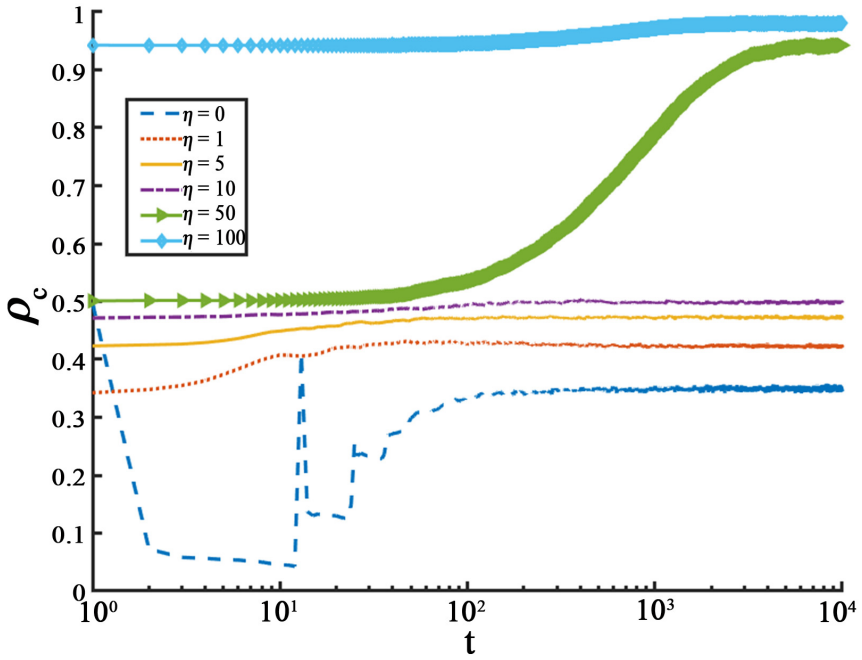

Figure 2. The evolution diagram of cooperation ratio ρ with time t when η takes different values. The remaining parameters $\alpha=0.8,\gamma=0.8,\varepsilon=0.02,b=1.02$

图 2. η 取不同值时合作比例 ρ 随时间 t 演变图。其余参数 $\alpha=0.8,\gamma=0.8,\varepsilon=0.02,b=1.02$

为了从微观角度理解促进合作的机制，图 3 描绘了动态平衡下方形网络的演化快照图。我们将背叛诱惑收益值固定 $b=1.02$。当时间尺度多样性较小时($\eta=1$，上图)，随着演化步骤的增加，这些玩家逐渐分散开，但背叛者占上风；当 $\eta=20$ 时(中图)，每个玩家都可以根据适应度适度调整自己更新策略的概率。最终，合作者和背叛者几乎持平。当 $\eta=100$ 时，合作者逐渐主导背叛者，最终合作者几乎占据了整个网络，只有少数分散的背叛者。每个玩家能在一次交互后将高收益的行为保持住，不更新自己的策略。这样的结果表明，若中心玩家为合作者时，合作者就能维持更长的时间，背叛者的数量就会减少；而若中心玩家是背叛者时，会从它的周围合作邻居那获得高额回报，故而与背叛者相邻的合作者倾向于在随后几轮采用背叛策略，反过来又减少了背叛者的收益，Q 学习可以引入目前人群中不存在的新策略，所以在这种周围都是背叛者状态中合作者仍然可以出现，从而导致背叛策略的短期存在。

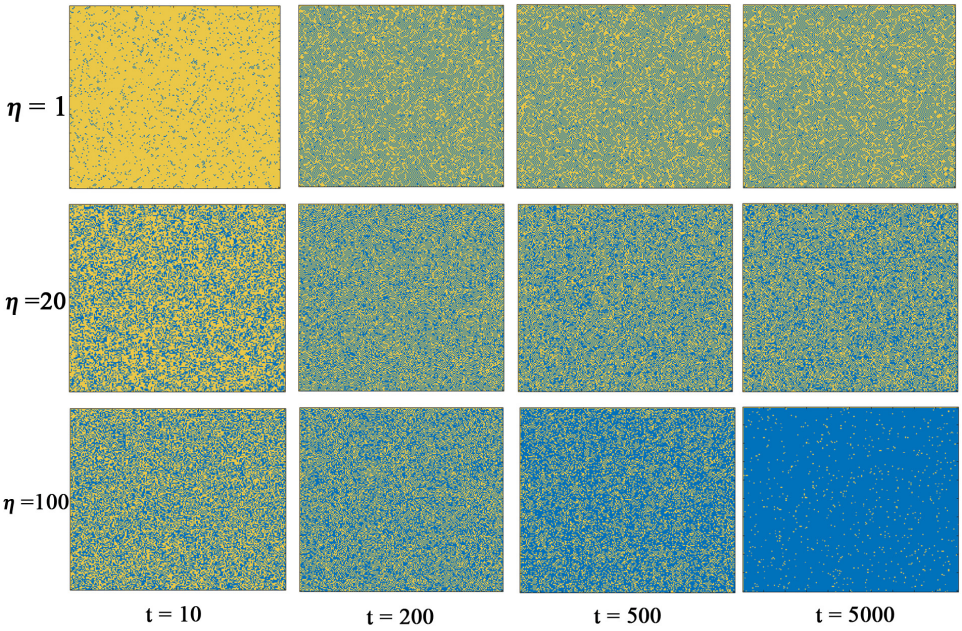

Figure 3. Evolutionary snapshots of cooperators (blue) and defectors (yellow) at different time scales and different iteration steps. From left to right, there are 10, 200, 500 and 5000 MCs, and the upper, middle and lower represent the cases of $\eta = 1, 20, 100$ respectively. The remaining parameters $\alpha = 0.8, \gamma = 0.8, \varepsilon = 0.02, b = 1.02$

图 3. 不同时间尺度和不同迭代步下合作者(蓝色)和背叛者(黄色)的演化快照图。从左至右依次为 10、200、500、5000 个 MC 的情形，上、中、下分别代表 $\eta = 1, 20, 100$ 的情形。其余参数 $\alpha = 0.8, \gamma = 0.8, \varepsilon = 0.02, b = 1.02$

上述我们已经证明了时间尺度机制结合强化学习如何有效的促进合作。接下来我们来探讨一下 Q 学习的自身参数($\alpha - \gamma$)在结合时间尺度机制下又会如何影响合作的演变。如图 4 所示，左右两图分别代表传统情形和有时间尺度多样性的情形。从图中可以看出来，随着学习率 α 升高，合作者比例也会随着升高，说明较快的遗弃旧值可以加强合作。γ 代表玩家远见水平的折扣系数(小 γ 意味着玩家更关注当前的收益)，左图中可以看出 γ 过大过小都是不利于合作的实现的，但当考虑时间尺度多样性后(右图) γ 最优取值范围更广一些，都反应了适度的关注未来的奖励才是明智的。

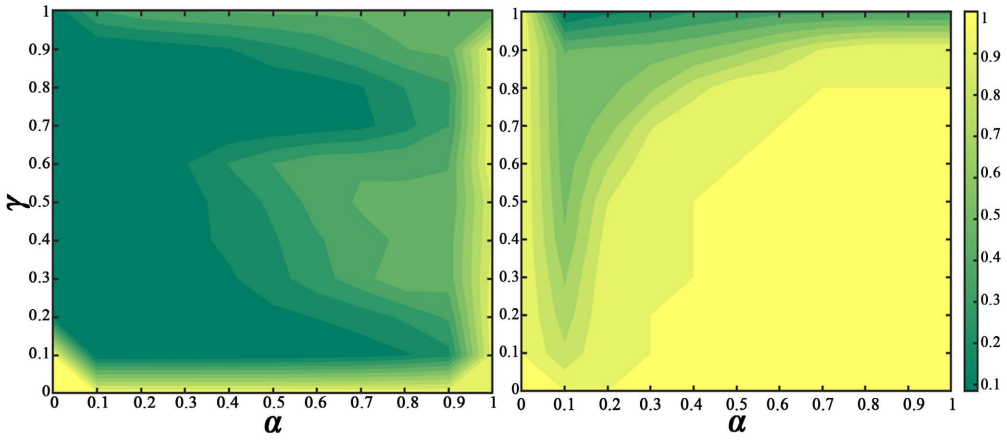

Figure 4. The two-dimensional parameters of $\alpha - \gamma$ jointly affect the contour map of the proportion of collaborators. Left Figure: $\eta = 0$; Right Figure: $\eta = 100$, The remaining parameters $\varepsilon = 0.02, b = 1.02$

图 4. $\alpha - \gamma$ 二维参数共同影响合作者比例的等高线图。左图：$\eta = 0$，右图：$\eta = 100$，其余参数 $\varepsilon = 0.02, b = 1.02$

最后，我们研究一下初始设置是否会影响合作的稳定性(即方法的鲁棒性)。为此，我们固定演化早期合作者的占比分别为总人口的 10%、30%、50%、70%、90%，如图 5，图 6 所示，图 5 为传统情形($\eta=0$)，图 6 为有时间尺度多样性($\eta=5$)。由于玩家喜欢学习最大收益策略以避免被他人利用，合作水平几乎不受初始分布的限制，并且初始分布对于有无时间尺度机制都是鲁棒的。从这个角度来看，如果玩家只学习最大收益的动作，合作可以达到相同合作水平的稳定状态。

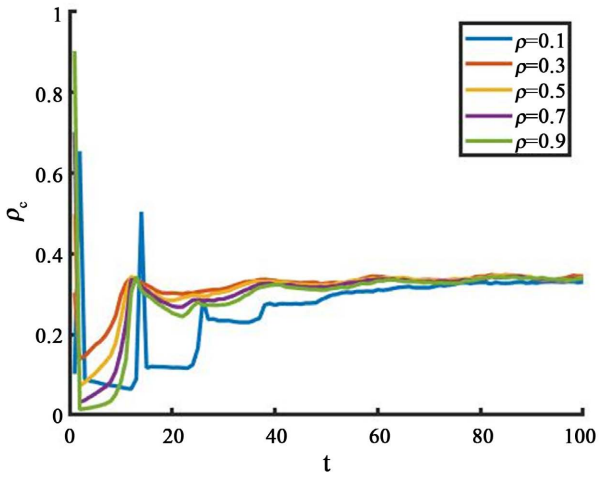

Figure 5. The evolution of cooperation rate with time under different initial proportion of collaborators ($\eta=0$). The remaining parameters $\alpha=0.8,\gamma=0.8,\varepsilon=0.02,b=1.02$

图 5. 不同合作者初始占比下合作率随时间的演化情况($\eta=0$)。其余参数 $\alpha=0.8,\gamma=0.8,\varepsilon=0.02,b=1.02$

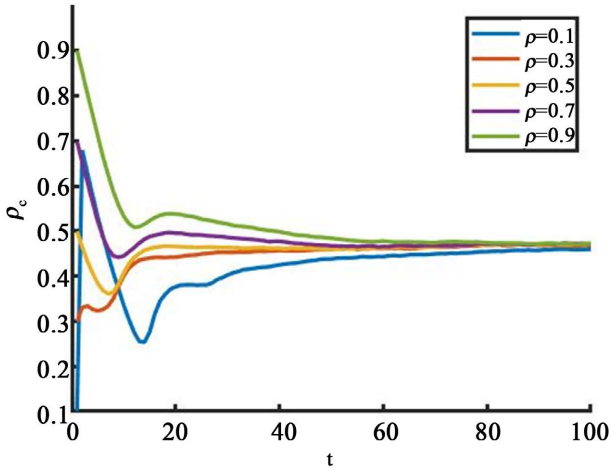

Figure 6. The evolution of cooperation rate with time under different initial proportion of collaborators ($\eta=5$)

图 6. 不同合作者初始占比下合作率随时间的演化情况($\eta=5$)

5. 结论

本文未考虑随机性因素，同时企业之间隐瞒信息导致在信息不对称下做决策是常有的事情，将是以后需要继续挖掘的课题。本文表明，将时间尺度机制与强化学习策略更新规则相结合能够极大的促进合作水平的提升。数值结果表明，即使诱惑很大，时间尺度机制依然能够促进合作的生存。这是因为当博

弈交互时间尺度大于策略更新时间尺度时，对于高收益的策略能够维持更长的时间，若中心玩家为合作者时，合作者就能维持更长的时间，背叛者的数量就会减少；而若中心玩家是背叛者时，会从它的周围合作邻居那获得高额回报，故而与背叛者相邻的合作者倾向于在随后几轮采用背叛策略，反过来又减少了背叛者的收益，Q 学习是一种创新规则，可以引入目前人群中不存在的新策略，所以在这种四周都是背叛者状态时合作者仍然可以出现，从而导致背叛策略的短期存在。不仅如此，我们研究了影响 Q 学习的参数，一般来说，高学习率和适度的折扣系数更有利于合作。最后，还验证了无论初始状态如何，合作将始终收敛到稳定水平。

我们的工作不同于参考文献[27]在混合良好的人群中进行，我们关注的是结构化人群，这更接近经验观察。也不同于参考文献[19] [20] [21] [28]将费米函数用作策略更新规则，揭示了自关注 Q 学习算法对于解决社会困境的特定积极作用。因此，我们相信，我们的工作能够为未来提高合作率的研究带来更多启示，从而推动社会困境解决机制的发展。

参考文献

[1] Pennisi, E. (2009) On the Origin of Cooperation. *Science*, **325**, 1196-1199. https://doi.org/10.1126/science.325_1196

[2] Axelrod, R. and Hamilton, W.D. (1981) The Evolution of Cooperation. *Science*, **211**, 1390-1396. https://doi.org/10.1126/science.7466396

[3] Zhu, Y., Zhang, Z., Xia, C. and Chen, Z. (2023) Equilibrium Analysis and Incentive-Based Control of the Anticoordinating Networked Game Dynamics. *Automatica*, **147**, Article ID: 110707. https://doi.org/10.1016/j.automatica.2022.110707

[4] Jian, Q., Li, X., Wang, J. and Xia, C. (2021) Impact of Reputation Assortment on Tag-Mediated Altruistic Behaviors in the Spatial Lattice. *Applied Mathematics and Computation*, **396**, Article ID: 125928. https://doi.org/10.1016/j.amc.2020.125928

[5] Li, W.J., Chen, Z., Jin, K.Z., Wang, J., Yuan, L., Gu, C., Perc, M., *et al*. (2022) Options for Mobility and Network Reciprocity to Jointly Yield Robust Cooperation in Social Dilemmas. *Applied Mathematics and Computation*, **435**, Article ID: 127456. https://doi.org/10.1016/j.amc.2022.127456

[6] Zhang, J., Zhang, C., Chu, T. and Perc, M. (2011) Resolution of the Stochastic Strategy Spatial Prisoner's Dilemma by Means of Particle Swarm Optimization. *PLOS ONE*, **6**, e21787. https://doi.org/10.1371/journal.pone.0021787

[7] Nowak, M.A. and May, R.M. (1992) Evolutionary Games and Spatial Chaos. *Nature*, **359**, 826-829. https://doi.org/10.1038/359826a0

[8] Wang, Z., Jusup, M., Wang, R.W., Shi, L., Iwasa, Y., Moreno, Y. and Kurths, J. (2017) Onymity Promotes Cooperation in Social Dilemma Experiments. *Science Advances*, **3**, e1601444. https://doi.org/10.1126/sciadv.1601444

[9] Amaral, M.A., Wardil, L., Perc, M. and da Silva, J.K. (2016) Stochastic Win-Stay-Lose-Shift Strategy with Dynamic Aspirations in Evolutionary Social Dilemmas. *Physical Review E*, **94**, Article ID: 032317. https://doi.org/10.1103/PhysRevE.94.032317

[10] Nowak, M.A. and Sigmund, K. (1992) Tit for Tat in Heterogeneous Populations. *Nature*, **355**, 250-253. https://doi.org/10.1038/355250a0

[11] Xia, C., Gracia-Lázaro, C. and Moreno, Y. (2020) Effect of Memory, Intolerance, and Second-Order Reputation on Cooperation. *Chaos*: *An Interdisciplinary Journal of Nonlinear Science*, **30**, Article ID: 063122. https://doi.org/10.1063/5.0009758

[12] Molenmaker, W.E., de Kwaadsteniet, E.W. and van Dijk, E. (2016) The Impact of Personal Responsibility on the (un) Willingness to Punish Non-Cooperation and Reward Cooperation. *Organizational Behavior and Human Decision Processes*, **134**, 1-15. https://doi.org/10.1016/j.obhdp.2016.02.004

[13] Wang, S., Chen, X. and Szolnoki, A. (2019) Exploring Optimal Institutional Incentives for Public Cooperation. *Communications in Nonlinear Science and Numerical Simulation*, **79**, Article ID: 104914. https://doi.org/10.1016/j.cnsns.2019.104914

[14] Li, X., Jusup, M., Wang, Z., Li, H., Shi, L., Podobnik, B., Boccaletti, S., *et al*. (2018) Punishment Diminishes the Benefits of Network Reciprocity in Social Dilemma Experiments. *Proceedings of the National Academy of Sciences of the United States of America*, **115**, 30-35. https://doi.org/10.1073/pnas.1707505115

[15] Luo, C., Zhang, X. and Zheng, Y. (2017) Chaotic Evolution of Prisoner's Dilemma Game with Volunteering on Inter-

dependent Networks. *Communications in Nonlinear Science and Numerical Simulation*, **47**, 407-415. https://doi.org/10.1016/j.cnsns.2016.12.004

[16] Guo, H., Song, Z., Geček, S., Li, X., Jusup, M., Perc, M., Wang, Z., *et al*. (2020) A Novel Route to Cyclic Dominance in Voluntary Social Dilemmas. *Journal of the Royal Society Interface*, **17**, Article ID: 20190789. https://doi.org/10.1098/rsif.2019.0789

[17] Gross, J. and De Dreu, C.K. (2019) The Rise and Fall of Cooperation through Reputation and Group Polarization. *Nature Communications*, **10**, Article No. 776. https://doi.org/10.1038/s41467-019-08727-8

[18] Pal, A. and Sengupta, S. (2022) Network Rewiring Promotes Cooperation in an Aspirational Learning Model. *Chaos*: *An Interdisciplinary Journal of Nonlinear Science*, **32**, Article ID: 023109. https://doi.org/10.1063/5.0071873

[19] Mao, Y., Rong, Z. and Wu, Z.X. (2021) Effect of Collective Influence on the Evolution of Cooperation in Evolutionary Prisoner's Dilemma Games. *Applied Mathematics and Computation*, **392**, Article ID: 125679. https://doi.org/10.1016/j.amc.2020.125679

[20] Han, W., Zhang, Z., Sun, J. and Xia, C. (2021) Emergence of Cooperation with Reputation-Updating Timescale in Spatial Public Goods Game. *Physics Letters A*, **393**, Article ID: 127173. https://doi.org/10.1016/j.physleta.2021.127173

[21] Mao, Y., Zhao, Q., Song, R., *et al*. (2021) Timescales Diversity Induces Influencers to Persist Cooperation on Scale-free Networks. *IEEE International Symposium on Circuits and Systems*, Daegu, 22-28 May 2021, 1-5. https://doi.org/10.1109/ISCAS51556.2021.9401147

[22] Perc, M., Jordan, J.J., Rand, D.G., Wang, Z., Boccaletti, S. and Szolnoki, A. (2017) Statistical Physics of Human Cooperation. *Physics Reports*, **687**, 1-51. https://doi.org/10.1016/j.physrep.2017.05.004

[23] Wang, Z., Kokubo, S., Jusup, M. and Tanimoto, J. (2015) Universal Scaling for the Dilemma Strength in Evolutionary Games. *Physics of Life Reviews*, **14**, 1-30. https://doi.org/10.1016/j.plrev.2015.04.033

[24] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Hassabis, D., *et al*. (2015) Human-Level Control through Deep Reinforcement Learning. *Nature*, **518**, 529-533. https://doi.org/10.1038/nature14236

[25] Hu, S., Leung, C.W. and Leung, H.F. (2019) Modelling the Dynamics of Multiagent q-Learning in Repeated Symmetric Games: A Mean Field Theoretic Approach. *Advances in Neural Information Processing Systems*, **32**, 12125-12135.

[26] He, Z., Geng, Y., Du, C., Shi, L. and Wang, Z. (2022) *Q*-Learning-Based Migration Leading to Spontaneous Emergence of Segregation. *New Journal of Physics*, **24**, Article ID: 123038. https://doi.org/10.1088/1367-2630/acadfd

[27] Zhang, J., Weissing, F.J. and Cao, M. (2016) Fixation of Competing Strategies When Interacting Agents Differ in the Time Scale of Strategy Updating. *Physical Review E*, **94**, Article ID: 032407. https://doi.org/10.1103/PhysRevE.94.032407

[28] Xu, X., Rong, Z., Tian, Z. and Wu, Z.X. (2019) Timescale Diversity Facilitates the Emergence of Cooperation-Extortion Alliances in Networked Systems. *Neurocomputing*, **350**, 195-201. https://doi.org/10.1016/j.neucom.2019.03.057