Journal of Systems Engineering and Electronics
Vol. 32, No. 3, June 2021, pp.642 – 657

A single-task and multi-decision evolutionary game model based on multi-agent reinforcement learning

MA Ye[1,*], CHANG Tianqing[1], and FAN Wenhui[2]

1. Academy of Army Armored Force, Beijing 100072, China;
2. Department of Automation, Tsinghua University, Beijing 100084, China

Abstract: In the evolutionary game of the same task for groups, the changes in game rules, personal interests, the crowd size, and external supervision cause uncertain effects on individual decision-making and game results. In the Markov decision framework, a single-task multi-decision evolutionary game model based on multi-agent reinforcement learning is proposed to explore the evolutionary rules in the process of a game. The model can improve the result of a evolutionary game and facilitate the completion of the task. First, based on the multi-agent theory, to solve the existing problems in the original model, a negative feedback tax penalty mechanism is proposed to guide the strategy selection of individuals in the group. In addition, in order to evaluate the evolutionary game results of the group in the model, a calculation method of the group intelligence level is defined. Secondly, the Q-learning algorithm is used to improve the guiding effect of the negative feedback tax penalty mechanism. In the model, the selection strategy of the Q-learning algorithm is improved and a bounded rationality evolutionary game strategy is proposed based on the rule of evolutionary games and the consideration of the bounded rationality of individuals. Finally, simulation results show that the proposed model can effectively guide individuals to choose cooperation strategies which are beneficial to task completion and stability under different negative feedback factor values and different group sizes, so as to improve the group intelligence level.

Keywords: multi-agent, reinforcement learning, evolutionary game, Q-learning.

DOI: 10.23919/JSEE.2021.000055

1. Introduction

Reinforcement learning is a model of machine learning. It can actively sense the environment via different behaviors or actions, evaluate the actions and adjust subsequent actions. It is a learning technology mapping different environmental states into actions [1]. Reinforcement learning mainly aims to choose the optimal action of the agents when they complete goals. It is widely used in robot control systems [2,3], intelligent decision-making [4,5], nonlinear optimal control [6−8] and other fields. A single agent generally has no proper decision-making ability or the ability to sense the environment, so it cannot respond to complex actual problems. Therefore, the concept of multi-agent was proposed at the end of the 20th century. It is a collection of multiple agents and also called multi-agent system. It belongs to the forefront of distributed artificial intelligence and is mainly used to explore the coordination, cooperation, communication, and conflicts among agent groups [9].

Reinforcement learning is also applied in multi-agent learning. In recent years, the combination of multi-agent learning and reinforcement learning has become a research hotspot [10−12]. One of the prominent achievements is AlphaGO, the Chinese game of go system based on reinforcement learning, which has defeated the top human players in the game and shows a great advantage. It immediately attracts the attention of all walks of life and more researchers have participated in the field of multi-agent reinforcement learning [13].

Game is a theory of action, used to study the strategic choices driven by multiple interests among multiple individuals [14]. Due to the development of the game theory, many research methods have been combined with the game theory, such as genetic algorithms [15], particle swarm optimization [16], and multi-agent systems. The idea of cooperation among agents in a multi-agent system can well reflect the game process of individuals in social groups in their work. The multi-agent system has been gradually introduced into the game field and achieved better results [17,18]. The game involves not only the action of an agent, but also the states of other agents, which increase the complexity of the system and learning, so the

Manuscript received March 03, 2020.
*Corresponding author.
This work was supported by the National Key R&D Program of China (2017YFB1400105).

convergence rate cannot be guaranteed. The introduction of reinforcement learning can better guide the process of the game. Bendor et al. [19] used reinforcement learning to study the steady-state convergence problem of games. Jacob et al. [20] improved the reinforcement learning algorithm and solved the poor player strategy selection problem in two-player tasks. As one of the reinforcement learning algorithms, the Q-learning algorithm does not require the dynamic environment and the characteristics suitable for long-scenario tasks in advance, so it is widely used in the game field. For example, Littman et al. [4] proposed the minimax algorithm. This algorithm can well solve the two-player zero-sum game, but it cannot be applied in games with more than two players. Jun et al. [21] proposed a random game Q-learning algorithm to search for the optimal strategy. In addition, according to the different effects of the reward function in the game task, the algorithm can be divided into three different types: fully cooperative, fully competitive, and mixed types [22]. The reward functions for different agents in the fully cooperative algorithm are the same and can be used in multi-intelligence systems with the same goal. The classic algorithms are distributed Q-learning algorithms [23] and team Q-learning algorithms [24]. The agents in the fully competitive algorithm are in a state of competition with each other in order to maximize their own returns while minimizing others' returns. Its classic algorithms include the Minimax-Q learning algorithm [25] and the Nash-Q learning algorithm [26]. The return function in the hybrid algorithm is not related to each other and there is no deterministic rule. It is suitable for the study on the equilibrium solution in the game theory. Its classic algorithms include the fuzzy Q-learning algorithm [27] and the correlated Q-learning algorithm [28].

In the study, through the full consideration of the advantages of multi-agents and reinforcement learning in the field of gaming, a single-task multi-decision evolutionary game model is proposed based on multi-agent reinforcement learning to explore the evolution of groups. This model combines the evolutionary game theory to perform a three-player evolutionary game. A negative feedback tax penalty mechanism is proposed and an improved Q-learning algorithm is used to optimize the effect of this mechanism. The algorithm in this paper can take into account the individual's incomplete rationality. In addition, a calculation method of group intelligence level is defined to evaluate the results of the group evolutionary game. The simulation results show that the introduction of the reinforcement learning algorithm can improve the guiding role of the negative feedback tax penalty mechanism and promote the evolution of decision-making group towards the direction of cooperation.

2. Key technical principles of the model

This paper proposes an evolutionary game model based on multi-agent reinforcement learning to explore the evolution rules in the game process. The key technical principles involved in the model are described below.

2.1 Evolutionary game theory

The evolutionary game theory is an extension of the classical game theory. The significant difference between them is that individuals in the evolutionary game theory can be non-rational [29,30]. The main research object of evolutionary games is the group game, namely, the game of multiple individuals. In the game, the interactive characteristics of the group strategy are described below. Firstly, there is no identity difference among individuals and the only difference among individuals is the selected strategy. Secondly, the group has only one strategy set and the number of strategies is limited. Individuals choose strategies from the strategy set. Individuals adopting the same strategy have the same rewards and their rewards depend entirely on the currently selected strategy. Thirdly, the rewards of each strategy are related to the number or proportion of the choice of the corresponding strategy.

The evolutionary game model consists of two main parts: group game and group state update, as shown in Fig. 1. The group game consists of three parts: the number of individuals, the set of strategies, and the reward utility function. Let the total number of individuals in the group be N, and the strategy set be $s=\{1,2,\cdots,m\}$ (m is the total number of strategies); x_i is the total number of individuals who choose strategy $i \in S$. The group state is $x=(x_1,x_2,\cdots,x_m)$, where $x_i \in \mathbf{N}$ and $\sum_{i\in S} x_i = N$. The group's state set is $X=\left\{x|x_i \in \mathbf{N}, \sum_{i\in S} x_i = N\right\}$. The reward of each individual is represented by the reward utility function $U_i \in X \to \mathbf{R}$, which corresponds to its chosen strategy. It represents the mapping from the state to the set of real numbers.

The change of the group state caused by the change of group time is the core of the evolutionary game. The group's decision-making action in the game process can be analyzed according to the change of the group state [31].

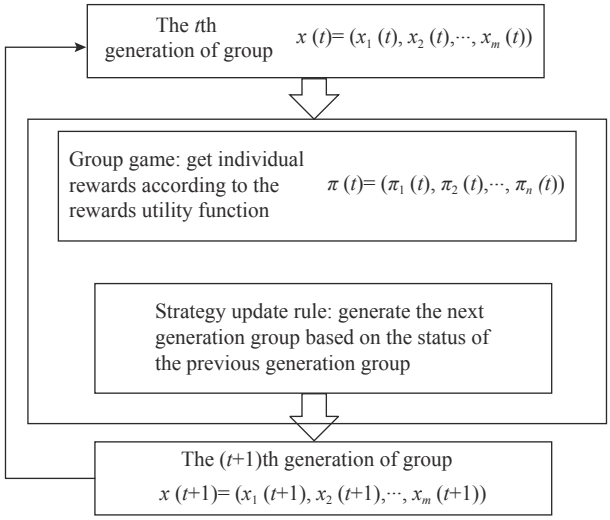


Fig. 1 Components of the evolutionary game model

2.2 Framework of game learning

In the evolutionary game, the group state is updated by adjusting its own strategy according to the game learning rules, which generally include individual game rules and information comparison rules with other individual strategies and rewards. The above process is also called game learning [32]. In each time step t, each individual $v_i \in v$ will continuously update its own strategy $s_i(t) \in S_i$ during the game cycle, where $s(t) = (s_1(t), s_2(t), \cdots, s_m(t)) \in S$ and $s(t)$ is the current strategies combination of all individuals. Each individual gets rewards $\pi_i(t) = U_i(s(t))$. Therefore, the discrete-time evolutionary game is defined as a triple $\Gamma = (v, \{S_i | v_i \in v\}, \{U_i | v_i \in v\})$. The framework of the game learning is shown in Fig. 2.

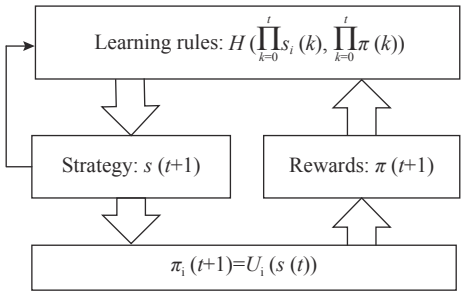


Fig. 2 Framework of game learning

The game learning rules are generally expressed as

$$s_i(t+1) = H_i\left(\prod_{k=0}^{t} s_i(k); \prod_{k=0}^{t} s_{-i}(k); U_i\right). \tag{1}$$

According to (1), all information utilized by each individual $v_i \in v$ has its own historical strategies $\prod_{k=0}^{t} s_i(k)$, other individuals' historical strategies $\prod_{k=0}^{t} s_{-i}(k)$, its own reward function U_i and its own historical rewards $\prod_{k=0}^{t} \pi_i(k)$. Among them, the individual learning rule H_i can be divided into a deterministic function or a random function according to the situation. Each individual can determine the next strategy according to rules $s_i(t+1) \in S_i$. The above learning rules are based on the assumption that all individuals are completely rational and can obtain all game information. In reality, individuals may not be consistent with the above assumptions. Therefore, the learning rules for bounded rationality and limited information acquisition ability are expressed as

$$s_i(t+1) = H_i(s_i(t); s_{-i}(t); U_i). \tag{2}$$

In other words, the memory ability of each individual is changed from infinite memory ability to limited memory ability, which is closer to the reality.

2.3 Learning framework of evolutionary game based on multi-agent

Multi-agent is introduced into the evolutionary game and each individual in the group is treated as an agent. Each agent can interact with each other and freely choose a strategy. Their own models, methods, and knowledge bases form the basic agent structure. When an agent is learning a game, it is defined as the main decision agent. Through the agent structure, each agent can extract the information required for its strategy selection from the environment and other individuals during game learning and simultaneously store the information into the knowledge base to establish its model library. Finally, based on the information in the method library and game learning information, a comprehensive analysis is performed to complete the strategy selection. The remaining agents are ordinary agents responsible for providing information to the main decision agent. The framework of the evolutionary game based on multi-agent is shown in Fig. 3.

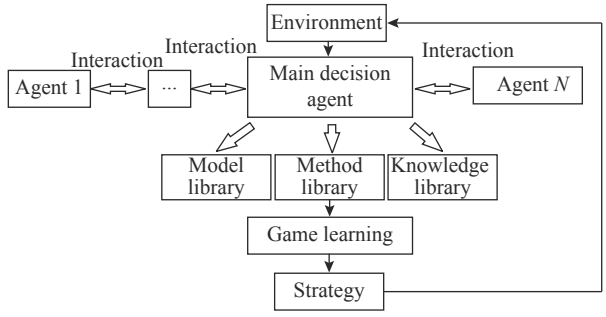


Fig. 3 Framework of evolutionary game based on multi-agent

2.4 Evolutionary game flow based on multi-agent and reinforcement learning

Reinforcement learning is a special branch of machine learning and shares the features of supervised learning and unsupervised learning. Its core is the information interaction between agents and the environment [33,34]. The mathematical framework of reinforcement learning is based on the Markov decision process (MDP). The Markov process consists of five key parts:

(i) Agent state is $X = \left\{ x | x_i \in \mathbf{N}, \sum_{i \in S} x_i = N \right\}$.

(ii) The strategy adopted by the agent for state transfer is $s(t) = (s_1(t), s_2(t), \cdots, s_m(t)) \in S$.

(iii) The agent's transition probability from the state x to the state x' according to the strategy s is $p^s_{xx'}$.

(iv) The probability that the agent who transits from the state x to the state x' according to the strategy s can obtain the reward is $R^s_{xx'}$.

(v) The discount factor controlling the reward is γ.

In the process of reinforcement learning, the agent will get corresponding rewards or rewards after the game learning is completed. In general, if the strategy is good, the reward is positive, otherwise it is negative. The agent always hopes to get the maximum reward. The calculation formula of the reward is as follows:

$$R_t = r_{t+1} + r_{t+2} + \cdots + r_T \tag{3}$$

where r_T is the reward for the agent's transition from one state to another state within time step T. If the task performed is a continuous task without a final state, a discount factor γ needs to be introduced to maximize the reward. The discount factor ranges from 0 to 1. Then, the calculation formula of the rewards can be expressed as

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}. \tag{4}$$

The purpose of reinforcement learning is to find the optimal strategy that enables each state of the agent to achieve the correct action. A value function is needed to represent the optimal degree of the agent in a specific state under the strategy π. The hypothetical value function is denoted as $V(s)$, which is the state value under a certain strategy. The function is expressed as

$$V^{\pi}(x) = \mathrm{E}_{\pi}[R_t | x_t = x]. \tag{5}$$

Equation (5) represents the expectation of reward under the strategy π and state x. Substituting (4) into (5) gives

$$V^{\pi}(x) = \mathrm{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | x_t = x \right]. \tag{6}$$

In order to represent the optimal degree of a particular action selected by an agent in a particular state under the strategy π, its state-action value function is defined as the Q function as follows:

$$Q^{\pi}(x, a) = \mathrm{E}_{\pi}[R_t | x_t = x, a_t = a]. \tag{7}$$

Equation (7) represents the expectation of reward for the action a under the strategy π and state x. Substituting (4) into (7) gives

$$Q^{\pi}(x, a) = \mathrm{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | x_t = x, a_t = a \right]. \tag{8}$$

In the evolutionary game process based on multi-agent reinforcement learning, when the agent chooses a strategy, if the environment gives the positive feedback (the reward value is good), the probability that the agent chooses the same strategy will increase in the next round, otherwise it will decrease. Therefore, the decision-making agent will acquire knowledge, learn from the acquired knowledge and the feedback given by the environment, and select a strategy. The flow of the evolutionary game based on multi-agent reinforcement learning is shown in Fig. 4.

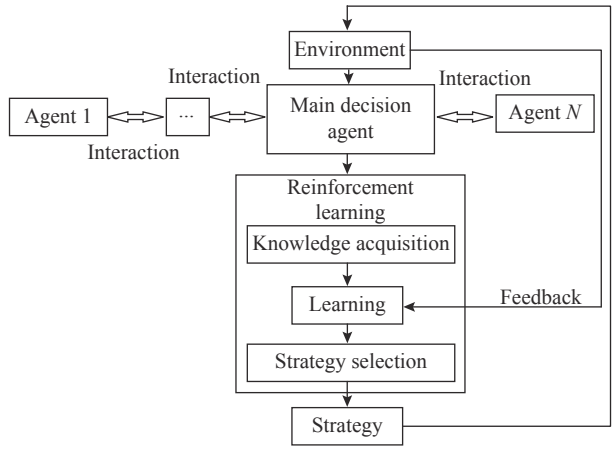


Fig. 4 Evolutionary game flow based on multi-agent reinforcement learning

3. Original model

In order to better describe the cooperative relationship and evolution rules in a group, a calculation method of the group intelligence level is defined. Firstly, it is assumed that each agent in the group has an intelligence level I, $I \in [0, 1]$. The intelligence level of the group is CI, which is the sum of the intelligence levels of individual agents in the group, $CI = \sum I$. When the individual agents in the group are playing an evolutionary game, the intelligence level of the group will be changed. Then, the

intelligence level of the group is $CI=\sum I+\Delta I$, where ΔI is the intelligent variation generated when individual agents participating in evolutionary games choose different strategies. According to this evolutionary game model, it is assumed that any three individual agents, namely, Agent i, Agent j and Agent k, respectively, have intelligence levels of I_i, I_j, and I_k. When they choose different strategies, intelligent variations are generated. The calculation formulas of intelligent variation for cooperation, competition and inaction strategies are provided.

The formula for calculating the intelligent variation of the cooperation strategy is

$$\Delta I=I_i+I_j+I_k+I_i\cdot I_j\cdot I_k-I_i-I_j-I_k=I_i\cdot I_j\cdot I_k. \tag{9}$$

The formula for calculating the intelligent variation of the competition strategy is

$$\Delta I=\max(I_i,I_j,I_k)-I_i-I_j-I_k. \tag{10}$$

The formula for calculating the intelligent variation of the inaction strategy is

$$\Delta I=0-I_i-I_j-I_k=-(I_i+I_j+I_k). \tag{11}$$

The intelligence level per capita is defined as the ratio of the group intelligence level to the total number of people N. According to (9)−(11), in the evolutionary game, the group intelligence level CI may be any value higher or lower than the sum of individual agent intelligence levels $\sum I, CI\in(-\infty,+\infty)$. According to the model and the definition of the intelligence level, when all the individual agents participating in the task adopt a cooperation strategy, the intelligence level of the group is the highest. The group has the lowest level of intelligence when the inaction strategy is adopted. When a competition strategy is adopted, the group has the medium intelligence level. The significance of defining the level of group intelligence is to provide an evaluation method for different results produced by different strategies adopted by the group in the process of game evolution. The intelligence level of the group is used as an indicator to measure the overall performance of the decision-making group when completing a task. Based on changes of the group intelligence level, the evolutionary rule of the game and the efficiency of collaboration between groups can be quantitatively analyzed. The changes of the group intelligence level imply that a decision-making individual not only applies the strategy in the three-person game, but also brings it into work and interaction with other individuals. In other words, an individual strategy is consistent with the behavior of an individual.

In order to explore the evolution rule of individual decision in a group, a simple original evolutionary game model is constructed. The original model is designed for an extreme situation and can more intuitively reflect the influences of key parameters in the model on the selection of the individual strategy and the evolution of the group intelligence level. The model assumes that a group completes a task together. The total number of individuals in the group is N and each individual is regarded as an agent. The cost of completing the task is C and the reward is R, where $R>C$. Each individual agent in the group can freely choose its own strategy during the evolutionary game and obtain the right to participate in the task by playing the game through the selected strategy. The set of strategies can be divided into three types: cooperation strategies (those who adopt the strategies are called cooperator, referred to as Co), competition strategies (those who adopt the strategies are called defender, referred to as D), and inaction strategies (those who adopt the strategies are called loner, referred to as L). After the individual agents determine their own strategies, the group will sequentially perform a non-repeating three-person random game and the winners in the game will jointly complete the task. The judgment results of game rules are shown in Table 1. The combination order is not considered in strategy combinations. For example, the combinations *Co Co D*, *Co D Co*, and *D Co Co* use the same game rules. The game rule is extended from the two-person game decision rule (when the strategy combination is *Co Co*, both win; when the strategy combination is *Co D*, *D* wins; when *Co L*, *Co* wins, when the strategy combination is *D D*, a random one wins; when the strategy combination is *D L*, *D* wins; when the strategy combination is *L L*, no one wins) and stipulates that two of the three are randomly selected to play the two-person game. The winner and the remaining person continue to play the game to select the final three-person game winner. The individual agent that chooses a competitive strategy can only win once, and the two of the three who choose the same strategy have the priority to play the game. For example, when a three-person game is played, two persons choose a cooperative strategy and one person chooses a competitive strategy. The two persons who choose a cooperative strategy will play the game first. Both of them will win, and then play a game with the remaining agent who chooses a competitive strategy. The agent will arbitrarily choose because we only care the number of the final winners. Whoever wins does not affect the number of the final winners. In the end, the winners of the three-person game are one of the agents that chooses the cooperative

strategy and the one that chooses the competitive strategy. In summary, this rule ensures that the final result of the strategy combination is not affected by the sequence of the combination. For example, the winners of the combinations of *Co Co D*, *Co D Co*, and *D Co Co* are all *Co D* (any one of *Co*).

Table 1 Game rules decision tables

Combination of strategies	Winner	Combination of strategies	Winner
Co Co Co	*Co Co Co*	*Co D L*	*D*
Co Co D	*Co D* (any one of *Co*)	*D D D*	*D* (any one of *D*)
Co Co L	*Co Co*	*D L L*	*D*
Co D D	*D* (any one of *D*)	*D D L*	*D* (any one of *D*)
Co L L	*Co*	*L L L*	none

In the model, it is assumed that the cost of completing the task is borne by all the agents participating in the task, whereas the rewards of the task are shared equally by all individual agents in the group. In other words, a "free-rider" action that an individual agent who does not participate in the task shares the reward is allowed. If the number of individual agent participating in the task is a, at time t, the rewards that can be obtained by the individual agent participating in the task are expressed as

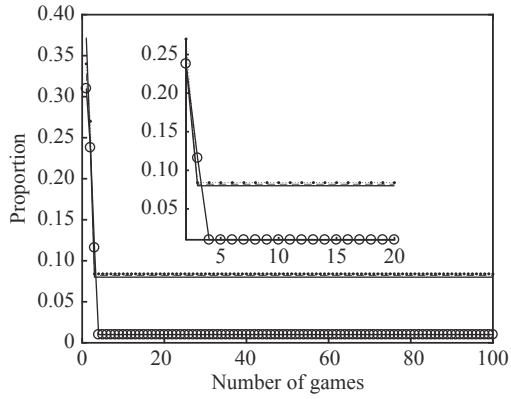
$$\pi(t) = \frac{R}{N} - \frac{C}{a}. \tag{12}$$

The rewards obtained by an individual agent who does not participate in the task can be expressed as

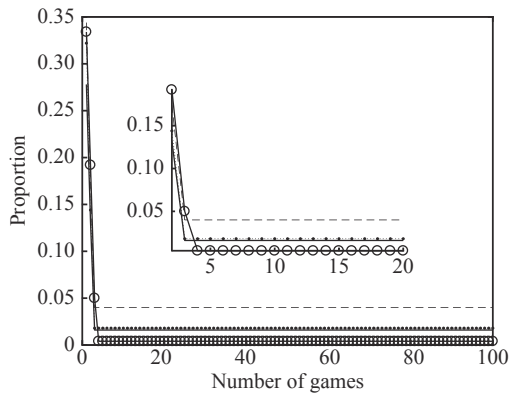
$$\pi(t) = \frac{R}{N}. \tag{13}$$

The evolutionary game learning process is described as follows. After one round of the game is completed, each agent will randomly select a certain number of subgroups for strategic comparison learning. The individual learning agent is also the main decision-making agent. If its own reward is less than the minimum value of the subgroup, the individual agent will copy the strategy of the individual agent with the largest reward in the subgroup. When all individual agents have completed the game learning, they will start the next round of the game and continue to advance the evolutionary game process until the game ends.

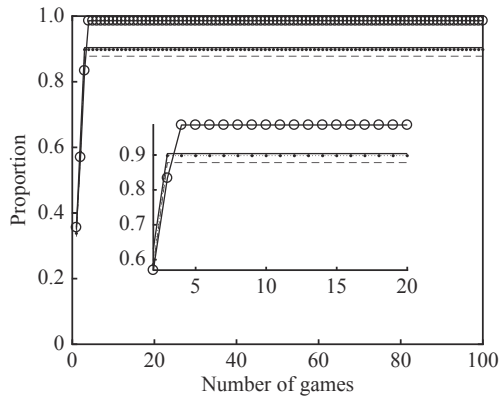
According to the model settings, the key parameters that affect the reward and game learning process include task cost C and task reward R. In order to facilitate the description of the relationship between rewards and cost, a simulation experiment is performed on the evolutionary game model with the reward-cost ratio R/C. In real life, the individual's ability to interact with other individuals is limited, so the number of subgroups is set to 4. The total number of individual agents N in the group is set to be 500 and the evolutionary game involves 100 rounds. The experiment is repeated 300 times. In the evolutionary game process, the proportions of individual agents choosing different strategies in the group and the group intelligence level under different reward-cost ratios are shown in Fig. 5 (for a clearer display of the changes in different reward-cost ratios, enlarged parts are added).



(a) Proportion of cooperation strategies



(b) Proportion of competition strategies



(c) Proportion of inaction strategies

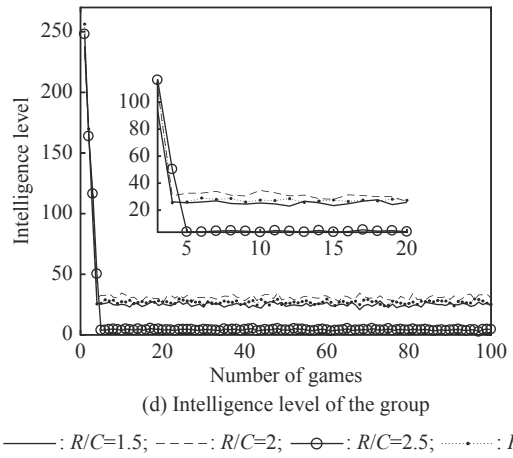


Fig. 5 Proportions of different strategies and the intelligence level of the group in the evolutionary game

Fig. 5(a)−Fig. 5(c) show the changing trend of the proportions of individuals choosing different strategies in groups with different reward-cost ratios. In the evolutionary game, under different reward-cost ratios, the proportion of choosing cooperation and competition strategies shows an obvious downward trend, whereas the proportion of choosing the inaction strategy shows an obvious upward trend. For the cooperation strategy, when the reward-cost ratios are at a small or intermediate value (R/C=1.5, R/C=2 or R/C=2.5), the proportion of the cooperation strategy in the population is suppressed. When the reward-cost ratio R/C is large (R/C=3), the proportion of the cooperation strategy in the population has an advantage. The smaller the selected proportion of the cooperation strategy in the group is, the larger the selected proportions of the inaction strategy and the competitive strategy in the group are. When the evolution is stable, the intelligence level of the group with different reward-cost ratios shows a significant decline (see Fig. 5(d)). When the reward-cost ratio is 3, the intelligence level of the group is the highest and the group collaboration effect is also the best.

In summary, due to the acquiescence of the "free-rider" action, the task income is unfairly distributed and the individual agents in the group are always prone to choose the inaction strategy regardless of the change in the reward-cost ratio, so as to ensure their own rewards. However, the task participation rate remains to be at a low level. Due to the sharp decline in the proportion of choosing cooperation strategies, the group intelligence level will be greatly reduced, thus negatively affecting the completion of tasks. In the setting of the original model, when all individuals choose a cooperation strategy, the group has the highest intelligence level and the best task completion effect, and the "free-rider" action does not promote the group to evolve towards the cooperation direction. Therefore, the model needs to be improved in such a way that it can guide the evolutionary game direction of the group to the ideal situation and reduce the proportion of choosing the inaction strategy.

4. Improved model based on the negative feedback tax penalty mechanism

The simulation results of the original model indicate that it is necessary to restrict the "free-rider" action in the group. Therefore, we optimize the reward rules of individual agents, increase taxes to appropriately reduce the rewards of individuals who do not participate in the task, and reward the taxes to the individual agents who participate in the task. In this way, the group is guided to evolve towards the cooperation direction and the group intelligence level is improved. In reality, the formulator of the reward rules may be the leaders of enterprises, institutions and government departments. According to (12) and (13), individuals who have not participated in the task can obtain rewards without any cost. In order to punish the individual agents who do not participate in the task, the model increases the tax rate T ($0 \leqslant T \leqslant 1$) to tax the individuals who do not participate in the task and transfer the tax equally to the individuals who participate in the task. In this way, the secondary distribution of the rewards is realized. The smaller the value of T is, the lighter the punishment effect on the "free-riding" action is. Therefore, the rewards of individuals who are not involved in the task are higher than the rewards of individuals who participate in the task and the number of agents participating in the task decreases. This process is a negative feedback process. The purpose of introducing a negative feedback tax rate T is to reduce the "free-riding" action through the evolutionary game learning process from others and increase the task participation rate. Then, the group evolves towards the cooperative direction. If the punishment is too large, it brings out higher costs and is not conducive to the evolution stability of the group. Therefore, the value of the tax rate T depends on the result of the evolutionary game. If the number of individuals participating in the task is a, at time t, the rewards of the individuals participating in the task are expressed as

$$\pi(t) = \frac{R}{N} - \frac{C}{a} + \frac{RT(N-a)}{aN}. \tag{14}$$

The rewards of individuals who have not participated in the task are expressed as

$$\pi(t) = \frac{R(1-T)}{N}. \tag{15}$$

The model proposes a tax rate penalty mechanism with negative feedback characteristics as follows:

$$T = 1 - \frac{a^{\beta}}{N} \tag{16}$$

where N is the total number of people in the group; a is the number of people participating in the task and related to the result of each round of evolutionary games; β is a negative feedback factor, a preset parameter of external forces (its value is about 1 based on the consideration of the actual tax rate). Fig. 6 shows the variation of the tax rate T with the number of participants under N = 30 and β=1, 0.5 and 1.1.

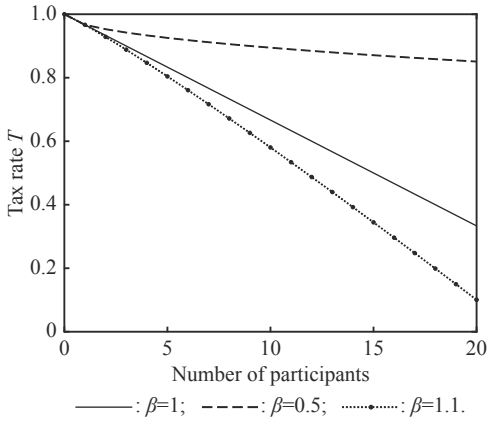
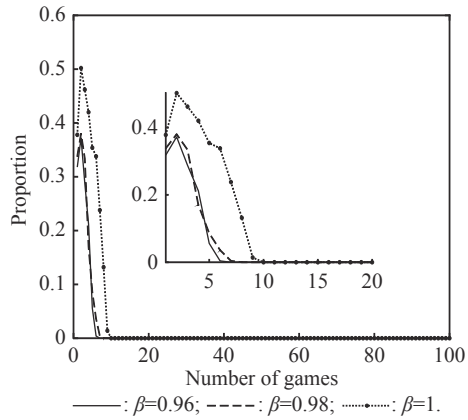


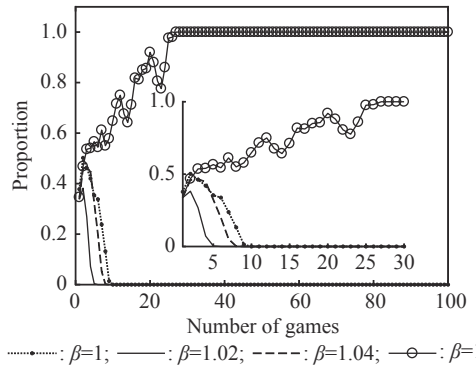
Fig. 6 Variation of the tax rate T with the number of participants

As shown in Fig. 6, when $\beta = 1$, the tax rate T is the standard negative feedback and its value decreases as the participation rate increases. When $\beta < 1$, the penalty effect of the tax rate T is gradually weakened. With the decrease in the number of people participating in the task decreases, the value of T decreases and the decreasing rate of T also decreases. When $\beta > 1$, the penalty effect of the tax rate T is gradually increased. The value of T decreases significantly with the increase in the number of participants, and the decreasing rate of T gradually increases. When t is 0, the model under the negative feedback tax penalty mechanism becomes the original model and the original model can be regarded as a special case.

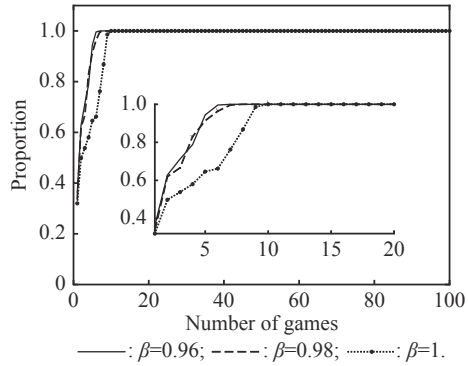
The simulation is performed with the improved model under the negative feedback tax penalty mechanism. The value of β is about 1, so the six values of β are respectively set as: 0.96, 0.98, 1, 1.02, 1.04, and 1.06. According to the analysis results of the original model, the reward-cost ratio of 3 is more conducive to the guidance of the cooperation strategy. Therefore, the reward-cost ratio is set as 3. The number of individual agents N in the group is 500. The evolutionary game involves 100 rounds. The evolution of the proportions of individual agents choosing different strategies in the groups under different β values and the group intelligence level are shown in Fig. 7.



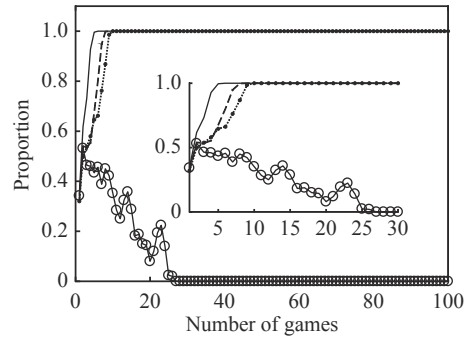
(a) Proportion of individual agents choosing cooperation strategies under β values of 0.96, 0.98, 1



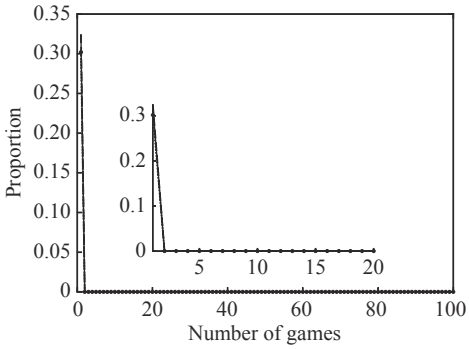
(b) Proportion of individual agents choosing cooperation strategies under β values of 1, 1.02, 1.04, 1.06



(c) Proportion of individual agents choosing competition strategies under β values of 0.96, 0.98, 1

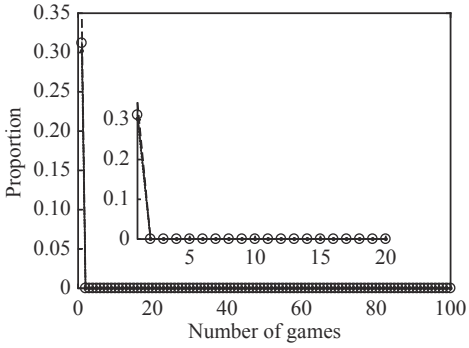


(d) Proportion of individual agents choosing competition strategies under β values of 1, 1.02, 1.04, 1.06



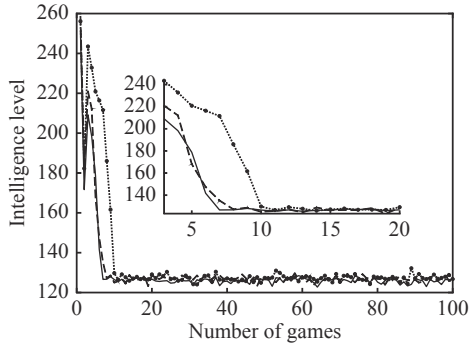
——: β=0.96; - - - -: β=0.98; ·····: β=1.

(e) Proportion of individual agents choosing inaction strategies under β values of 0.96, 0.98, 1



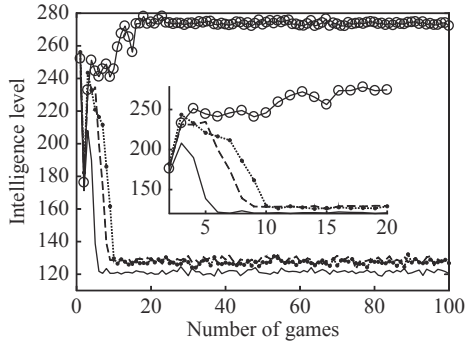
·····: β=1; ——: β=1.02; - - - -: β=1.04; —o—: β=1.06.

(f) Proportion of individual agents choosing inaction strategies under β values of 1, 1.02, 1.04, 1.06



——: β=0.96; - - - -: β=0.98; ·····: β=1.

(g) Intelligence level of the group under β values of 0.96, 0.98, 1



·····: β=1; ——: β=1.02; - - - -: β=1.04; —o—: β=1.06.

(h) Intelligence level of the group under β values of 1, 1.02, 1.04, 1.06

Fig. 7 Proportions of different strategies and the intelligence level of the group in the evolutionary game

Fig. 7(a)−Fig. 7(f) show the changing trend of the proportions of individual agents choosing various strategies in groups under different reward-cost ratios. When the negative feedback factor $\beta = 1.06$, the proportion of individual agents choosing a cooperation strategy is increasing significantly and eventually stabilized at a higher level. When the negative feedback factor β is 0.96, 0.98, 1, 1.02, and 1.04, the proportion of individual agents choosing a cooperation strategy firstly shows a temporary upward trend, then significantly declines and is finally remained at a lower level. When $\beta = 1.06$, the best guidance effect on the cooperation strategy is realized and the proportion of individual agents choosing a cooperation strategy is greatly increased. When other β values are set, during the learning process, the individual agent gradually finds that the rewards of inaction strategies in the task are more advantageous than the rewards of other strategies, and turns to other strategies that are more beneficial to the rewards. When the evolution is stable, cooperation strategies are rarely used. The changing trend of the proportion of the competitive strategy is opposite to that of the cooperation strategy. When $\beta = 1.06$, the competitive strategy is almost not adopted. When other β values are taken, the competitive strategy is the main strategy adopted by individual agents. The inaction strategy has the least probability to be adopted under the new game learning rules.

As shown in Fig. 7(g) and Fig. 7(h), when the evolution is stable, under different β values except $\beta = 1.06$, the group intelligence level firstly increases significantly and then remains stable. The group intelligence level under other β values decreases significantly. The change of the intelligence level is related to the trend of the proportions of cooperation strategies. If the proportion of individual agents choosing a competition strategy shows an upward trend, the group intelligence level also shows an upward trend.

In a word, the game learning rules of the negative feedback tax penalty mechanism play a more significant role in limiting the adoption of inaction strategies. Regardless of the value of β, the proportion of individual agents choosing an inaction strategy is almost 0, because the secondary distribution of rewards is not good for agents who are not involved in the task. Under the majority of β values, the dominant strategy is the competition strategy. Compared with the original model, the improved model increases the number of people participating in the task, but it has a certain inhibitory effect on the intelligence level of the group. When $\beta = 1.06$, the dominant strategy is the cooperation strategy. The effect of guiding the group to evolve towards the cooperation mode is the best and the group intelligence level is also improved. When

the value of β is less than 1.06, the penalty effect is weakened. Although the group has a tendency to choose a cooperation strategy in the initial stage of the evolution, it is eventually replaced by a competition strategy, which limits the number of people who actually participate in the task and is not conducive to the improvement of the group intelligence level. It can be seen that although the negative feedback tax punishment mechanism increases the number of people participating in the task and has a certain chance to change the group evolution towards the cooperation direction, the dominant strategy is still the competition strategy. In other word, the negative feedbacktaxpunishmentmechanismcannotalwaysguidetheevolutionary game direction of the group towards the ideal situation. In reality, the formulators of the reward rule need to reasonably formulate the relevant parameters in the negative feedback tax penalty mechanism in order to better guide the group.

5. Improved model based on reinforcement learning algorithms

In the evolutionary game of individual agents in the group, an agent continuously exchanges information with other agents and makes decisions. The agent has a certain ability of autonomous learning. If the model is improved by combining the evolutionary game process with reinforcement learning, it can better guide the direction of the evolutionary game and realize a more ideal evolutionary situation. In the original model, all individual agents are assumed to be fully rational individuals. In reality, game individuals may not always be completely rational. Players sometimes do not follow the rules of game learning when choosing strategies. Therefore, in the improved model based on reinforcement learning, the bounded rationality of individuals will be reflected to some degree.

The original model is built in the Markov decision framework and recorded as the Markov process in a discrete finite state, $< S, A, r, p >$, where S and A are respectively discrete state space and action space, r is the reward function of the agent individual. When the individual participates in the task, r is determined by (14). When the individual is not involved in the task, r is determined by (15). p is a transition function determining the transition from one state to another when an individual chooses a certain strategy. However, in the evolutionary game process, the transfer function of the model is unknown, so a special reinforcement learning method, Q-learning algorithm, is required. It can learn without the known transfer function and be suitable for the combination with evolutionary games. In the Q-learning algorithm, the state value is not considered, but the value of the state-action pair $Q(s,a)$, namely, the role of selecting the action a in a certain state s, should be considered. The Q value is updated from time 1. At time t, the Q value of time $t-1$ is updated according to the following formula:

$$Q(s_{t-1}, a_{t-1}) = Q(s_{t-1}, a_{t-1}) + \alpha(r + \gamma \max Q(s_t, a_t) - Q(s_{t-1}, a_{t-1})) \quad (17)$$

where $\alpha \in [0,1]$ is the learning rate; γ is the discount factor; s_t and a_t are the states and behaviors at time t. Based on the model setting, the values of s_t and a_t are taken from corresponding state space S and behavior set A according to the above rules. The Q-learning algorithm generally uses ε greedy strategy to update strategy selection. In order to combine the Q-learning algorithm with the evolutionary game model, based on the consideration of the bounded rationality states of different individuals, the traditional ε greedy strategy is improved to obtain a bounded rationality evolutionary game strategy. The principle is shown in Fig. 8.

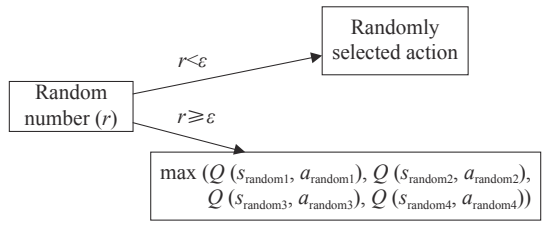


Fig. 8 Bounded rationality evolutionary game strategy

Under this strategy, all behaviors are selected with a non-zero probability ε. Due to the bounded rationality of an individual agent, it does not always learn from other individuals through the comparison of Q value in the selection of strategies. Therefore, in the evolutionary game strategy of bounded rationality, the individual randomly chooses the strategy in the next round of game with the probability of ε and compares the Q value with that of other four random agents with the probability of $1-\varepsilon$. Finally, the strategy with the largest Q value is chosen as the strategy in the next round of game.

In the single-task and multi-decision evolutionary game model based on multi-agent and reinforcement learning, an individual agent selects the strategy used in each round of the game through the Q-learning algorithm. According to the evolutionary game rules, the state of the reinforcement learning algorithm is set as a three-person strategy combination. According to Table 1, the action set or strategy set is A={<*Co Co Co*>, <*Co Co D*>, <*Co Co L*>, <*Co D D*>, <*Co L L*>, <*Co D L*>, <*Co Co Co*>, <*D D D*>, <*D L L*>, <*D D L*>, <*L L L*>}, ten types in total. There is no difference of the order among strategy combinations. For example, the game rules used by *CoCoD*, *CoDCo*,

and *DCoCo* are the same. The state space is $S = \{0.96, 0.98, 1.1, 1.02, 1.04, 1.06\}$, six types in total. The state space refers to the different values of β. The reward is determined by (14) and (15). The steps of the evolutionary game process are provided as follows:

Step 1 Initialize the Q value table.

Step 2 Play a non-repeating three-person random game.

Step 3 Calculate the rewards of all individual agents according to (14) and (15).

Step 4 Select the strategy for the next round of games based on the bounded rationality evolutionary game strategy.

Step 5 Update Q table according to (17).

Step 6 Repeat Steps 2−5. Continue to play a new round of games until the specified number of rounds and then stop the game process.

The simulation is performed with the improved model under the reinforcement learning algorithm. Six values of β (0.96, 0.98, 1, 1.02, 1.04 and 1.06) are simulated respectively. The number of individual agents in the group is 500 and the reward-cost ratio is 3. The evolutionary game involves 1 000 rounds and the simulation is repeated 300 times.

The error of Q value under different β values can converge. In one simulation, the change of Q value error under the β value of 1 is shown in Fig. 9.

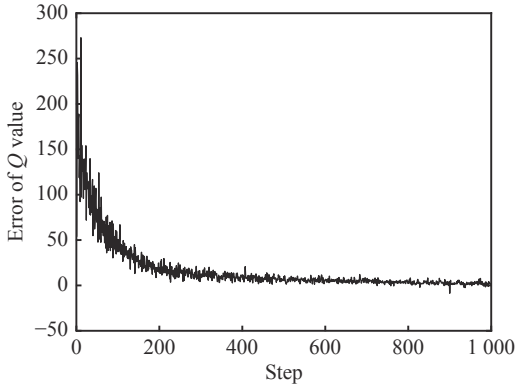


Fig. 9 Error of Q value

It can be seen from Fig. 9 that the Q value error firstly fluctuates greatly, then gradually decreases and finally becomes stable with the increase in the rounds of the evolutionary game.

In a simulation of the evolutionary game under the β value of 0.96, Fig. 10 shows the proportions of individual agents choosing different strategies in the group. It can be seen that in the evolutionary game process, the proportions of individual agents choosing each strategy fluctuate slightly. The proportion of cooperation strategies shows a clear upward trend, whereas the proportions of competition and inaction strategies show significant downward trends. The strategy with the highest proportion is the cooperation strategy and the strategy with the lowest proportion is the inaction strategy. The above results show that the introduction of reinforcement learning algorithms can effectively guide the group to evolve towards the direction of cooperation. In order to clearly show the proportion of individual agents choosing each strategy in the stable evolutionary game process under different values of negative feedback factor β, the average of the proportion of individual agents choosing each strategy in the last 100 rounds of the evolutionary game is used as the final evolutionary game stability result.

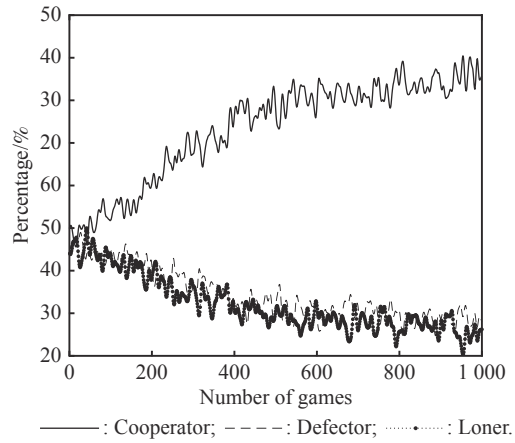
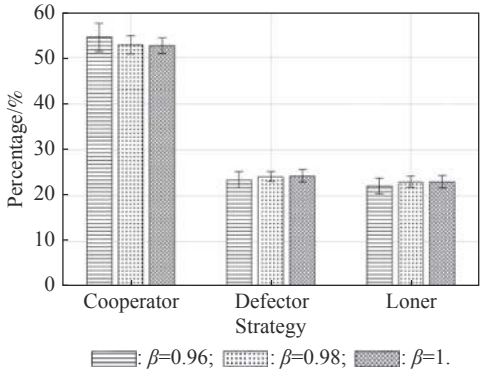
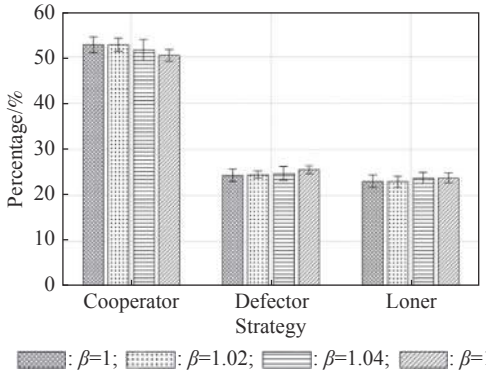


Fig. 10 Proportions of individual agents choosing different strategies in the evolutionary game

The proportions of individual agents choosing each strategy is shown in Fig. 11. It can be seen from Fig. 11(a) and Fig. 11(b) that under different β values, the proportion of individual agents choosing the cooperation strategy is significantly higher than those of individual agents choosing the competition and inaction strategies when the evolutionary game is stable. Moreover, the stable results of the proportions of individual agents choosing different strategies are not significantly different. The differences among the proportions of the same strategy under different β values are smaller after the reinforcement learning algorithm is introduced. The proportion of individual agents choosing the competition strategy tends to decrease as the value of β increases. When the evolutionary game is stable and β is 0.96, the proportion of individual agents choosing the competition strategy is the highest. When β is 0.96, the improved model allows the best guiding effect on the group evolutionary game. When β is 1.06, the proportion of individual agents choosing the competition strategy is the lowest.



(a) Proportions of individual agents choosing each strategy under different β values of 0.96, 0.98, and 1



(b) Proportions of individual agents choosing each strategy under different β values of 1, 1.02, 1.04, and 1.06

Fig. 11 Proportions of individual agents choosing each strategy under different β values in the stable evolutionary game process

When the value of β is 0.96, the evolutionary game effect of the model is the best. The evolutionary game results under the fixed selection value are compared with those based on the Nash-Q learning algorithm, the Monte Carlo method and the genetic algorithm. All the algorithms have 1 000 rounds of evolutionary games and are repeated 300 times. Among them, the genetic algorithm takes every 100 rounds as a generation, and the game learning at the end of each generation adopts the classic uniform crossover operation of the genetic algorithm. When the evolutionary game of different methods is stable, the proportion of each strategy is obtained (see Table 2).

Table 2 Percentage of each strategy

Algorithm	Proportion of cooperation strategies	Proportion of competition strategies	Proportion of inaction strategies
Algorithm of this article	0.54	0.23	0.23
Nash-Q learning algorithm	0.51	0.25	0.24
Monte Carlo method	0.47	0.27	0.26
Genetic algorithm	0.45	0.26	0.29

It can be seen from Table 2 that the algorithm proposed in this paper has the best effect on the evolutionary game of the model since its cooperation strategy accounts for the highest proportion.

According to the previous analysis, the trend of the intelligence level of the group is related to the trend of the proportion of individual agents choosing the competition strategy. In the evolutionary game under the same β value, the proportion of individual agents choosing the competition strategy shows the increasing trend, so the group intelligence level also increases. In order to clearly show the changes in the group intelligence level under different β values, the group intelligence level in the initial stage of the evolution is compared with that in the final stable stage. The average value of the group intelligence level in the first 100 rounds is taken as the result of the starting stage and the average value of the last 100 rounds is taken as the final stable result. The group intelligence level is shown in Fig. 12.

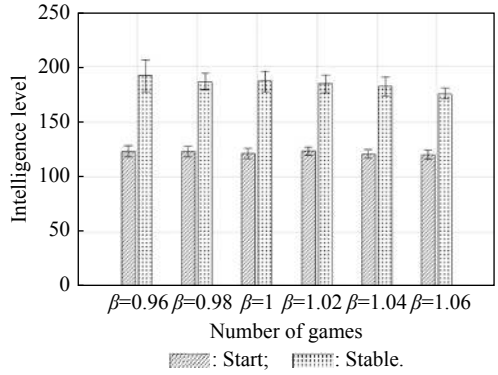


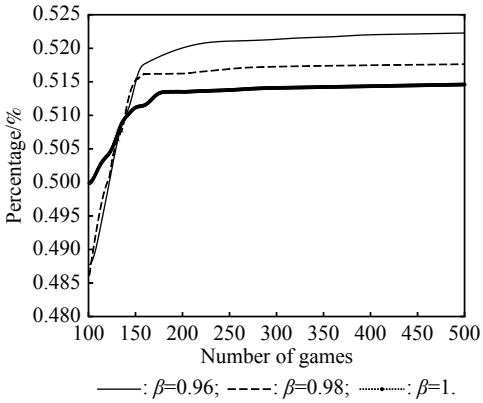
Fig. 12 Group intelligence level in the initial stage and end of the evolutionary game under different values of β

It can be seen from Fig. 12 that under different β values, the group intelligence level in the stable stage of the evolutionary game is significantly higher than that in the initial stage of the evolutionary game. The group intelligence level is the highest under the β value of 0.96 and the lowest under the β value of 1.06.

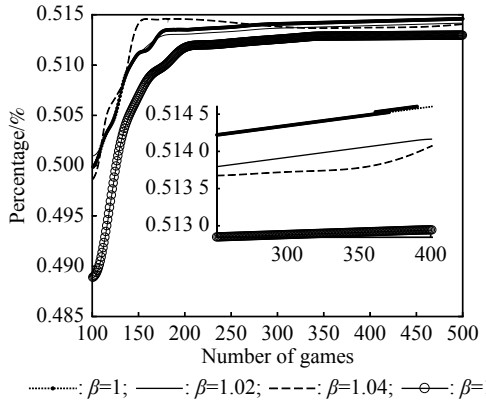
The simulation results show that the single-task multi-decision evolutionary game model based on multi-agent and reinforcement learning can effectively guide the group's evolutionary game direction to evolve towards the ideal situation under the different effects of the negative feedback tax penalty mechanism. The improved model improves the group intelligence level and promotes the completion of the task. In reality, if the group performs an evolutionary game according to this model, the formulator of the income rules may not pay too much attention to relevant parameters in the negative feedback tax penalty mechanism, which can effectively guide the evolutionary game effect of the group. When the task completion requirements are high, relevant parameters in the negative feedback tax penalty mechanism should be

considered in the formulation of reward rules.

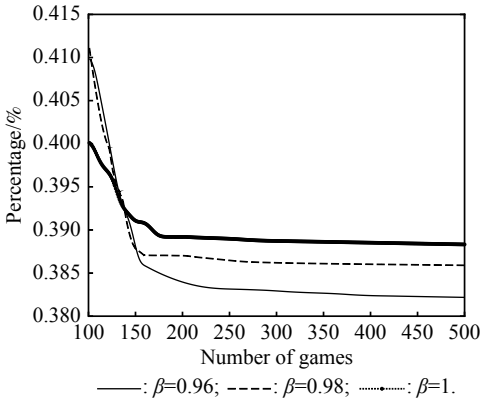
The above simulation results are based on a group composed of a fixed number of people. In order to study the effect of the size of the group on the simulation results under different feedback factor values, 1 000 rounds of evolutionary game simulation experiments in the range of [100,500] are performed and repeated 500 times. Negative feedback factor β is respectively set to be 0.96, 0.98, 1, 1.02, 1.04, and 1.06 and the reward-cost ratio is 3. The average of the last 100 evolutionary game results is used as the stable evolution result and 500 repeated experimental results are averaged. Fig. 13 shows the changes in the proportion of individual agents choosing each strategy as well as the average individual intelligence level in the stable evolution under different values of the feedback factor β. In order to more clearly display the changes under different reward-cost ratios, partial enlarged views are added in Fig. 13(b), Fig. 13(d), Fig. 13(f), and Fig. 13(h).



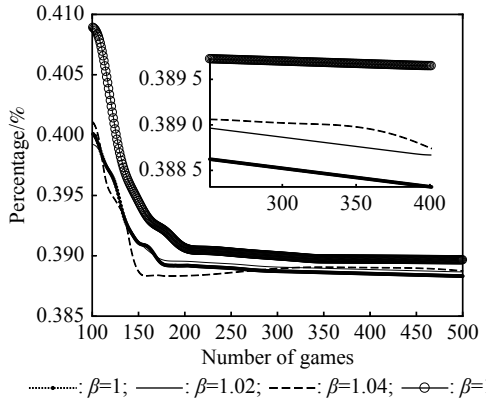
(a) Proportions of individual agents choosing the cooperation strategy under different β values of 0.96, 0.98, and 1



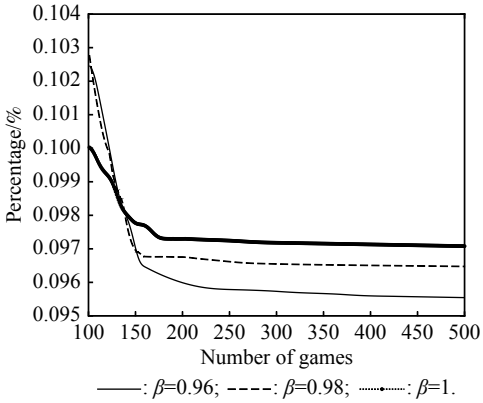
(b) Proportions of individual agents choosing the cooperation strategy under different β values of 1, 1.02, 1.04, and 1.06



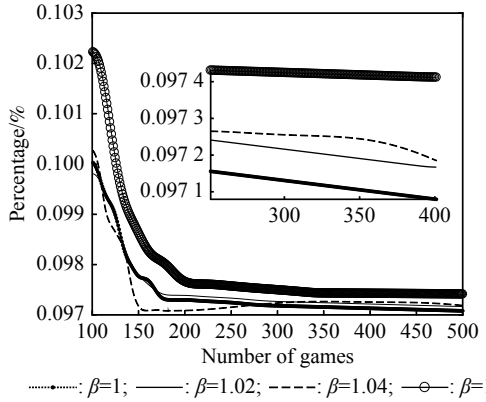
(c) Proportions of individual agents choosing the competition strategy when β is 0.96, 0.98, 1



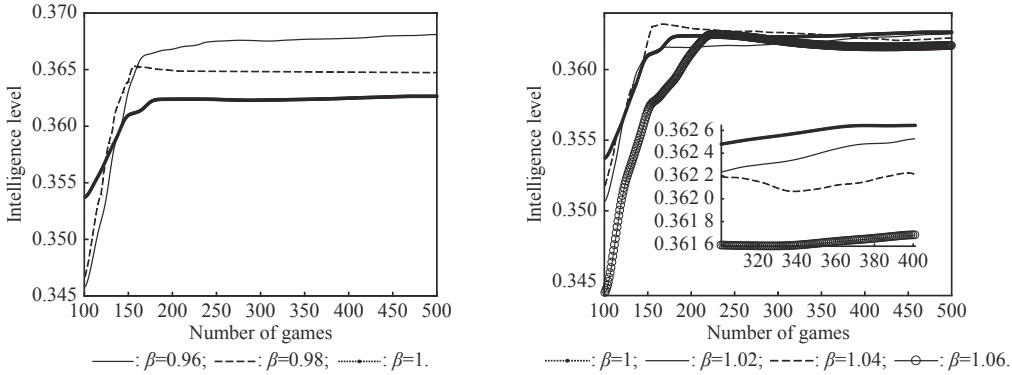
(d) Proportions of individual agents choosing the competition strategy under different β values of 1, 1.02, 1.04, and 1.06



(e) Proportions of individual agents choosing the inaction strategy under different β values of 0.96, 0.98, and 1



(f) Proportions of individual agents choosing the inaction strategy under different β values of 1, 1.02, 1.04, and 1.06



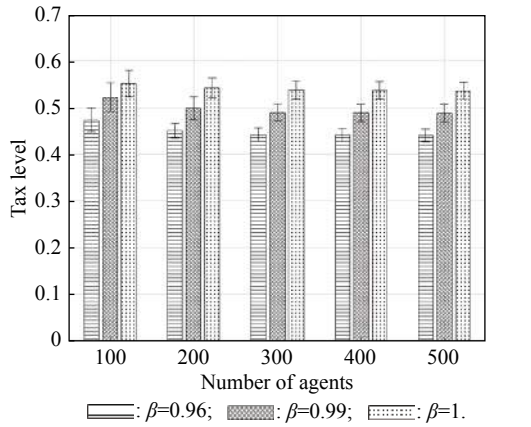
(g) Average individual intelligence levels under different β values of 0.96, 0.98, and 1

(h) Average individual intelligence levels under different β values of 1, 1.02, 1.04, and 1.06

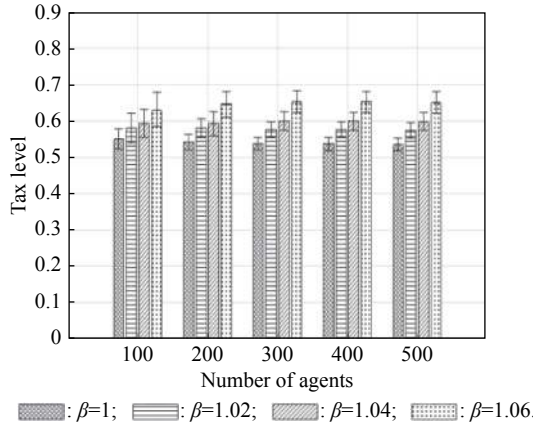
Fig. 13 Variations of the proportions of individual agents choosing each strategy and the average individual intelligence level with group size under different β values

Fig. 13(a)−Fig. 13(f) show that under different β values, the proportion of individual agents choosing the cooperation strategy in the stable evolution increases slightly as the group size increases. The proportions of individual agents choosing the competition strategy and the inaction strategy show downward trends. With the increase in the group size, the increasing or decreasing trend of the proportion of each strategy increases and the advantages of cooperation strategies become more significant. When the group size increases to 300 or more, the proportions of individual agents choosing different strategies remain stable. The further increase in the number of the individual agents will no longer affect the final proportion of each strategy. The dominant strategies under different β values are cooperation strategies and the strategies with the smallest proportion are inaction strategies. As the value of β increases, the proportion of cooperation strategies decreases, whereas the proportions of competition and inaction strategies increase. When the evolutionary game is stable and β is 0.96, individual agents choosing the cooperation strategy accounts for the largest proportion. When β is 0.96, the improved model shows the best guiding performance in the group evolution game. When β is 1.06, the proportion of individual agents choosing the cooperation strategy is the lowest. It can be seen from Fig. 13(g) and Fig. 13(h) that when the evolution is stable, with the increase in the group size, the average individual intelligence level increases slightly and remains stable. When the group size increases to above 300, the average individual intelligence level is no longer affected by the group size. When the evolutionary game is stable and the value of β is 0.96, the individual intelligence level is the highest and the rules of the game at this time have the best effect on the group's cooperation and guidance.

Fig. 14 shows the variations of the penalty tax rate in the evolutionary game process with the group size in the stable evolution under different β values.



(a) Tax rate level under different β values of 0.96, 0.98, and 1



(b) Tax rate level under different β values of 1, 1.02, 1.04, and 1.06

Fig. 14 Variation of tax rate levels with group size under different β values

As shown in Fig. 14, as the value of β increases, the penalty tax rate also increases and its value slightly decreases with the increase in the group size. If more people actually participate in the task, the value of the penalty tax rate decreases, but the effect of the decrease is relat-

ively weak.

As shown in Fig. 13 and Fig. 14, the increase in the penalty tax rate increases the proportion of individual agents choosing the cooperation strategies and the average individual intelligence level, and the change in the group size has a limited impact on the level of the penalty tax rate and the proportion of individual agents choosing each strategy. With the increase of the group size, the evolutionary game direction of the group can be better guided towards the ideal situation, but the guiding effect reaches its limit when the group size increases to a certain degree.

6. Conclusions

In this paper, a multi-decision evolutionary game task is constructed to study the evolution rules of groups in the game process. After introducing the concept of multi-agents into the evolutionary game process, a tax rate penalty mechanism with negative feedback characteristics is proposed to guide the selection of individual strategies in the group. In addition, a calculation method of the group intelligence level is defined to evaluate the result of the group evolution game. The simulation results show that the value of the negative feedback factor determines the strategy selection result of individual agents to a certain degree. When the value of the negative feedback factor β is 1.06, it can effectively increase the proportion of individual agents choosing cooperation strategies as well as the group intelligence level and guide the group to evolve towards the ideal evolution direction. The tax rate punishment mechanism has limited guidance in the evolutionary game process. Most of the values of negative feedback factor β cannot effectively improve the evolutionary game results of the group. To solve this problem, this paper proposes a single-task multi-decision evolutionary game model based on multi-agent and reinforcement learning. The model combines multi-agents with Q-learning algorithms in the process of evolutionary games, improves the selection strategy of Q-learning, and proposes a bounded rationality evolutionary game strategy. This learning strategy not only reflects the rules of evolutionary games, but also takes into account the bounded rationality of individual agents. The simulation results of the model show that in the stable stage of the evolutionary game, different values of the negative feedback factor β can guide the evolutionary game to evolve towards the cooperation direction and improve the group intelligence level. The simulation results also confirm the impact of the group size on the model. The increase in the group size can increase the proportion of individual agents choosing cooperation strategies as well as the group intelligence level to a certain degree. However, when it increases to a certain scale, it will no longer affect the results of evolutionary games.

In summary, the model proposed in this paper can effectively guide the evolution direction of the group in the single-task multi-decision game and explores the penalty tax rate and the group size in the evolutionary game.

References

[1] SUTTON R, BARTO A. Reinforcement learning: an introduction. Cambridge: MIT Press, 1998.
[2] AWHEDA M D, SCHWARTZ H M. The residual gradient FACL algorithm for differential games. Proc. of the Canadian Conference on Electrical and Computer Engineering, 2015: 1006−1011.
[3] JELAI Z. Reinforcement learning based human-prosthetic robot interaction control in movement therapy. Proc. of the International Conference on New Technologies, Development and Application, 2020: 172−181.
[4] LITTMAN M L. Markov games as a framework for multi-agent reinforcement learning. Proc. of the 11th International Conference on Machine Learning, 1994: 157−163.
[5] LI Y, HAN W, WANG Y Q. Deep reinforcement learning with application to air confrontation intelligent decision-making of manned/unmanned aerial vehicle cooperative system. IEEE Access, 2020, 8: 67887–67898.
[6] DEPTULA P, BELL Z I, DOUCETTE E A, et al. Data-based reinforcement learning approximate optimal control for an uncertain nonlinear system with control effectiveness faults. Automatica, 2020, 116: 108922.
[7] GOTTSCHALK S, BURGER M. Differences and similarities between reinforcement learning and the classical optimal control framework. Proceedings in Applied Mathematics and Mechanics, 2019, 19(1): e201900390.
[8] LIAO H C, LIU J S. A model-based reinforcement learning approach to time-optimal control problems. Proc. of the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, 2019: 657−665.
[9] SHI H B, ZHAI L J, WU H B, et al. A multi-tier reinforcement learning model for a cooperative multi-agent system. IEEE Trans. on Cognitive and Developmental Systems, 2020, 12(3): 636–644.
[10] NGUYEN N D, NGUYEN T, NAHAVANDI S. Multi-agent behavioral control system using deep reinforcement learning. Neurocomputing, 2019, 359(24): 58–68.
[11] QIE H, SHI D, SHEN T, et al. Joint optimization of multi-UAV target assignment and path planning based on multi-agent reinforcement learning. IEEE Access, 2019, 7: 146264–146272.
[12] FIRDAUSIYAH N, TANIGUCHI E, QURESHI A G. Modeling city logistics using adaptive dynamic programming based multi-agent simulation. Transportation Research Part E: Logs and Transportation Review, 2019, 125: 74–96.
[13] REN Y, FAN D M, FENG Q, et al. Agent-based restoration approach for reliability with load balancing on smart grids. Applied Energy, 2019, 249: 46–57.
[14] MYERSON R B. Game theory: analysis of conflict. Cambridge: Harvard University Press, 1997.
[15] NIE L, WANG X G, PAN F Y. A game-theory approach based on genetic algorithm for flexible job shop scheduling problem. Journal of Physics: Conference Series, 2019, 1187: 032095.

[16] WANG X H, ZHONG X X, LI L, et al. PSOGT: PSO and game theoretic based task allocation in mobile edge computing. Proc. of the IEEE 21st International Conference on High Performance Computing and Communications, 2019. DOI: 10.1109/HPCC/SmartCity/DSS. 2019.00318.
[17] XU L, HU B, GUAN Z Z, et al. Multi-agent deep reinforcement learning for pursuit-evasion game scalability. Proc. of the Chinese Intelligent Systems Conference, 2020: 658−669.
[18] ABDOOS M. A cooperative multi-agent system for traffic signal control using game theory and reinforcement learning. IEEE Intelligent Transportation Systems Magazine, 2020. DOI: 10.1109/MITS. 2020.2990189.
[19] BENDOR J, MOOKHERJEE D, RAY D. Reinforcement learning in repeated interaction games. Advances in Theoretical Economics, 2001, 3(2): 159–174.
[20] CRANDALL J W, GOODRICH M A. Learning to compete, coordinate, and cooperate in repeated games using reinforcement learning. Machine Learning, 2011, 82: 281–314.
[21] HU J L, WELLMAN M P. Multiagent reinforcement learning: theoretical framework and an algorithm. Proc. of the 15th International Conference on Machine Learning, 1998: 242−250.
[22] LIU H, LI J F, GE S Y, et al. Coordinated scheduling of grid-connected integrated energy microgrid based on multi-agent game and reinforcement learning. Automation of Electric Power Systems, 2019, 43(1): 58–66. (in Chinese)
[23] XU L, ZHO Z J. Channel and power allocation algorithm based on distributed cooperative Q learning. Computer Engineering, 2019, 45(6): 166–170, 180. (in Chinese)
[24] MATTA M, CARDARILLI G C, NUNZIO L D, et al. Q-RTS: a real-time swarm intelligence based on multi-agent Q-learning. Electronics Letters, 2019, 55(10): 589–591.
[25] CHEN Y, LIU J M, ZHAO H. Social structure emergence: a multi-agent reinforcement learning framework for relationship building. Proc. of the 19th International Conference on Autonomous Agents and Multi-Agent Systems, 2020: 1807−1809.
[26] GE Y Y, ZHU F, HUANG W, et al. Multi-agent cooperation Q-learning algorithm based on constrained Markov game. Computer Science and Information Systems, 2020, 17(2): 647–664.
[27] DAEICHIAN A, HAGHANI A. Fuzzy Q-learning based multi-agent system for intelligent traffic control by a game theory approach. Arabian Journal for Science and Engineering, 2018, 43(6): 3241–3247.
[28] ULUSOY U, GUZEL M S, BOSTANCI E. A Q-learning-based approach for simple and multi-agent systems. Multi-Agent Systems-Strategies and Applications, 2020. DOI: 10.5772/intechopen. 88484.
[29] HOFBAUER J, SIGMUND K. Evolutionary games and population dynamics. Cambridge: Cambridge University Press, 1998.
[30] NOWAK M A. Evolutionary dynamics: exploring the equations of life. Cambridge: Harvard University Press, 2006.
[31] SMITH J M. Evolution and the theory of games. Cambridge: Cambridge University Press, 1982.
[32] KIMURA M. The neutral theory of molecular evolution. Cambridge: Cambridge University Press, 1983.
[33] CHEN Z H, YANG Z H, WANG H B, et al. Overview of reinforcement learning from knowledge expression and handling. Control and Decision, 2008, 23(9): 962–975. (in Chinese)
[34] GAO Y, CHEN S F, LU X. Research on reinforcement learning technology: a review. Acta Automatica Sinica, 2004, 30(1): 86–100. (in Chinese)

Biographies

MA Ye was born in 1993. She received her master 's degree from the Academy of Army Armored Force, Beijing, in 2017. Currently, she is a Ph.D. candidate at the Academy of Army Armored Force. Her research interests include intelligent technology of control system, and modeling and simulation of complex systems.
E-mail: mayegf@126.com

CHANG Tianqing was born in 1963. He received his Ph.D. degree in concurrent engineering from Tsinghua University in 1999. Since 2000, he has been a professor with the Academy of Army Armored Force. His current research interests include target detection and recognition, as well as navigation, guidance and control.
E-mail: oliver_chan1214@126.com

FAN Wenhui was born in 1968. He received his Ph.D. degree in control science and engineering from Zhejiang University in 1998. Currently, he is a professor, doctoral tutor, and vice-director in the Department of Automation, Tsinghua University. His research interests include modeling and simulation of complex systems, product information integration modeling technology, product life-cycle management technology, and collaborative design platform technology.
E-mail: fanwenhui@tsinghua.edu.cn