Finding a Prototype Form of Sustainable Strategies for the Iterated Prisoners Dilemma

Mieko Tanaka-Yamawaki and Ryota Itoi

Department of Information and Knowledge Engineering, Graduate School of Engineering,
Tottori University, 101-4 Koyamacho-Minami, Tottori, 680-8552 Japan
mieko@ike.tottori-u.ac.jp

Abstract. We deal with a multi-agent model of the iterated prisoners' dilemma with evolvable strategies, originally proposed by Lindgren that allows elongation of genes represented by one-dimensional binary arrays, by means of three kinds of mutations: the duplication, the fission, and the point mutation, and the strong strategies are set to survive according to their performance at every generation change. The actions that the players can choose are assumed to be either cooperation (represented by C) or defection (represented by D). We conveniently use {0,1} instead of {D,C}. Each player has a strategy that determines the player's action based on the history of actions chosen by both players. Corresponding to the history of actions, represented by a binary tree of depth m, a strategy is represented by the leaves of that tree, an one-dimensional array of length 2^m. We have performed extentive simulations until many long genes are generated by mutations, and by evaluating those genes we have discovered that the genes of high scores are constructed by 3 common quartet elements, [1001], [0001], and [0101]. Furthermore, we have found that the strong genes commonly have the element [1001 0001 0001 0001] that have the following four features:

(1) never defects under the cooperative situation, represented by having '1' in the fourth element of the quartet such as [***1],
(2) retaliates immediately if defected, represented by having '0' in the first element and the third element in the quartet such as [0*0*],
(3) volunteers a cooperative action after repeated defections, represented by '1' in the first element of the genes,
(4) exploits the benefit whenever possible, represented by having '0' in the quartet such as [*0**].

This result is stronger and more specific compared to [1**1 0*** 0*** *001] reported in the work of Lindgren as the structure of strong genes.

1 Introduction

In designing a system, we often ignore the necessity for individual based on the idea that a specific necessity for individual may not apply to the others. However, a design for everybody sometimes satisfies nobody's need. Given a sufficient speed and capacity of today's computers, we are now in the position to put the necessity for individual into a computer.

S. Yamamoto (Ed.): HIMI/HCII 2013, Part III, LNCS 8018, pp. 616–624, 2013.

Based on this thought, we have been studying the game theory simulations and the prediction of the price fluctuation using multi agent models in which the individual setting is allowed for each agent. We have discovered the fact that an evolutional program to simulate a game theory, in order to create a set of better strategies to win the game by examining the past rewards acquired by the players corresponding to the history of actions by both players, can be immediately converted into a program for predicting the next price by changing a small number of commands. For the sake of short term prediction, those elements must be considered independent of the prices. However, it is extremely difficult to incorporate into the program the elements other than the prices, such as human expectations and social conditions. Those elements are to be digested into the market prices after a long time, but it takes a while before they become reflected in the market prices.

In this paper, we consider a model of two-player-game in which strategies of the two players evolve by learning the performance in the past. We adopt a model of iterated prisoners' dilemma with evolving strategies originally proposed by Lindgren and perform extensive amount of simulations until a novel strategy stronger than TFT or Pavlov, by considering the past actions of the both players to the depth 5. This particular strategy is characterized by the 4 features such as, (1) cooperative by nature (2) reasonable (3) generous (4) cool

2 Iterated Prisoners‘ Dilemma

The prisoners‘ dilemma is defined by the payoff structure of both players shown in Table 1. We assume the players have only two ations to choose, to cooperate (C, hereafter) or todefect (D, hereafter). There are four parameters R, P, S, T which are set tp satisfy $S < P < R < T$ and $S + T < 2R$. The key point of the situation under which the two players are set in this model is the better choice for individual results in the worst choice of both. For example, if we assume B cooperates, A's rational choice is to defect because $R < T$. However, even if we assume B defects, A's rational choice is still to defect because $S < P$. Thus A is supposed to defect whatever B chooses. The situation is the same for B. Thus both A and B end up with choosing to defect. However, the payoff P is smaller than R. How can they choose the better option of mutualo cooperation ?

Table 1. The payoff table of the prisoners' dilemma($S<P<R<T$ and $S+T<2R$)

(A's payoff, B's payoff)	B's action is C	B's action is D
A's action is C	(R, R)	(S, T)
A's action is D	(T, S)	(P, P)

The poor soluion (P,P) is inevitable for a single game, unless they promise to start with the cooperative actions. When they repeat the game by starting with the cooperative actions, then the best choice for both of them is to continue to cooperate except the last match. Because onece each of them defects, then the opponent will retaliate in the next match. Therefore if they know the time to end the repeated game, they will defect at the last match. For this reason, the iterated prisoners dilemma game (abbreviated as IPD, hereafter) is played without fixing the time to end. In such a game, a particular strategy called Tit-For-Tat (TFT, hereafter) wins over the other strategies. In general, good strategies including the TFT, share the following three features:

(1) to cooperate as long as the opponent cooperates
(2) to retaliate immediately if defected
(3) to offer cooperation after continuous defections.

However, it has been known that the Pavlov strategy (PAV, hereafter) is better than the TFT under a certain condition. The PAV keeps the same action after getting T or R which are the good payoff, and changes the action from the pervious one after getting S or P which are the poor payoff. This strategy is stronger than the TFT in a model allowing errors in actions in which the player chooses an opposite action from the one chosen by the strategy [7].

This situation is depicted in an example shown in Table 2. In this case, both (TFT,TFT) and (PAV, PAV) begin the game from the cooperative relationships at the time t=1. Suppose if an error occurs at t=2 in the second player, then the TFT pair immediately fall into pose if an error occurs at t=2 in the second player, then the TFT pair immediately fall into (C, D) a series of (C, D) and (D, C), while the PSV pair can recover the original cooperative situation of (C, C). Thus the TFT is not always the best under errors.

Table 2. Actions of the TFT/PAV pair when the second player commits an error at t=2

Time t	(TFT,TFT)	(PAV,PAV)
1	(C, C)	(C, C)
2	(C, 'D')	(C, 'D')
3	(D, C)	(D, D)
4	(C, D)	(C, C)
5	(D, C)	(C, C)

3 Evolvable Strategies in the IPD

In the framework of the artificial life (ALIFE), a new scheme of searching for the better strategies was preseted in Ref. [6] in a multi-agent model of evolvable strategies, in which the strategies grow like genes. Here the strategies are represented by one-dimensional binary strings.

The two actions, the cooperation and the defection, {D,C}, are represented by {0,1}. Each player has a strategy that determines the player's action based on the history of actions chosen by both players in each game. Corresponding to the history of actions, represented by a binary tree of depth m, a strategy is represented by the leaves of that tree, an one-dimensional array of length 2^m. It is convenient to set the two edges of the binary tree to have 0 in the left edges and 1 in the right edges.

For example there are four strategies repsresented by [00], [01], [10], [11], for m=1 corresponding to a model to simply count the opponent's previous action as the history. The strategy [00] is called as ALLD because only D is chosen irrelevant to the opponent's past action. Likewise, [11] is called as ALLC. The strategy [01] is the TFT because D is chosen only when the opponet's action of the immediate past is D. Likewise the strategy [10] is called as anti-TFT (abbreviated as ATFT).

If we count the actions of both players as the history, that is the case of m=2 and the corresponding strategy becomes a binary string of length 4. For example, a strategy [1001] chooses C if the past actions of both players are the same, i.e., both C or both D, and chooses D if the past actions of both players were not the same, i.e., when one player's action was C, the other player's action was D. This corresponds to the PAV. A strategy [0101] is the same as [01] because D is chosen for the opponent's defective action and C is chosen for the opponent's cooperative action irrelevent to the past action of the other side. Likewise, the strategy [0000] is the same as [00]. A strategy represented by [0001] chooses C only when the past actions of both sides were C. We call this strategy as the retaliation-oriented TFT (abbreviated by RTFT).

For larger m, the history and the corresponding strategy can be written as $h_m = (a_m, ..., a_2, a_1)_2$ and $S_m = [A_1 A_2 ... A_n]$ for $n = 2^m$. An example of the strategy for the case of m=3 represented by a string of 10010001 is shown in Fig. 1. Out of all the possible strategies, good ones are chosen by employing the genetic algorithm. The typical job-flow of this mecahnism is illustrated in Fig.2.

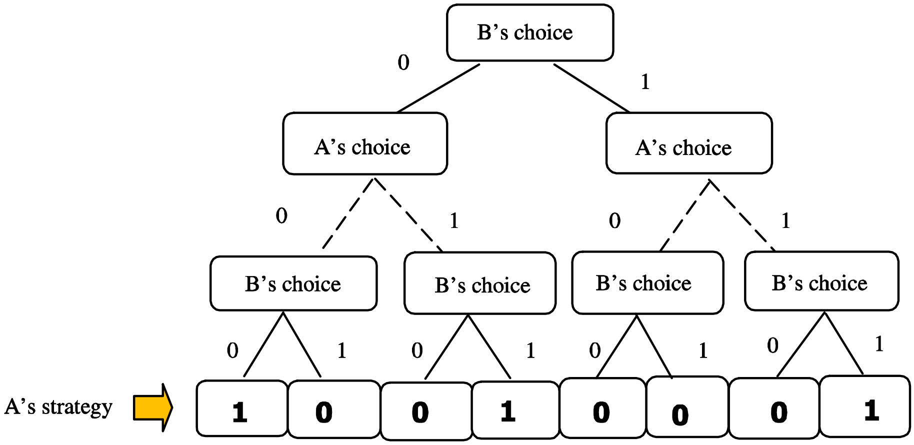

Fig. 1. A strategy of length 8 and a binary tree of the history of depth 3

Starting from the initial population of agents, which could be the entire set of possible strings or a randomly sampled subset of the entire set, pairs of agents play the IPD of indefinite length. After all the agents playing the game with all the other agents, their total payoff are counted and their population is renewred according to the

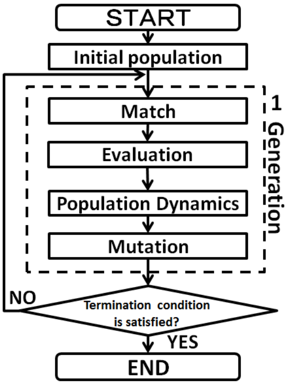

Fig. 2. The Job-flow of the evolvable IPD simulation

population dynamics explained below. Subseqeuently the mechanism of three types of mutation, (1) point mutation (2) doubling, and (3) fission, are applied in order to grow the strategy strings to create new patterns and the new lengths that the previous generation did not know.

We have followed the senario written by Lindgren [6], except for the two points: the first point is the stochastic ending of IPD, and the second point is that we have performed extensive amount of simulations. As a result, we have discovered the type of gene structure of sustainable strategies in more specific manner compared to [1**1 0*** 0*** *001] suggested in Ref.[6]. The reason that we have chosen the stochastic ending is as follows. It is well known that the defective action is the optimum choice for a single-time PD. If the players know the ending time, they are bound to choose to defect (if they are rational), because the situation at the last match is exactly the same as the single match. If the players know their choice to defect at the n-th match, they do not have to consider the effect of their current choice on the later games. In other words, IPD of the length n is equivalent to the IPD of the length n-1, if their choices of the last game are fixed from the beginning and they cannot avoid taking the defective choice at the (n-1)-th match. Thus the players are bound to take the defective choice throughout the IPD, if they know the time to end the iteration of the games. In the IPD game with stochastic ending, on ther other hand, the players do not know the time of ending and they have to consider which action to choose each time.

4 Simulation Result

We have run our program by the following conditions. We have tried two different initial conditions to start the simulation. The first type consists of the four m=1 strategies , [00], [01], [10], and [11] with equal polulations of 250 each, and the second type consists of 1000 random sequences of length 32. Either case, the total population of agents is kept unchanged from the initial value of 1000 throughout the simulation. The number of simulations are 50 for the first type and 40 for the second type. The rate of point mutation, the duplication rate, and the split rate are are set to be 2×10^{-5}, 10-6, 10-6, the same as in Ref. [6]. We also assume the rate of error, i.e., with which the opposite action prescribed by the gene is executed, to be 0.01. The payoff parameters in Table 1 are also chosen to be S=0, P=1, R=3, and T=5.

The length of each game is not fixed in order to avoid the convergence to the ALLD dominance, but the end of the game is announced with the probability of 0.005.

We show simulation results of the Type I initial populations in Fig. 3 in which the horizontal axis shows the generation and the vertical axis shows the population of strategies. Both cases exhibit drastic changes of dominating stratesies as the generation increases.

An interesting feature is observed in Fig. 3. Namely, the [01] (=TFT) dominance followed by the [1001] (=PAV) dominance, then the [0001] (RTFT) dominance comes and the [01] dominance. This particular pattern is observed in 37 examples out of 50 independent runs of the first type initial condition, and this triplet pattern of TFT=>PAV=>RTFT is sometimes repeated for many generations. However, as the length of the genes reaches the size of 16 or 32, this triplet pattern disappear and the [1001000100010001] dominates.

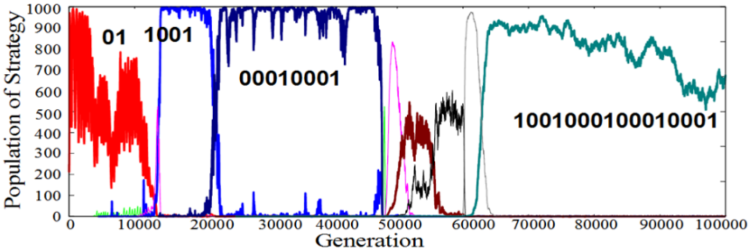

Fig. 3. The TFT-PAV-RTFT triplet is observed in the Type I condition

Fig.4 is an example of the triplet pattern of TFT-PAV-RTFT repeated for three cycles. Fig.5 shows a case of the triplet pattern washed away by the emergence of the longer and the stronger strategies.

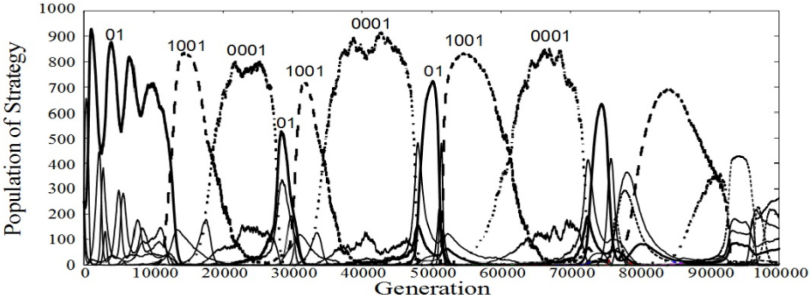

Fig. 4. An example of the TFT-PAV-RTFT triplet repeated by 3 times

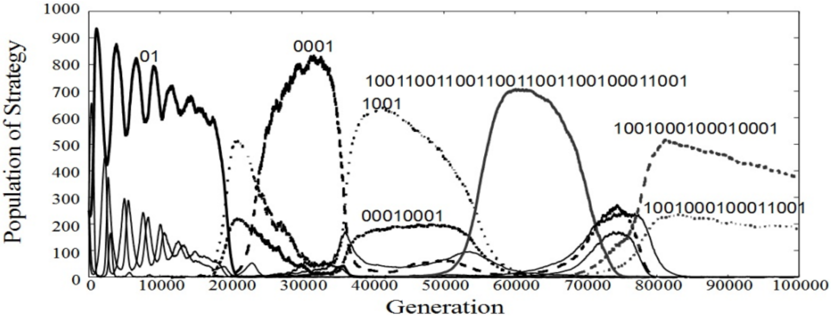

Fig. 5. A collapse of the triplet by the emergence of longer strategies

5 Evaluation of the Sategies

We try to quantify the degree of sustainability of those strategies by means of a fitness parameter W_i defined by the accumulated sum of population throughout the total generation. The 8153 strategies emerged in the 45 simulations of Type I initial condition and the 11753 strategies emerged in the 50 simulations of Type II initial condition are sorted in the descending order of W_i in Table 3. The strategies having positive values of fitness are chosen as 'good' strategies and selected for further analysis. The number of 'good' strategies, satisfying the of the positive fitness condition, was 340 out of 8153 for the case of Type I initial condition, and 785 out of 11753 for the case of Type II initial condition.

Table 3. Evaluation of the strategies

Type I initial strategies(fixed)		**Type II initial strategies(random)**	
Strategy	W_i	**Strategy**	W_i
1101 1001	0.123	1011	0.078
0101 1001	0.077	0000 0011	0.070
1101 0110	0.064	1101 1010	0.059
1010 0011	0.050	1001 1001	0.049
1101 0100	0.047	1101 1011 1101 1011	0.045
1001 0001 0001 0001	0.041	0001 0011	0.038
0001 1011	0.040	1101 0101 0001 1001	0.036
0100 1001	0.039	1101 1101 0000 0111	0.032
1101 0111	0.029	1000 0000 0100 0001	0.029
1001 1011 1001 1011	0.028	1111 0101 0101 1110	0.027

We search for a possible characteristic feature of those strategies selected by using the goodness criterion. We first set the length of all those strategies to the equal length (=32), by doubling and count the frequency of symbol '1' at each site, as illustrated in Fig. 8. The rates of '1' for all the 32 sites are shown in Fig. 9 for the Type I initial population and in Fig. 10 for the Type II initial population. This structure can be assumed to be the prototype strategy. The result shows that both Type I and Type II derived the same structure of [1001 0001 0001 0001].

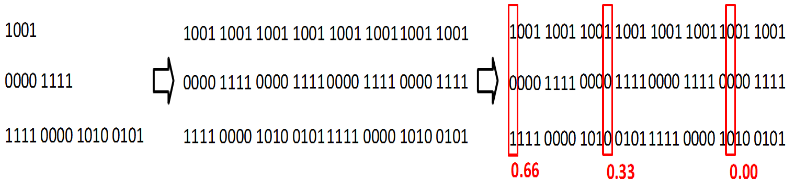

Fig. 6. Compute the rate of occurrence of '1' at each site

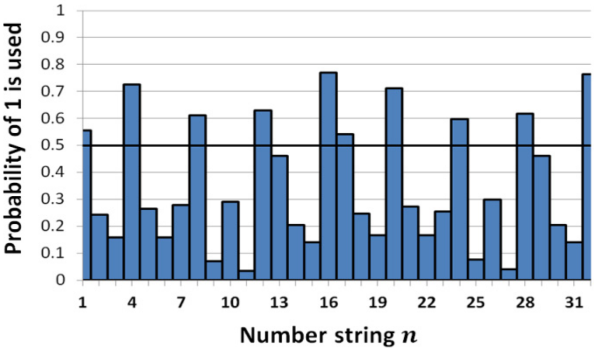

Fig. 7. The rates of '1' at each site of total 32 sites for Type I initial population

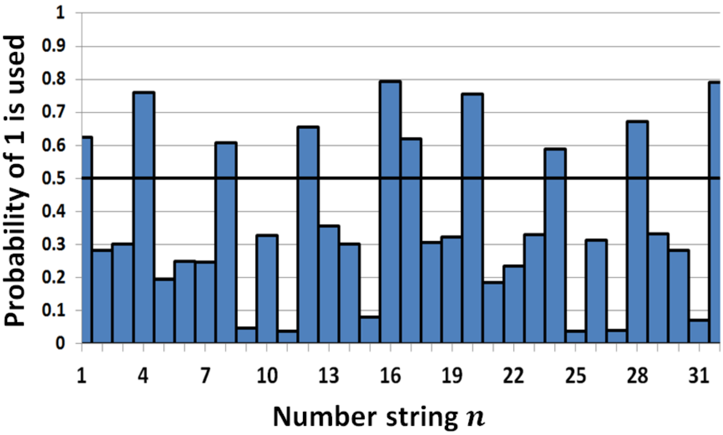

Fig. 8. The rates of '1' at each site of total 32 sites for Type II initial population

6 Discussion

Based on the result of our simulations, 'good' strategies who survive longer with larger population compared to the others have a common prototype gene structure of [1001 0001 0001 0001]. Moreover, this result was irrelevant to the initial population. This gene structure is characterized by the following 4 features.

1. cooperate if the opponent cooperates (This feature is seen in common to [***1 ***1 ***1 ***1], TFT, PAV, and [1**1 0*** 0*** *001], etc.)
2. immediately retaliate if defected (This feature is seen in common to [0*0* 0*0* 0*0* 0*0* 0*0*], TFT, PAV, but not in [1**1 0*** 0*** *001].)
3. generous
   This feature is seen in common to [1*** **** **** ****], PAV, and [1**1 0*** 0*** *001], but not in TFT. Also, the structure [1*** **** **** ****] has an advantage over PAV for being more robust against ALL-D due to longer term of patience.
4. coolness
   This feature is in common to [*0** *0** *0** *0**] having 0 against the opponent's cooperative action. TFT, [1**1 0*** 0*** *001] do not have such a feature.

7 Conclusion

We have performed extensive simulations of IPD and analyzed to determine the prototype structure of 'good' genes having a structure of [1001 0001 0001 0001]. Although this is a specific example of the structure of strong gene, [1**1 0*** 0*** *001], suggested in Ref. [6], our analysis have reached much stronger specification of the gene structure of the strategy 'better' than TFT. This prototype consists of two types of quartets corresponding to PAV and RTFT. In other words, this strategy acts like the Pavlov when the actions of both players were 'Defect' at the game before the last, but acts like RTFT for the other three cases. This strategy has stronger tendency of retaliation against the opponent's defection compared to the Pavlov strategy. The advantage of this strategy compared to TFT is based on the structure of starting with '1', which helps to offer cooperation under defective situations, which is considered to be a key to solve the dilemma structure of many social problems.

References

1. Novak, M.A., Sigmund, K.: Evolution of indirect reciprocityby image scoring. Nature 393, 573–576 (1998)
2. Roberts, G., Sherratt, T.N.: Development of cooperative rela-tionship through increasing investment. Nature 394, 175–178 (1998)
3. Yao, X., Darwen, P.: How important is your requtation in amultiagent environment. In: IEEE-SMC 1999, pp. 575–580 (1999)
4. Axelrod, R.: The Evolution of Cooperation (1984)
5. Tanaka-Yamawaki, M., Murakami, T.: Effect of reputation on the formation of cooperative network of prisoners. In: Nakamatsu, K., Phillips-Wren, G., Jain, L.C., Howlett, R.J. (eds.) New Advances in Intelligent Decision Technologies. SCI, vol. 199, pp. 615–623. Springer, Heidelberg (2009)
6. Lindgren, K.: Evolutionary Phenomena in Simple Dynamics; Articial Life II, pp. 295–312. Addison-Wesley (1990)
7. Nowak, M.A., Sigmund, K.: A strategy of win-stay, lose-shift that outperforms tit-for-tat in Prisoner's Dilemma. Nature 364, 56–58 (1993)