Learning and Evolution Affected by Spatial Structure

Masahiro Ono[1] and Mitsuru Ishizuka[1]

[1]Graduate School of Information Science and Technology,
The University of Tokyo,
7–3–1 Hongo, Bunkyo–ku, Tokyo 113–8656, Japan
{mono, ishizuka}@mi.ci.i.u-tokyo.ac.jp

Abstract. In this study, we explore the roles of learning and evolution in a non-cooperative autonomous system through a spatial IPD (Iterated Prisoner's Dilemma) game. First, we propose a new agent model playing the IPD game; the game has a gene of the coded parameters of reinforcement learning. The agents evolve and learn during the course of the game. Second, we report an empirical study. In our simulation, we observe that the spatial structure affects learning and evolution. Learning is not effective for achieving mutual cooperation except under certain special conditions. The learning process depends on the spatial structure.

Key words:game theory, prisoner's dilemma, small world network

1 Introduction

From the scientific viewpoint, learning, evolution, and the interaction between them have been studied. One of the famous earlier works is the Baldwin effect [1], which hypothesized that the characteristics of how an individual learns affect the evolution of a species.

Moreover, these two concepts are worth studying from an engineering viewpoint. Recently, the autonomous agent has been studied, and there are systems that consist of many autonomous individuals, such as the multi-agent system. In general, agents adapt to the environment owing to learning in addition to their hard-coded program. These two functions correspond to learning and evolution (a part of the hard-coded program), as described earlier. In this area, the focus has mainly been on the manner in which the agents cooperate. However, the agents are not always cooperative, for example, in the case of agents that belong to different organizations and whose payoffs conflict. We forecast that there would be two types of systems-the cooperative and non-cooperative. Studies based on the abstract model can provide useful insights for many possible application systems both in the cooperative and non-cooperative cases.

One of the non-cooperative relations is PD (Prisoner's Dilemma) from game theory. This game has been studied thus far because the game itself arouses our curiosity and is suitable for studying learning and evolution. Further, the evolutionary game theory pertains to evolution. In general, the equilibrium point of the game has been studied in the field of game theory. The evolutionary game theory assumes an alternative equilibrium point ESS (Evolutionary Stable Strategy). In addition, numerical as well as the analytical approaches exist. One of the numerical approaches is that given by Lindgren [2]. His model expresses a meta-strategy that determines the subsequent moves based on the game history. The learning game has been studied in accordance with the development in the area of reinforcement learning. For example, it is shown that the reinforcement learning agents in the multi-agent system cause instability due to mutual learning [3]. There are a few studies that deal with both learning and evolution. Thus far, the combination of learning and evolution were adapted to the stochastic game [4] and Baldwin effect [5] under the simple learning condition.

With regard to the original concern of IPD (Iterated Prisoner's Dilemma), the basic issue pertains to the contradiction between the mathematical solution where the non-cooperative strategy is stable and the phenomenon in the real world, i.e., we often cooperate with each other. This issue has been studied for a long time, and a spatial structure was introduced in the evolutionary game theory. Players are located on the spatial structure, for example, a two-dimensional regular lattice [6][7] or a small-world network model [8][9][10]. They evolve in every generation after they play games across neighborhoods. In this case, the emergence of a cooperative strategy is observed after a period of dominance of the non-cooperative strategy. Thus far, investigation on the learning and evolution of the spatial structure through studies has been sufficient.

In this paper, we investigate the roles of learning and evolution in relation to the emergence of cooperation. In particular, we focus on the spatial structure that limits the communication among agents. This assumption is reasonable in the case of agents on the Internet. Our work may help design an agent to improve the system performance from an engineering viewpoint. In addition, it also contributes to science: the human characteristic to make a decision is affected by the spatial structure. In the subsequent sections, we first provide brief backgrounds of the Prisoner's Dilemma, mechanism of reinforcement learning, and small-world network model. We then describe a model that we have designed for combining the functions of learning and evolution. Experimental results are then presented and discussed. Finally, we present our conclusions and directions for future work.

2 Backgrounds

Here, we describe IPD, reinforcement learning, and the small-world network model. We consider learning and evolution in the IPD game. Reinforcement learning is the method of learning the selection of a move in the game. The small-world network is a model that expresses the spatial structure.

2.1 Iterated Prisoner's Dilemma

The Prisoner's Dilemma is the most popular game in game theory because it is an elegant model to express many social phenomena. The name and the typical payoff matrix of this game were given by Albert Tucker in the 1950s.

In a symmetric two-player game, the payoff matrix of Prisoner's Dilemma is expressed as Table 1, where R, T, S, and P represent the reward, temptation, sucker, and punishment payoffs, respectively. The payoff relations ($T > R > P > S$, $2R > T + S$) that exist among them raise a dilemma.

Both players in the game would select the defect strategy if they are assumed to be rational. Player 1 considers that he should defect and earn a higher payoff whenever player 2 cooperates or defects. Player 2 would also defect after the same consideration. In the end, each player defects, and (D, D) is the only Nash equilibrium in this game. However, this state is a pareto inferior and therefore it is not an optimal strategy for either player. Hence, this game poses a dilemma.

Table 1. The payoff matrix of Prisoner's Dilemma

		Player2	
		C (cooperate)	D (defect)
Player 1	C	R / R	T / S
	D	S / T	P / P

($T > R > P > S$, $2R > T + S$)

2.2 Reinforcement Learning

Reinforcement learning is a type of machine learning. Assuming that an agent takes an action and receives a reward in return from the environment, the reinforcement learning algorithm attempts to find a policy for maximizing the agent's cumulative reward.

Essentially, an agent has inner states. The learning process is as follows: the agent selects an action in a state according to each value function, then the action in the state is evaluated and the value function is updated. The agent performs a state transition and selects the next action in the state. This process is repeated and the agent continues to refine the action-value function. Although there are several algorithms for reinforcement learning, in this study, we consider the one given by Sarsa [11].

2.3 Small-World Network

In 1998, Duncan Watts defined the small-world network using two characteristic parameters–characteristic *characteristic path length* and *clustering coefficient* [12].

The *characteristic path length* L is the average of the shortest path length between any two vertices on the network. The *clustering coefficient* C indicates the extent to which vertices adjacent to any vertex are adjacent to each other as an average.

If it is assumed that the number of vertices and edges on a graph are fixed, the structure changes as the randomness parameter p of the network varies. Each vertex is connected to its neighborhoods mutually in the case of $p = 0$. The edges change stochastically as p increases. In case of $p = 0$, the network is referred to as a regular network and both L and C are large. On the other hand, a random network appears at $p = 1$, where L and C are small. In the middle, between these two extremes, the network exhibits the property of a small-world network, where L is small and C is large.

3 Proposed Model

We propose a new agent model that learns and evolves for the IPD game. The process by which the agent selects the next move is as follows:

We assume that the agent remembers the moves from the previous game in iterated games and he selects the next move in the manner of the first-order meta-strategy. The agent has four inner states (own previous move and opponent's previous move) = CC, CD, DC, and DD derived from the possible combinations of the moves from the previous game. In other words, the agent selects the next move in a state based on the previous game.

The agent has a gene in which two components are coded. One is the part that expresses the initial action-values for each of the possible combinations of actions and states. The other is the part for parameters of reinforcement learning. The created agent has initial action-values and a reinforcement learning function with parameters derived from the coded gene. He can play the IPD games and update the action-values. In other words, he can learn the game.

Primarily, when the agent selects a move in the game, he compares the action-values of each action in the states and picks up the greater one; as shown in Eqn. (1):

$$\text{next move} = \begin{cases} C, & Q(s, C) \geq Q(s, D) \\ D, & Q(s, C) < Q(s, D) \end{cases} \tag{1}$$

where $Q(s, a)$ evaluates the action a in the state s, and s is a possible state. For example, in the state CC, the next move is C in the condition $Q(CC, C) \geq Q(CC, D)$. Since the agent has information on all the combinations of actions and states, all possible first-order meta-strategies can be expressed. The examples are listed in Table 2.

The agent selects a move regardless of the previous moves in the case of all C and D. Tit-for-Tat (TFT) created by Anatol Rapoport is a famous strategy that repeats the opponent's previous move. This strategy won the famous iterated prisoner's dilemma tournament organized by Robert Axelrod in the year 1981

Table 2. First-order meta-strategy examples

PM	OPM	Strategy examples: All C	All D	TFT	Pavlov
C	C	C	D	C	C
C	D	C	D	D	D
D	C	C	D	C	D
D	D	C	D	D	C

PM: Previous Move, OPM: Opponent's Previous Move

[13]. The Pavlov strategy [14]-also known as the "Win-Stay, Lose-Shift" strategy-selects the move opposite to the previous move when the agent cannot earn a high reward.

In return for the action, the agent receives a reward (the payoff of the game) and updates the action-values. The rule for updating the action-values is denoted by Eqn.(2). $Q(s_t, a_t)$ evaluates the action a_t in the state s_t including the next action-value $Q(s_{t+1}, a_{t+1})$:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right] \tag{2}$$

where α is the learning rate; r_{t+1}, a reward in return for the action a_t; and γ, the discount rate. Here, a_{t+1} is determined by the ϵ-greedy method in which the next move primarily depends on the action-values but is selected randomly according to a possibility ϵ. The agent obeys the action-value completely; it selects a move in a deterministic manner in the case of $\epsilon = 0$. Otherwise, the agent selects randomly and the selection depends on ϵ to some extent.

This process of updating is repeated during each selection. This implies that the agent learns through the iterated games. Since the strategy depends on the action-values as described before, the updating of the action-values indicates a change in the strategy, for example, from TFT to Pavlov. The characteristics of the learning process are the speed of learning, dependence on the reinforcement learning parameters, etc.

In this model, the eligibility trace parameter λ is also taken into account. The agent has a gene for these four parameters α, γ, λ, and ϵ.

4 Experimental Studies

The purpose of this experiment is to investigate the influence of spatial structure on learning and evolution. We carry out the experiment in two different spatial structure cases–a pool case and network case. Our simulation is performed as follows:

4.1 The network case

This is the case with a spatial structure that limits the communication among agents to some extent. The small-world network, in which randomness parameter p is given as an initial value and fixed during the simulation, is structured at first. The network consists of n agents, each of which has m links. These parameters are fixed at

The agents act as the players of the IPD game. They play games with other agents having a direct connection on the network. A unit of a game is an IPD game that is iterated 100 times. This does not pose a problem as the agents do not employ backward induction because they cannot determine the end of the IPD game in this simulation. It is assumed that a random noise involved in move selection by the agent is reversed according to a noise probability. The order in which the agents play games is determined randomly because the fixed order in each generation would influence the learning.

A generation change occurs after all the games in every generation end. A total of 20% of the agents are to die stochastically according to their gain earned in the generation. Then a new agent is to be located there in lieu of the dead agent; the new agent is generated as a copy of the agent that earned the most in the first neighborhoods of the dead agent. The mutation of the gene is generated at a copying probability of 0.02%. The payoff parameters in the simulation are set as $(T, R, P, S) = (5, 3, 1, 0)$ in Table 1.

4.2 The pool case

Here, we assume a pool of agents without a spatial structure. In this case, the agents are not linked and they have the opportunity to meet any agent. Although the general procedure is similar to the network case as described before, there are two differences:

The first one determines the opponents they play games with. Since the spatial property is without a structure, there is a possibility that they might play a game with only one agent amongst all agents. However, an agent plays games m times with randomly selected agents in order to provide a chance for learning with an equivalent frequency for the network case.

The second one is related to the generation change. In this case, we adopt tournament selection (tournament size: 2), two-point crossover, and mutation. Essentially, the parameters in relation to the generation change are the same as that in the network case.

In this study, we fix two parameters, $\epsilon = 0.1$ and $\lambda = 0$ as the first step. A total of 2,000 generations are repeated in the simulation. The average of a trial is the value that is obtained by averaging the values from 1,000 to 2,000 generations. The results shown below are the average of over 10 tries.

4.3 Results

We conducted simulations by varying the parameters of noise probability.

At first, we focus on the average gain, which is the index indicating the number of cooperators existing in the population. An approximate index value of 3 indicates the existence of many cooperators.

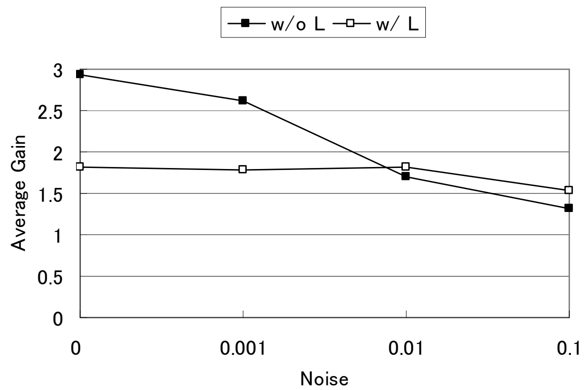

Fig. 1. Average gain in the pool case

The average gain in the pool case is shown in Fig. 1. "L" in the legend represents "learning." The difference between the cases with and without learning is clear in this figure. In the case with learning, the value remains almost constant regardless of the noise probability. On the other hand, in the case without learning, the average gain is about 3 when the noise does not exist. This implies that almost all agents cooperate with each other. The line decreases monotonically from the value 3, as the noise increases. It is observed that the lines intersect when the noise is approximately 1.0×10^{-2}. This implies that the agents cooperate well based on a condition. In this case, the condition is the noise probability.

The average gain in the network case is shown in Fig. 2. The number in the legend represents the randomness parameter of the small-world network. It is evident that all points in the no-learning case are higher than the ones in the learning case. The lines of the no-learning case decrease monotonically and vary when noise $= 0.1$. However, the lines of the learning case remain almost constant in the range 2.0$\sim$2.4, irrespective of the change in the noise or the randomness parameter of the network structure. With regard to the randomness of the network, the higher value lines are lower than the lower value lines in both learning and no-learning cases.

The fact that no-learning case is better than learning case, except under a few special conditions, implies that learning does not always make agents happy. In our experiments, the exception is in the limited condition that is in the range of higher noise in the pool case. The agents with learning ability tend to be extremely greedy; they tend to defect to get higher rewards. In the end, they select Nash equilibrium in the traditional PD game. Learning is a useful

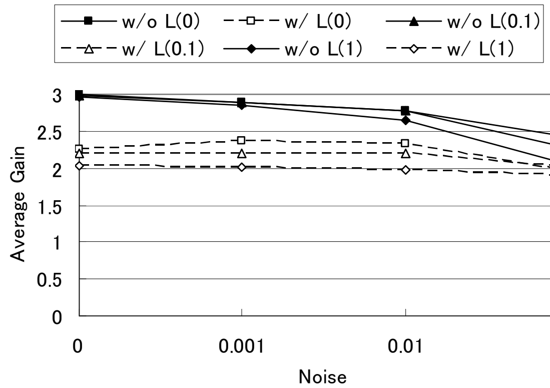

Fig. 2. Average gain in the network case ($p = 0.1$)

ability, but learning without additional wisdom, such as a general social rule or anticipation of the opponent's action, makes agents non-cooperative.

Next, the focus is on the learning process. In addition to the reinforcement parameters α and γ, we define an index "ACR (action change rate)" to investigate the effect of learning. ACR is the rate of the action difference in each possible state between the strategy changed by learning and the initial strategy decided by the gene.

Figs. 3 and 4 depict the case of pool and the small-world network of $p = 0.1$ respectively.

The ACR line is maintained at about 0.3 in both cases, as shown in Figs. 3 and 4. This implies that there are no stable strategies by nature. Learning always plays a constant role to adapt to the environment.

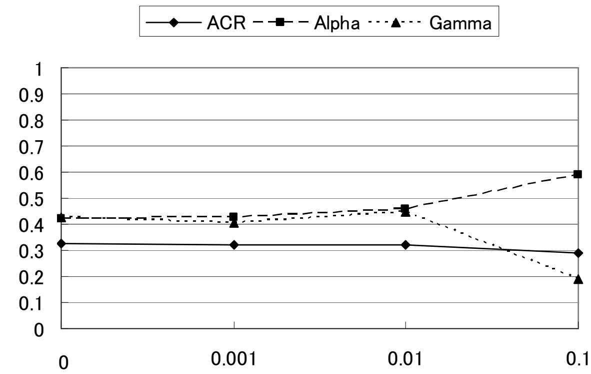

Fig. 3. Reinforcement learning parameters in the pool case

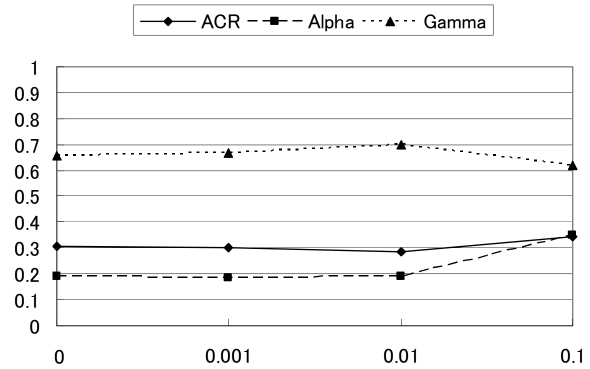

Fig. 4. Reinforcement learning parameters in the network case

With regard to α, γ of the variable reinforcement parameters of agents, their variations are shown in Figs. 3 and 4. α could be considered as the speed at which the value function changes. γ expresses the evaluation of the gain that would be received in the future. In the pool case, α and γ are approximately 0.4 when the range of noise rate is less than 0.01. At noise = 0.1, α increases and γ decreases suddenly.

On the other hand in the network case, until noise = 0.01, α is 0.2 and γ is about 0.7. Then, similar to the pool case, α increases and γ decreases.

Comparing the two cases, the values of α and γ are clearly different. These data imply that, in the network case, agents evaluate the gain that would be received in the future and they do not change their strategy rapidly; this is because they tend to maintain their relationships with other agents and they can expect that there is a high possibility that the opponent uses the same strategy. On the other hand, in the pool case, the agents evaluate the gain received at the moment, and they change their strategy rapidly because they face unknown opponents in almost all games. It is noteworthy that the learning process differs considerably depending on the spatial structure, regardless of the fact that each ACR is almost identical.

The characteristics of the lines represent the behavior of the noise. At a lower noise, the values are almost constant; however, α increases and γ decreases at the highest noise point. These results are explained by the same reason, that is, uncertainty of the opponent's strategy.

5 Conclusion

We studied the Prisoner's Dilemma game on the spatial structure played by agents who learn and evolve.

We proposed an agent model incorporating learning and evolution of strategy. These experiments revealed that the spatial structure of the agents influences cooperation and learning and evolution. In comparison with the learning case, agents in the no-learning case perform better, except under certain special conditions against our intuition. There are no stable strategies by nature and learning always plays a role in adaptation to the environment. The learning process is quite different depending on the spatial structure regardless of the fact that each ACR is almost identical.

For further study, the condition for effective learning is required to be clarified. In addition, the dynamic network case has to be considered. The next step is to bridge the gap between the pool and the network cases.

Although our study is still in a preliminary stage, it will contribute to the agent or system design policy in the future.

References

1. Baldwin, J.M.: A new factor in evolution. American Naturalist **30** (1896) 441–451
2. Lindgren, K.: Evolutionary phenomena in simple dynamics. Artificial Life II: Proceedings of the Workshop on Artificial Life held February 1990 in Santa Fe, New Mexico (1992) 295–312
3. Sandholm, T.W., Crites, R.H.: Multiagent reinforcement learning in the iterated prisoner's dilemma. Biosystems **37** (1996)
4. Hingston, P., Kendall, G.: Learning versus evolution in iterated prisoner's dilemma. In: Proceedings of the Congress on Evolutionary Computation, CEC2004. Volume 1., Portland, Oregon (2004) 364–372
5. Suzuki, R., Arita, T.: Interactions between learning and evolution: Outstanding strategy generated by the baldwin effect. Biosystems **77** (2004) 57–71
6. Nowak, M.A., May, R.M.: Evolutionary games and spatial chaos. Nature **359** (1992) 826–829
7. Lindgren, K., Nordahl, M.G.: Evolutionary dynamics of spatial games. Physica D **75** (1994) 292–309
8. Watts, D.J.: Small Worlds. Princeton University Press (1999)
9. Masuda, N., Aihara, K.: Spatial prisoner's dilemma optimally played in small-world networks. Physics Letters A **313** (2003) 55–61
10. Ono, M., Ishizuka, M.: Prisoner's dilemma game on network. In: Proceedings of the Eighth Pacific-Rim International Workshop on Multi-Agents, PRIMA2005, Kuala lumpur, Malaysia (2005) 9–22
11. Sutton, R.S., Barto, A.G.: Reinforcement Learning. MIT Press (1998)
12. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393** (1998) 440–442
13. Axelrod, R.: The Evolution of Cooperation. Basic Books (1985)
14. Nowak, M.A., Sigmund, K.: A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. Nature **364** (1993) 56–58