The Fifteenth International Symposium on Artificial Life and Robotics 2010 (AROB 15th '10),
B-Con Plaza, Beppu,Oita, Japan, February 4-6, 2010

A Study on Q-learning Considering Negative Rewards

Takayasu FUCHIDA Kathy Thi Aung
Atsushi SAKURAGI
Graduate School of Science and Engineering, Kagoshima University
1-21-40 Kohrimoto Kagoshima-city, 890-0065 Japan
Email: fuchda@ibe.kagoshima-u.ac.jp

Abstract: In the reinforcement learning system, the agent obtains positive reward such like 1 when it achieves own goal. Positive rewards are propagated around the goal area and the agent gradually becomes to success to reach his goal. If you want to avoid some situations such like dangerous places or poison items, you might want to give negative reward to the agent. But conventional Q-learning, negative rewards are not propagated more than one state. In this paper, we propose a new way to propagate negative rewards. This is very simple and efficient technique for Q-learning. At last, we show the results of computer simulations and the effectiveness of proposed method.

Keywords: reinforcement learning, Q-learning, negative rewards

I INTRODUCTION

Reinforcement learning (RL) methods are powerful and hopeful way to control the agents like an autonomous robot [1], [2]. In the RL, a Q-learning is the very popular way to construct the intelligence about the given environments [3].

Normally, a state which is the goal for the agent gives the reward 1 to the agent, and other states will give the reward 0 to the agent. When we want to show that a state which is bad for the agent, we would want to give the reward -1 to the agent. But in the normal Q-learning, only the maximum Q-value in the next state must be selected even if a negative Q-value would exist. So if a state has a negative Q-value which absolute value is maximum value, the positive Q-value will be selected and negative Q-value is not propagated. When we could use the negative reward in order to indicate the bad state for the agent, the agent would possibly avoid the states such like this.

Reinforcement learning using reward and punishment was proposed by OKADA in 2000 [4] but in this paper experimental results were not shown.

In this paper, we propose a new learning method in which the negative Q-value can be propagated, and show the effectiveness of this method. The reason that the negative Q-value is not propagated is that the learning equation of normal Q-learning method uses the maximum value of the next state. So we changed the term of the equation from selecting the maximum value of next state to using the absolute value of Q-values.

As the results of computer simulations, we illustrate that the propagation of the negative Q-values forces the agent to avoid the bad states and go to the goal effectively.

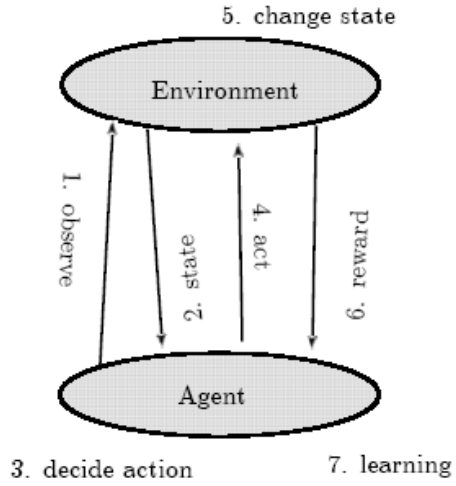


Figure 1: Interactions of RL

II REINFORCEMENT LEARNING

Figure 1 illustrates the interactions of the reinforcement learning. In the reinforcement learning, first, the agent observes the environment and obtains the state. The agent decides the next action using the state and do it to the environment. The environment is influenced by that action and changes their own state. After changing the state, the environment returns the reward to the agent. Finally, the agent learns with received reward.

During the learning period, this process is repeated and the agent accumulates the knowledge about its circumstances as the value of Q.

1 Q-learning

In the Q-learning, the worth of selecting an action in the specified state is quite important. The Q-value is the expected value of returns which is the discounted sum of the rewards the agent received, and is used as the worth of action and state pair.

On the general Q-learning, every action and state pair have own Q-value. These values are initialized to small random numbers and gradually change toward the optimal values through the learning.

At the time t, the agent observe the state s_t and executes the action a_t. As the result, the agent obtained the reward r_t, and finally the state of environment turned into s_{t+1}, then the Q-value $Q(s_t, a_t)$ is updated as following equation (1).

$$\begin{aligned} Q(s_t, a_t) \quad \leftarrow \quad & Q(s_t, a_t) \\ & +\alpha \left(r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right) \end{aligned} \tag{1}$$

In this equation, α is the learning factor and γ is the discount factor.

2 Deciding action

The agent selects next action which has highest Q-value. The action which has large Q-value is considered as the good way to achieve the goal. But selecting highest Q-value continually decreases the opportunities to find better way. Therefore the agent sometimes selects next action randomly. This random selection is useful for exploring the state space and finding the new better way which has not been found yet.

3 Propagation of worth

In the equation (1), the term $\max_a Q(s_{t+1}, a)$ selects the highest value of next state s_{t+1}. This term has an role to spread the worth of Q-value.

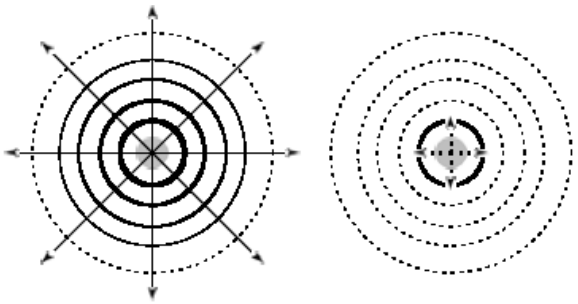
Generally, in the reinforcement learning, the agent acts many times before it reaches to the goal. Excepting the goal state, other states usually give the reward $r_t = 0$ to the agent. So the update equation becomes as (2).

$$\begin{aligned} Q(s_t, a_t) \quad \leftarrow \quad & Q(s_t, a_t) \\ & +\alpha \left(\gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right) \\ = \quad & (1-\alpha) Q(s_t, a_t) + \alpha\gamma \max_a Q(s_{t+1}, a) \end{aligned} \tag{2}$$

This means that Q-value gradually decreases and is added the discounted maximum value of next state.

For example, if the state s_1 had large Q-value and the state s_2 was the previous state that could move to s_1 with one action. The agent staying the state s_2, without random action, must selects an action which leads it to the state s_1 and This means that the worth of s_1 is propagated to s_2.

In this way, the worth of goal area is propagated step by step into the whole state space and the agent will become to be able to reach to goal efficiently.



(a) positive worth (b) negative worth

Figure 2: Propagation of worth on conventional Q-learning

III PROPOSED METHOD

1 Problem in conventional way

If the objective of the agent would be only reaching the goal, it was enough just to put the positive reward in the state space. But if you want the agent to avoid the obstacles and to reach the goal, you might be want to put some negative rewards in the state space. The negative reward represents the bad result and is to be avoided by the agent. Because of the agent acts in order to increase the sum of rewards, the states which have negative rewards are not suitable for the agent.

But in the conventional Q-learning, the positive worth can be propagated around the goal area but the negative worth can not be propagated. This is because you select the highest Q-value of next state s_{t+1} in the equation (1).

Figure 2 illustrates the situations of spreading positive and negative worth. The center points of these tow figures are the states giving their own rewards. The positive worth will be propagated around the reward state like fig.2-(a), but the negative worth will not be propagated more than one movement length like fig.2-(b).

2 Propagation of negative worth

In the case that some negative reward states exist, it is better that the negative worth is also propagated. If the negative worth would not be propagated, the agent could not notice that the dangerous state was approaching to him until the danger was close to just next to him. If the negative worth would be propagated, the agent might become to be able to avoid the dangerous zone in early time.

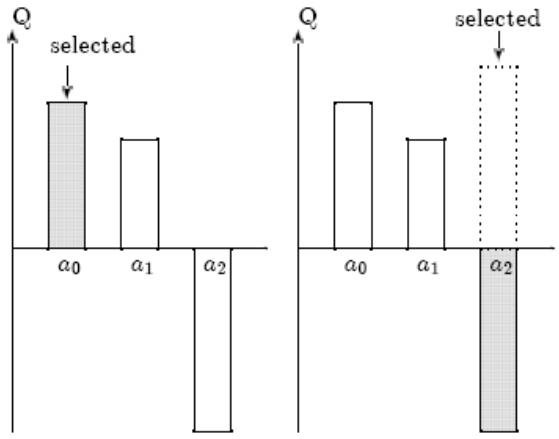


Figure 3: The difference of eqn.(1) and (3)

In this paper, we propose a new learning method to be able to propagate the negative worth. The learning equation is shown in (3).

$$\begin{cases} p = \arg\max_a |Q(s_{t+1}, a)| \\ Q(s_t, a_t) \leftarrow Q(s_t, a_t) \\ \qquad +\alpha\left(r_t + \gamma Q(s_{t+1}, p) - Q(s_t, a_t)\right) \end{cases} \tag{3}$$

In this equation, p is the argument of action index which has the largest absolute Q-value in next state s_{t+1}.

Figure 3 illustrates the difference of equation (1) and (3). In this example, 3 actions a_1, a_2 and a_3 can be selected in 1 state. For equation (1), the action a_1 which has largest Q-value is selected, but for equation (3), even though the largest Q-value is a_1, the action a_3 is selected because a_3 has largest absolute Q-value. In the actual learning, absolute value is not used.

Therefore, if the state has large negative Q-value, the negative worth is propagated around this state.

IV COMPUTER SIMULATIONS

In order to confirm the efficiency of proposed method, we experiments several computer simulations.

1 Problem

The problem treated here is as follows.

A bug moves in the closed 2 dimensional world. This bug is an agent in this model. We call this world "bait world", because several bait areas for the bug are put on this world. The bug eats the bait when he enters into the bait area. There are two kinds of bait. One is very good taste for the bug and gives positive reward to him, and the other is very bad and

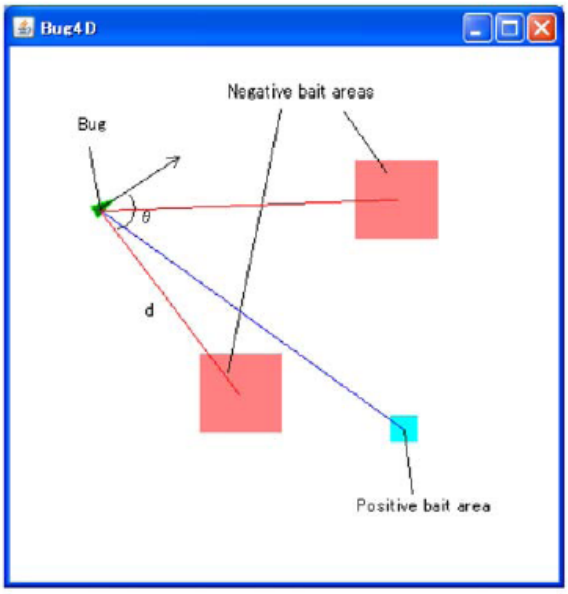


Figure 4: Image of experiments

negative reward is given. The objective of the bug is to eat as mush as baits during a specific period.

Figure 4 illustrates an image of "bait world". In this figure, the bug is represented as a triangle and the arrow of that triangle shows the direction of the bug. A large rectangular area is represented a negative bait area and a small one is positive area.

The agent observes angle and distance to each area. In the case of fig.4, since the number of bait area is 3, the agent observes 6 parameters. These 6 parameters construct the 6 dimensional state space.

The actions of the agent are 1:go straight 2:turn left and 3:turn right.

If the agent eat a bait, the position of the agent is randomly changed in the world.

2 Simulation conditions

In the simulations, "one turn" means a cycle of reinforcement learning — from an observation of the agent to an update of the Q-value. 100,000 turns make "one period". We executed 200 periods in "one experiment". we count the number of bait that the agent has eaten during one period.

Q-values are initialized to small random numbers. Positive reward is 1 and negative reward is -1. Other state gives reward 0.

Learning rate $\alpha = 0.1$ and discount factor $\gamma = 0.9$.

3 Results

We done following two types of simulation experiments.

exp1 1 positive area and 1 negative area are placed in the bait world.

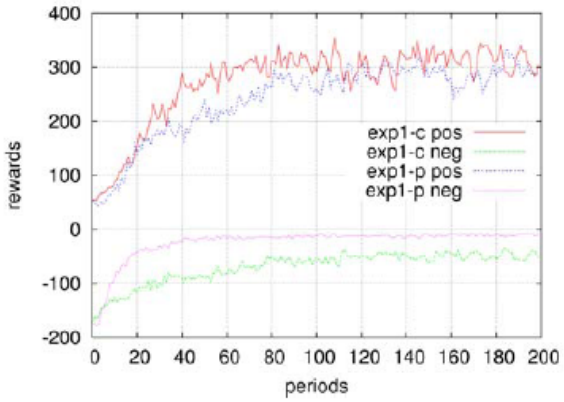


Figure 5: Result of 1 positive and 1 negative

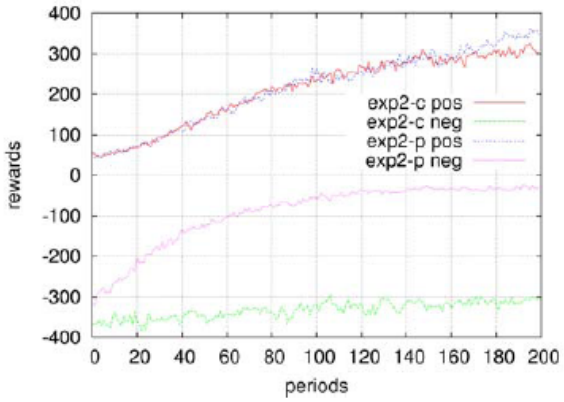


Figure 6: Result of 1 positive and 2 negatives

exp2 1 positive area and 2 negative areas are placed in the bait world.

Each experiment is done for 5 times and the results are averaged.

Figure 5 shows the result of exp1. "exp1-c" shows the conventional learning method and "exp1-p" shows the proposed method.

From this graph, the positive rewards are almost same on conventional method and proposed method, but the negative rewards are different. The number of negative rewards using proposed method is less than the one using conventional method. This means that proposed method is efficient for avoiding the bad situations and reaching to good goal.

Figure 6 show the result of exp2. In this case, the number of negative area is more than exp1. So the probability to trap the bad area is larger than exp1. From this result, you realize that the proposed method is much more efficient for avoiding the unworthy situations.

V CONCLUSION

1 Conclusion

In this paper, we proposed a new learning method which can propagate the worth of negative and showed that this method can useful for avoiding that unsuitable situations. Since this method works to propagate the negative worth around the bad state, the agent can be able to know the dangerous item at the far point of such negative areas.

And also this method is very simple way. The idea of this method is to use the absolute value of next state. By using absolute value when we select the next Q-value in learning, if the state has large negative value, it become to be able to propagate the negative worth.

The results of computer simulations, it is shown that this proposed method is very efficient for avoiding the bad situations.

2 Future works

In our experiments, the result of negative rewards have been improved but the positive one is not so much. This reason may be that in order to avoid the bad area, the agent went a long way to reach the positive goal.

Over spreading of negative worth often causes a wasteful action for the agent. To avoid this, we may have to control adaptively the discount factor γ. In this simulation, we use the constant number for γ.

References

[1] Richard S. Sutton and Andrew G. Barto, "Reinforcement Learning — An Introduction —", The MIT Press, 1998

[2] Leslie P. Kaelbling, Michael L. Littman and Andrew W. Moore, "Reinforcement Learning: A Survey", Journal of Artificial Intelligence Research 4 pp.237-285, 1996

[3] Soushi YAMADA, Takeshi OHASHI, Takichi YOSHIDA and Toshiaki EJIMA, "Research about Reinforcement Learning for Autonomic Robot in Multi-Agent Environment", IPSJ Kyushu Conference 2B-2 pp.177-186, 1998

[4] Hiroyuki OKADA, Hiroshi YAMAKAWA and Takashi OMORI, "Reinforcement Learning by Reward and Punishment", Technical Report of IEICE, NC99-100 pp.55-62 March 2000