Construction of a Learning Agent Handling Its Rewards According to Environmental Situations

Koichi Moriyama
koichi@nm.cs.titech.ac.jp

Masayuki Numao
numao@cs.titech.ac.jp

Department of Computer Science, Graduate School of Information Science and Engineering
Tokyo Institute of Technology
2-12-1, Ookayama, Meguro, Tokyo, 152-8552, Japan

ABSTRACT

The authors aim at constructing an agent that learns appropriate actions in a Multi-Agent environment with and without social dilemmas. The agent ought to voluntarily give up its profit in a dilemma situation and it should keep its profit in another situation. We divide the environment into three situations and introduce reward-handling manners for learning actions, which are effective in each situation. Since the agent must select an effective manner for the situation, the authors contrive criteria for recognizing the situation. This paper shows that the agent having the manners and the criteria acts well in two of the three Multi-Agent situations composed of homogeneous agents.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*intelligent agents, multiagent systems*; I.2.3 [**Artificial Intelligence**]: Learning

General Terms

Algorithms, Experimentation, Performance

Keywords

Reinforcement Learning, Social Dilemma.

1. INTRODUCTION

We deal with problems about social dilemmas in a Multi-Agent environment. Social dilemmas are troubles caused by the friction between individuals and their society. In social dilemma situations, individuals should act nonrationally and in fact we humans can act independently like that.

Therefore, we now construct a nonrational agent which gets over dilemmas independently. Since there is much work on rational learning methods, we give the agent a rational learning method *with handled rewards*.

However, if we give the agent a fixed reward-handling manner, it does not act well in a Multi-Agent environment with and without social dilemmas. Thus, the authors equip the agent with criteria for recognizing the situation to change reward-handling manners *by itself*.

In this paper, we divide a Multi-Agent environment composed of homogeneous agents into three situations, and observe that the agent can get over a dilemma scenario, while it can take efficient actions in a non-dilemma one.

2. THREE MULTI-AGENT SITUATIONS

We divide a Multi-Agent environment into three situations. Suppose that there are n agents in a society each of which takes two actions: *greedy* (G) and *disinterested* (D). A combination of their actions is expressed as a vector $\boldsymbol{a} \equiv (a_1, a_2, ..., a_n)$ $a_i \in \{\mathrm{G}, \mathrm{D}\}$, and r_i is the i-th agent's reward. Then, the first situation, *isolating*, is defined as follows:

$$\arg\max_{\boldsymbol{a}} \sum_{i=1}^{n} r_i = (\mathrm{G}, \mathrm{G}, ..., \mathrm{G}),$$

and the second situation, *morassy*, is defined as follows:

$$\arg\max_{\boldsymbol{a}} \sum_{i=1}^{n} r_i = (\mathrm{D}, \mathrm{D}, ..., \mathrm{D}).$$

The third situation, *competing*, gains the maximum summed reward when there are at least one G-actor and at least one D-actor.

The authors aim at constructing an agent which acts well in all the three situations.

3. AGENT HANDLING ITS REWARDS

In this paper, we use Q-learning [3] which is a representative method of reinforcement learning. Since it is a rational learning method for maximizing rewards, we must consider how we use Q-learning for learning proper actions in a dilemma situation, i.e. in a morassy and a competing one.

The authors propose that an agent A_i learns by Q-learning with a *handled reward* $r'_{i,t+1}$ at time t, which is the sum of a reward $r_{i,t+1}$ and a parameter $\lambda_{i,t+1}$. We omit the subscript i showing "the agent itself (A_i)" in the following.

$$r'_{t+1} = r_{t+1} + \lambda_{t+1} \tag{1}$$

First we introduce

$$\lambda_{t+1} = \sum_{A_k \in N_i \setminus A_i} r_{k,t+1} \tag{2}$$

AAMAS'02, July 15-19, 2002, Bologna, Italy.

as an effective λ_{t+1} for a morassy situation. $N_i \backslash A_i$ is a set composed of A_i's neighbors N_i without A_i itself. It is effective in a situation in which the neighbors suffer when A_i takes a self-interested action. Second we introduce

$$\lambda_{t+1} = r_{t+1} - r_t \tag{3}$$

as an effective λ_{t+1} for a competing situation. This emphasizes the temporal difference of rewards and is also effective in an isolating situation.

These λ_{t+1}'s are ineffective in unfit situations. Thus, we introduce two perception criteria to select these λ_{t+1}'s correctly as follows. The agent regards the current situation as a morassy when at least one of these criteria is hold.

$$Q(s_t, a) < 0 \quad \text{for all } a \in \mathcal{A}_t. \tag{4}$$

$$r_{t+1} < Q(s_t, a_t) - \gamma \max_{a \in \mathcal{A}_{t+1}} Q(s_{t+1}, a). \tag{5}$$

$\mathcal{A}_t$ is a set of actions the agent may take in the state s_t. Formula 4 means that all actions in $\mathcal{A}_t$ cause punishments in the future. In Formula 5, since both sides are equal when Q-learning is converged, the right side is an estimate of r_{t+1} learned so far. In that, this inequality means that a real reward is less than an estimate. Note that the left side of Formula 5 should be r'_{t+1}, because Q-learning in this paper uses a handled reward r'_{t+1}. However, since the formula decides which λ_{t+1} is used in Formula 1, we cannot use r'_{t+1} here and we use r_{t+1} instead.

In this paper, the authors propose that the agent handles its own reward by using Formula 2 if it recognizes the situation as a morassy, otherwise by using Formula 3.

4. EXPERIMENTS

We use a game that models the tragedy of the commons [1] as a morassy situation and as an isolating situation. A competing situation is omitted in this paper due to limitations of space. Each agent takes three actions: *selfish*, *cooperative*, and *altruistic*. Each action $a_{i,t}$ of an agent A_i brings a base reward $r(a_{i,t})$ and a cost for the society $c(a_{i,t})$. A base reward $r(a_{i,t})$ is 3, 1, and -3 followed by a selfish, a cooperative, and an altruistic action, respectively. Each agent obtains a reward $r_{i,t+1} \triangleq r(a_{i,t}) - \sum_j c(a_{j,t})$ after all the agents take actions.

A Multi-Agent environment is comprised of ten homogeneous agents. There are five types of agent for comparison, i.e. taking random actions (Random), conducting normal Q-learning (Normal), always using Formula 2 (SumReward), always using Formula 3 (TempDiff), and having the proposed method (AutoSelect).

A set of neighbors N_i appearing in Formula 2 is defined as $N_i \triangleq \{A_k \mid k = (i+j) \bmod 10, j = 0, 1, 2, 3\}$. A_i's states are defined by the combination of actions of agents in N_i. Q-learning parameters, viz. the learning rate α and the discount factor γ, are both set to 0.5. Action selection is the Boltzmann selection with temperature $T = 1$.

4.1 Morassy Situation

In a morassy situation, we set $c(a_{i,t})$ to 1, 0, and -1 followed by a selfish, a cooperative, and an altruistic action, respectively. Thus, a summed reward becomes maximum when all agents take *altruistic* actions. However, a *selfish* action always produces the largest reward from an agent's point of view. The result is shown in Figure 1.

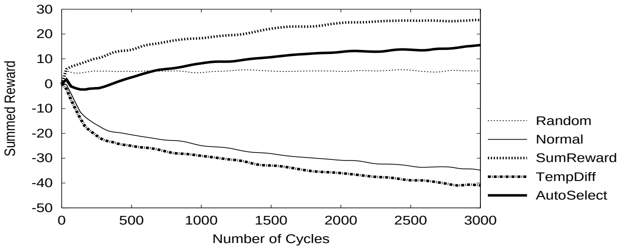

Figure 1: The result of a morassy situation

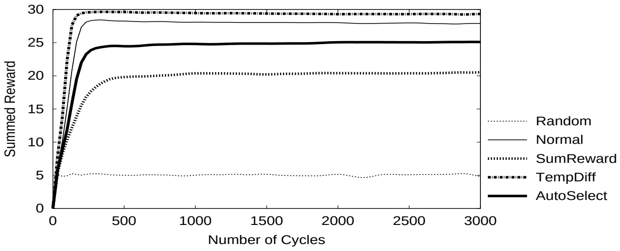

Figure 2: The result of an isolating situation

4.2 Isolating Situation

In an isolating situation, we set $c(a_{i,t}) = 0$ for all $a_{i,t}$'s. Thus, a summed reward becomes maximum when all agents take selfish actions. The result is shown in Figure 2.

5. DISCUSSION

In the experiments, SumReward and TempDiff agents tend to take altruistic and selfish actions, respectively. AutoAdapt agents generally tend to take altruistic actions, but sometimes they take selfish actions. Therefore, there is no type of agents which obtains more summed rewards than AutoAdapt agents' in both situations.

This research is composed of two parts: devising a learning method effective for a specific situation and constructing a learning method effective for several situations by combining those specific methods. The latter means how meta-rules are designed and it is a point of this research.

6. CONCLUSION

We first divided a Multi-Agent environment into three situations: isolating, morassy, and competing. Then we constructed an agent which learns proper actions in the three situations by the handled-reward Q-learning with two perception criteria. This paper shows that the agent acts well in morassy and isolating situations composed of homogeneous agents.

7. REFERENCES

[1] G. Hardin. The Tragedy of the Commons. *Science*, 162:1243–1248, 1968.

[2] K. Moriyama and M. Numao. Constructing an Autonomous Agent with an Interdependent Heuristics. In *Proc. PRICAI-2000* (LNAI 1886), 329–339, 2000.

[3] C. J. C. H. Watkins and P. Dayan. Technical Note: Q-learning. *Machine Learning*, 8:279–292, 1992.