Occurrence of state confusion in the learning of communication using Q-learning

Masanobu Nakanishi, Masanori Sugisaka, Katsunari Shibata
Dept. of Electrical and Electronic Engineering, Oita University
700 Dannoharu Oita 870-1192 Japan

Abstract

The learning of one-way communication between two agents using Q-learning was investigated. A transmitter agent learned what communication signals should be transmitted and a receiver agent learned to generate appropriate actions from the signals. We discovered that when there exists a non-looped branch in the receiver's state transition, and the optimal action in a detour is the same as the optimal one in a state closer to the goal on the optimal path, there is a possibility that the receiver cannot take the optimal path because of state confusion. The main reason why the receiver agent falls into the state confusion can be considered that it is not reinforced for the transmitter agent to learn to transmit the state value to the receiver agent.

1. Introduction

Communication has a very important role in collision avoidance, cooperative action and the supplement to insufficient observation in multi-robot and multi-agent systems. In order to learn a purposive communication autonomously, evolutionary method[1] or reinforcement learning[2][3][4] has been used. A kind of simulation that shows autonomous acquisition of one-way communication to supply the receiver's insufficient observation was employed. However, it was not examined what kind of information should be transmitted, or whether the optimal communication can be acquired in any cases.

We have investigated what communication signals a transmitter agent should learn to transmit and whether a receiver agent can learn to generate appropriate actions from the signals. We discovered the case in which a receiver agent falls into POMDP(Partially Observable Markov Decision Process) and its action does not become optimal one because of state confusion that is caused by the communication signal obtained by reinforcement learning. In this paper, some simulation results are introduced, and it is considered empirically why and on which condition the state confusion happens and blocks to learn the optimal path.

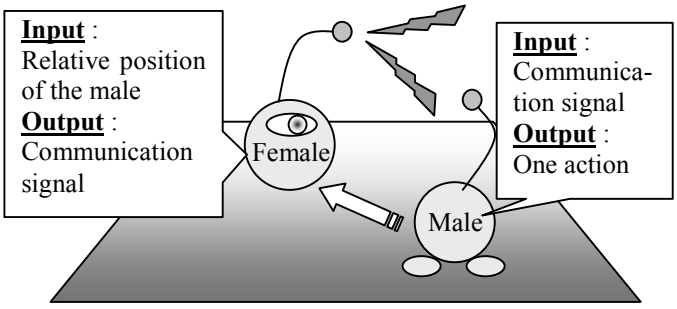

Fig.1 Learning of one-way communication

2. Learning of one-way communication

2-1 Task description

In this paper, the learning of one-way communication that supplies the receiver's insufficient observation is focused on. The simulation environment was decided referring to [1][2]. Fig.1 shows the image of one-way communication learning. Two agents called a male and a female are assumed in a discrete environment. The male can move, but does not have sight. On the other hand, the female can't move, but can transmit some signals to the male. The female's input is the relative position of the male, and it's output is a communication signal. The male's input is the communication signal and its output is an action. If the male touches the female, a reward is given to the both agents. The meaning of the communication signal is not given to either agent at all beforehand. Therefore, a transmitter agent has to learn what communication signal should be transmitted and a receiver agent has to learn to generate appropriate actions from the signal. If some common language can be built up between the male and the female, the contact can be repeated efficiently.

2-2 Learning method for the both agents

For the learning of the both agents, Q-learning is used. In Q-learning, the state-action pair is evaluated, and the evaluation value is called Q-value. An agent chooses an action with the probability calculated from the Q-values. It is performed on a discrete environment and action space.

The algorithm of Q learning is as follows.

(1) An agent observes a state.
(2) The agent selects and executes an action.
(3) The agent observes the state after the transition.
(4) A reward r_{t+1} is received from the environment.
(5) Q-value is modified as

$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha\left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a)\right] \quad (1)$$

where α is a learning rate $(0 < \alpha \leq 1)$
γ is a discount factor $(0 \leq \gamma < 1)$
(6) t -> $t+1$, and it is returned to the step(2).

For the female, the state s is the male's relative position, and the action a is the communication signal. For the male, the state s is the communication signal, the action a results in a state transition, and a is 0.1, γ is 0.9 and all the initial Q-values are 0 here.

2-3 Action selection

An action is selected using Boltzmann selection here. When the state is x, the probability of the action a is calculated as,

$$p(a \mid x) = \frac{\exp(Q(x,a)/T)}{\sum_{i \in A} \exp(Q(x,i)/T)} \quad (2)$$

where A is a set of actions, and T is a temperature coefficient. An action is selected randomly when T is large. As opposite to it, when T is close to 0, a little difference of Q-value has great influence on the action selection, in other words, the action selection is almost greedy. In this simulations, the initial value of T is 1.0, and it is gradually decreased exponentially to 0.01 in 80% of trials. In the rest of trials, it was fixed at 0.01.

2-4 Flow of the learning

These agents act in accordance with the following cycle.

(1) The female detects the male's state.
(2) The female's Q-value at t-1 is modified.
(3) The female transmits a signal to the male.
(4) The male receives the female's signal.
(5) The male's Q-value at t-1 is modified.
(6) The male makes an action.
(7) If the both agents touch each other, the trial finishes, and they get a reward. In that case they learns their Q-value at t with $\max Q(s_{t+1}, a) = 0$ according to Eq.(1) and the flow returns to (1). If the trial finished, t=0, otherwise t -> t+1.

The step(2)(5) is not executed when t=0.

3. Simulation

3-1 Case1

At first, an example, in which the agent could not learned the optimal path, that we found, is introduced.

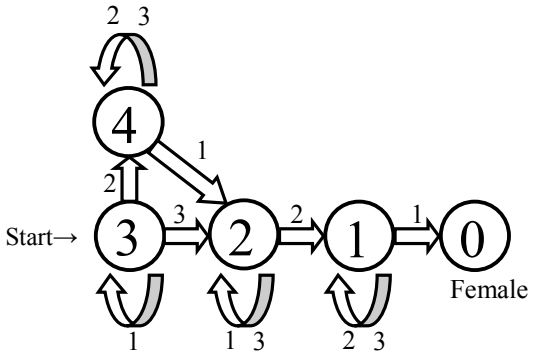

Fig.2 Simulation environment 1

The diagram of the male's state transition is shown in Fig.2. A number in the circle indicates the state, and an arrow indicates a state transition that is always deterministic. The male can take one of three actions(1,2,3), and the number along the arrow indicates the taken action. The female's input is the male's present state(one of from 0 to 4), and the female's signal has four kinds(from 1 to 4). On the other hand, the male's input is the communication signal from the female. Therefore, the number of the signals is more by one than that of the actions. The male's initial state is 3, and the state 0 is the goal where the female exists. If the male reaches the state 0, a reward is given to the both agents, and then the male is located on the initial state 3 again. A trial is defined as a sequence from the initial state to the goal.

Q-values of the both agents after 10000 trials of learning are shown in Table 1. The maximum Q-value for each state is hatched in the table. From the table, it can be seen that the female transmits the signal 2 on the state 1 or 4, and it transmits the signal 1, 3 or 4 on the state 2 or 3. On the other hand, the male selects the action 2 on the signal 1, 3 or 4, and it selects the action 1 on the signal 2.

Since, the male selects the action 2 on the state 3, the optimal action could not be learned even though the number of the signals is more than that of the male's actions. Even when the number of the communication signals is increased, the agents still could not learn the optimal path, because one of the communication signal was assigned to the action 1, and the other signals are all

assigned to the action 2 in the male. The process that the agents learned non-optimal action has been considered as shown in Fig.3.

Table 1 Q-value after learning

Female's Q-value [state][signal]	Male's Q-value [signal][action]
f_q[1][1]=0.610	m_q[1][1]=0.677
f_q[1][2]=1.000	m_q[1][2]=0.754
f_q[1][3]=0.727	m_q[1][3]=0.678
f_q[1][4]=0.670	m_q[2][1]=0.849
f_q[2][1]=0.900	m_q[2][2]=0.699
f_q[2][2]=0.794	m_q[2][3]=0.680
f_q[2][3]=0.900	m_q[3][1]=0.678
f_q[2][4]=0.900	m_q[3][2]=0.757
f_q[3][1]=0.729	m_q[3][3]=0.678
f_q[3][2]=0.653	m_q[4][1]=0.677
f_q[3][3]=0.729	m_q[4][2]=0.758
f_q[3][4]=0.729	m_q[4][3]=0.683
f_q[4][1]=0.726	
f_q[4][2]=0.810	
f_q[4][3]=0.729	
f_q[4][4]=0.728	

State 1 When the female transmits a signal A, the male learns to select the action 1 and reaches the goal, then the female also learns to transmit the signal A.

↓

State 4 Female learns to transmit the signal A because the male chooses the action 1 when the female transmit the signal A.

↓

State 4 The value of the state 4 increases for the male because it receives the same signal at the state 1.

↓

State 3 The male learns to choose the action 2 because the value of the action 2 is higher than the action 3.

↓

Agents could not learn the optimal path.

Fig.3 Considered process by which non-optimal path was learned.

From this, the main problem is that the signal which represents the state value was not reinforced. Furthermore, all the signals except for the one signal are assigned to the action 2 in the male. The reason is considered as follows.

(1) State 2 : When the female transmits a signal B which is not A in Fig.3, the male chooses the action 2, and learns to move to the state 1.

(2) State 3 : The female learns to transmit another signal C except for A or B because the female wants the male to take the action 3.

(3) State 3 : The male learns to choose the action 2 even if it receives the signal C because the evaluation value of the state 4 is higher than that of the state 2 due to the state confusion as shown in Fig.3.

(4) State 3 : The Q-value for the signal C goes down, and the female comes to transmit the signal D.

Then, the male learns to allocate all the signals except for the signal A to the action 2 by repeating (2), (3), (4).

3-3 Case 2

Next, an example in which the optimal actions can be learned in a similar environment to that in the case 1 is shown in Fig.4. Some important Q-values are also shown.

Female's Q-value f_q[state][signal]	Male's Q-value m_q[signal][action]
f_q[2][1]=0.900	m_q[1][1]=0.667
f_q[2][2]=0.793	m_q[1][2]=0.743
f_q[2][3]=0.900	m_q[1][3]=0.671
f_q[2][4]=0.900	m_q[3][1]=0.667
f_q[3][1]=0.770	m_q[3][2]=0.743
f_q[3][2]=0.616	m_q[3][3]=0.672
f_q[3][3]=0.742	m_q[4][1]=0.667
f_q[3][4]=0.748	m_q[4][2]=0.743
	m_q[4][3]=0.674

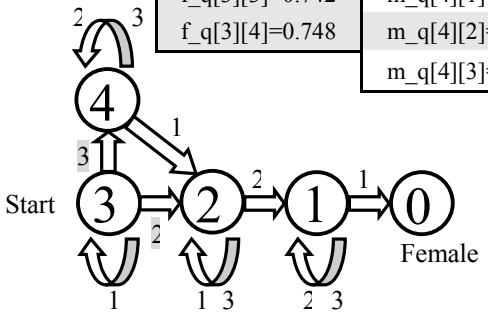

Fig.4 Simulation environment 2

Between the case 1 and case 2, only the state transition from the state 3 is different. In this environment, when the made selects the action 2, it moves to the state 2. When it selects the action 3, it moves to the state 4. The action and state transition pair is inverted from the environment 1 as shown in Fig.2.

Though there is non-looped branch in the receiver's state transition and the optimal action in the state 4 on

the detour is the same as the optimal one in the state 1 that is closer to the goal on the optimal path, the male could learn the optimal path to the goal. The reason why the agent could learn is that the optimal action in the state 3 is the same as the optimal one in the state2 that is closer to the goal on the optimal path.

From the above, it is suggested that the agent sometimes can learn the optimal actions even if there exists a non-looped branch in the receiver's state transition and the optimal action in a detour is the same as the optimal one in a state closer to the goal on the optimal path.

3-2 Case 3

Finally, the case in which non-looped branch does not exist.

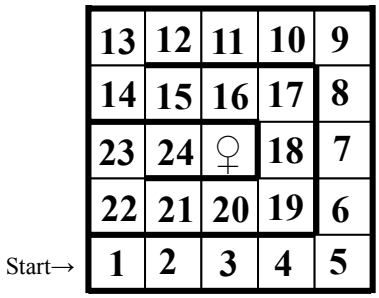

Fig.5 Simulation environment 3

Here, a simulation environment consisting of 25 discrete states as shown in Fig.5 is assumed. A bold line shows a wall. The male starts at the state 1, and female in the center square. The male's action has four kinds, and the male moved to the next cell in each of 4 directions. When the action to go to the wall is selected, the agent does not move.

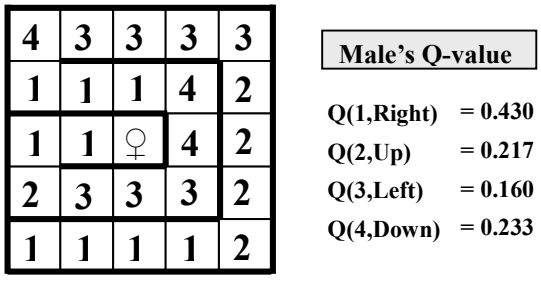

Fig.6 Result of simulation environment 3

The number in each square in Fig.6 is the communication signal assigned by the female after learning. These signals became correspond to the optimal action to the goal, such that the agent should go to the right when the signal is 1. This environment becomes POMDP for the male due to the state confusion.

After learning the Q-values of non-optimal action in each state becomes almost 0, and the optimal path could be learned. The reason might be that if the agent selects a looped branch, it will return to the state again, and that results in the decrease of the Q-value corresponding to the looped branch. Therefore, it can be thought that if there is no branch except for a loop the optimal actions could be learned, even if the state value that is closer to the goal is smaller due to the state confusion.

4. Conclusion

In this paper, it was shown that when there exists a non-looped branch in the receiver's state transition, and the optimal action in a detour is the same as the optimal one in a state closer to the goal on the optimal path, there is a possibility that the receiver cannot take the optimal path because of the male's state confusion. Then, the main factor for this problem might be that the transmitter agent's signal is not reinforced to represent the state value. It was also suggested that there is a possibility that the optimal action could be learned even if state confusion happened on the optimal path, if there is no branch except for a loop even if it happens that the evaluation value of a state that is closer to the goal is smaller due to the state confusion.

The method by which the transmitter learns to send a signal representing the state value has to be developed in the future research.

Acknowledgement

This research was partially supported by the Japan Society for the Promotion of Science, Grants-in-Aid for Scientific Reearch, #14350227 and #15300064.

Bibliography

[1] G.M. Werner & M. G. Dyer : Evolution of Communication in Artificial Organizing System, Proc.of Artificial life II, 1-47(1991)

[2] N.Ono, T. Ohira, and A. T. Rahmani: Emergent Organization of Interspecies Communication in Q-Learning Artificial Organism,*Advances in Artificial Life*, pp. 396-405 (1995)

[3] K. Shibata and K. Ito: Learning of Communication for Negotiation to Avoid Some Conflicts of Interests-Learning of Dynamic Communication Using a Recurrent Neural Network-Trans. of SICE Vol.35, No.11, 1346-1354 (1999) (in Japanese)

[4] K. Shibata and K. Ito Emergence of Communication for Negotiation by a Recurrent Neural Network, Proc. of ISADS '99, pp, 294-301. (1999)