The Sixteenth International Symposium on Artificial Life and Robotics 2011 (AROB 16th '11),
B-Con Plaza, Beppu,Oita, Japan, January 27-29, 2011

A Method for Finding Multiple Subgoals for Reinforcement Learning

Fuminori Ogihara and Junichi Murata

Kyushu University, 744, Motooka, Nishi-ku, Fukuoka, Fukuoka, Japan
(Tel : 092-802-3675; Fax : 092-802-3692)
(ogihara@cig.ees.kyushu-u.ac.jp)

Abstract: This paper proposes a new method for discovering multiple subgoals automatically to accelerate reinforcement learning. There have been proposed several methods for discovery of subgoals. Some use state visiting frequencies in the trajectories that reach the goal state. When a state visiting frequency is very high, this state is regarded as the subgoal. Because this kind of methods need that the goal state is reached many times to collect trajectories, they take a long time for discovering subgoals. In addition, they cannot discover the potential subgoals that will become appropriate subgoals when the goal state changes. On the other hand, some methods identify subgoals by partitioning local state transition graphs. But this kind of methods require large calculation amounts. We propose a new method that solves the above drawbacks. The new method utilizes state visiting frequencies. But we collect trajectories that go through particular non-goal states selected at random. For each particular state, trajectories are collected. Most of the trajectories reach the particular state more easily that the goal state. Therefore, it is expected that we can discover subgoals quickly and discover multiple subgoals together.

Keywords: reinforcement learning, subgoal discovery, the state visiting frequency, the particular state

I. Introduction

Reinforcement Learning is the method that decides proper actions at each state by trial and error. The trial and error method is effective when the environment is complicated or unknown. But learning by trial and error takes a long time. So it is important to accelerate the learning. There have been proposed several methods that accelerate reinforcement learning. Symmetrical-Actions [1] utilizes symmetry of the environment, and Macro-Actions [2] and Options [3] divide the environment to reduce the learning problem size.

Dividing environment by subgoals is one of the methods that speed up the learning, since the number of selections of states or actions reduces. To use this method, it is necessary to discover subgoals.

There have been proposed several methods for finding subgoals. These methods utilize the feature of subgoals [4]. Some utilize the feature that the trajectories that reach the goal state always go through subgoals. In this kind of methods, we collect only positive trajectories (the trajectories that reach the goal state) and ignore negative ones (the trajectories that do not reach the goal state). When the state visiting frequency, which is the value given by dividing the number of positive trajectories into the visiting counts of the state, is very high, it is judged that the state is the subgoal. However, it takes a long time to collect a number of positive trajectories. On the other hand, some methods utilize the transitions between states [5]. If the number of transition paths between two states is small, the learning environment can be cut between those states and they are judged as subgoals. However, this kind of methods must count the transitions for all neighboring pairs of states, and judge if each count of transitions is lower than the threshold. So they require large calculation amounts.

In this paper, we propose a new method for finding multiple subgoals. The purpose of this paper is discovery of multiple subgoals. In addition, we solve the drawbacks of the existing methods. The new method utilizes the state visiting frequencies. However, we collect not only positive trajectories but also negative ones that go through particular non-goal states chosen at random. These trajectories can be collected faster and thus the subgoals can be found more quickly.

The structure of this paper is as follows. In Section 2 we explain subgoals. In Section 3 we introduce a new method for finding multiple subgoals. Experiment of the use of the new method and the results are written in Section 4. And Section 5 gives conclusions.

II. Subgoal Overview

Subgoals are the states that the agent must go through before the task goal is reached. In

reinforcement learning, if the environment is large and many states or actions are involved, learning tends to be slow owing to trial and error. So it is very effective for reduction of the number of states or actions taken to divide the task environment at subgoals. For example, when an agent goes from one room to another room by way of the doorway, it is very difficult that the agent goes to another room with random actions. But it is easier that the agent goes to the doorway. In this case, the agent must go through the doorway to reach another room. So the doorway becomes a subgoal, which divides the whole task into two smaller tasks.

The methods of Macro-Actions and of Options also divide the environment for speeding up learning. They can be readily constructed once subgoals have been discovered. So finding subgoals is very useful.

III. A new method for finding multiple subgoals

The conventional methods that utilize the state visiting frequency for finding subgoals have several drawbacks [4]. The goal state must be reached many times to count the visiting frequencies, and finding subgoals takes a long time in large or complicated environments. Besides, if the goal state changes, subgoals that were found are not effective for the new situation. In addition, because the trajectories go through the states near the start state very often, the visiting frequencies of these states may be high and these states are judged as subgoals by mistake.

We propose a new method for finding multiple subgoals that solves the above drawbacks. The new method utilizes the state visiting frequencies. But in this method, the definition of the state visiting frequency is different from the past definition in terms of the two points.

First, the way of collecting trajectories is different. The agent takes random actions for several episodes and then selects several particular states among the states that have already been visited. We collect the trajectories that go through the particular state regardless of whether the task goal is reached or not. The state visiting counts are counted up for the states between the start state and the particular state. Figure 1 shows the way of collecting trajectories. Here is one positive trajectory, so we can collect only one trajectory for the existing method. But if we use the new method, we can collect three trajectories including two negative ones. Therefore, we can obtain enough data to accurately calculate the visiting frequencies and the doorway between two rooms will be selected as the subgoal more quickly. Besides for each particular state, trajectories are collected. So it is expected that multiple subgoals are found.

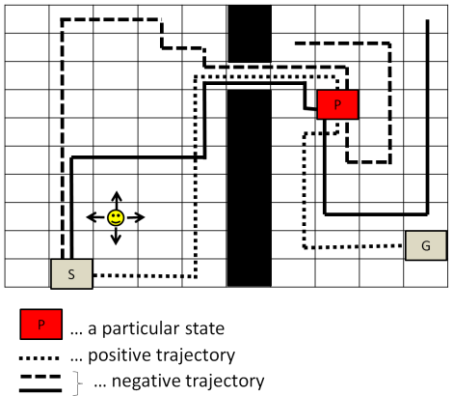

Fig.1 The trajectories that go through a particular state.

Second, in this method, the visiting counts are divided by the average state visiting frequency. The average state visiting frequency is the expected frequency that the trajectory goes through the state when the agent selects an action with equal probability at any state. Figure 2 shows the average state visiting frequency of each state around the start state. The average visiting frequencies of the states around the start state tend to be high. On the other hand, as the state becomes far from the start states, the average state visiting frequency becomes low. So dividing the observed visiting frequency by the average state visiting frequency reduces the seemingly high visiting counts of the states around the start state, and prevents that these states from being judged as subgoals by mistake

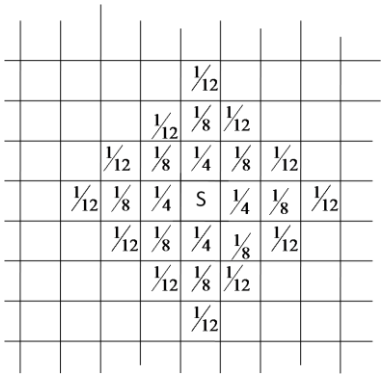

Fig.2 The average state frequency of each state in a gridworld environment.

The visiting frequency F_i of state i for the particular state p is defined by

$$F_i = \frac{\text{the visiting count of state } i}{\frac{\text{\# of trajectories that go through the particular state } p}{\text{the average visiting frequency of state } i}}$$

The Algorithm of the new method is sketched below:

1. At the first several episodes, the agent selects random actions at any states.
2. The agent selects several particular non-goal states at random among the states it has already visited. In addition, if the goal state has been reached during these episodes, the states belonging to a positive trajectory and the goal state are selected as the particular states, because this leads to easy finding of the effective subgoals.
3. The agent collects trajectories that go through the particular state while it learns the policy and selects actions with Q-learning.
4. For each particular state, the state visiting frequencies are calculated.
5. For each particular state, if the state visiting frequency of a state is the maximum over all the states on the way to the particular state and higher than the threshold, the state is judged as the subgoal.

IV. Experiment

We test the new method for finding multiple subgoals. Figure3 shows the environment used in the experiment. It consists of three rooms. The start state and the goal state are indicated by S and G, respectively, in Fig.3. Besides, the horizontal axis and the vertical axis are indicated by *x* and *y*, respectively, and the state is represented by (*x*,*y*). The location of the upper doorway is (3,10) and the right doorway is (10,5). The agent can select from four actions: going up, down, right, and left. It uses Q-learning with ε-greedy with ε that is 0.3 at the first episode and is reduced by 0.001 at every episode down to 0.1. The learning rate α in Q-learning is 0.1, and the discount rate γ is 0.9.

The agent takes random actions for the first 25 episodes. Then the agent uses Q-learning, and collects the trajectories that go through the particular non-goal states for 300 episodes. The fixed threshold 36 is used.

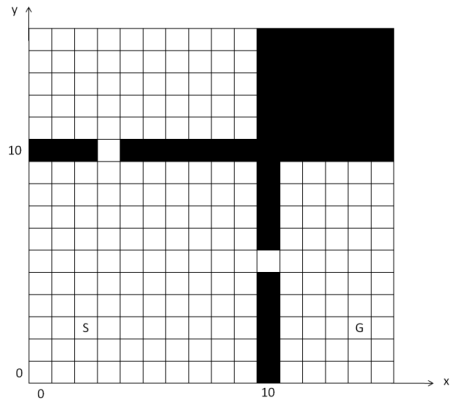

Fig.3 The environment.

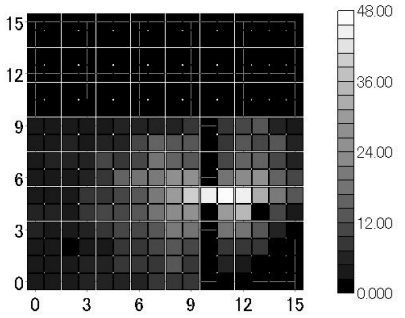

Fig.4 State visiting frequencies for the particular state (13,4).

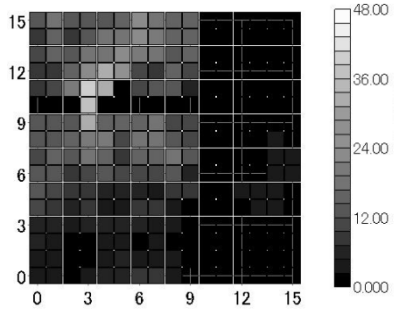

Fig.5 State visiting frequencies for the particular state (5,11).

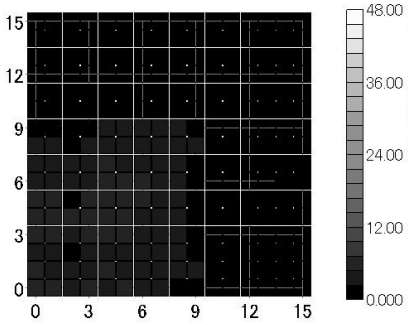

Fig.6 State visiting frequencies for the particular state (2,5).

Figure 4 through Fig.6 show the state visiting frequencies. The lighter the state's color is, the higher the state visiting frequency is. In Fig.4 the particular state is located in the right room where the goal state locates. This particular state is situated near the right doorway, so collecting the trajectories that go through the particular state is easier than collecting the trajectories that reach the goal state. In fact, the number of the trajectories that go through the state (13,4) is 24, which is higher than the number of positive trajectories that is 15. Figure 4 shows that the visiting frequencies of the states around the right doorway are high. The maximum visiting frequency is the visiting frequency of the state (11,5) that is 48 and higher than the threshold. So the state (11,5) is selected as the subgoal.

In Fig.5 the particular state is located in the upper room and near the upper doorway. For this particular state, the visiting frequencies of the states around the upper doorway are high. It is not useful to select one of these states as the subgoal in this environment. But for example, when the goal state moves to the upper room, the upper doorway will become the good subgoal. In other words, even if the goal state moves to the upper room, the agent will not need to repeat the discovery of subgoals. The maximum visiting frequency is the visiting frequency of the state (3,11) that is 40 and higher than the threshold. So the state (3,11) is selected as the subgoal. Again, the number of the trajectories that go through the state (5,11) is 20, which is higher than the number of positive trajectories that is 15.

In Fig.6 the particular state is located in the start room. Because the trajectories often go through this particular state before the agent gets out the start room, the visiting frequency of the doorway state is very low, and the visiting frequency of one of the states in the start room is the maximum. But the visiting frequencies overall are generally low. The maximum visiting frequency is the visiting frequency of the state (3,4) that is about 6.8 and lower than the threshold. So no state is selected as the subgoal from the trajectories passing through this particular state.

VI. Conclusions

In this paper we propose a new method for finding multiple subgoals that solves the existing drawbacks. We tested it in the three-room gridworld environment.

The experiment shows that our new method can solve the conventional methods' drawbacks. First, the method can collect the trajectories that go through the particular state more than the positive trajectories. So, it can find subgoals more quickly. Second, it can find the state that may become an effective subgoal when the goal state changes. Third, the use of the average state visiting frequencies suppresses the erroneous selection of the states around the start state as the subgoals.

But our method has a drawback. The particular non-goal states are selected at random. If all the particular states happen to be located in the start room, any doorways will not be selected as the subgoals. To solve the drawback, there need some devises, for example, increasing the number of the particular states, or selecting each particular state to be distant from other particular states.

References

[1] M.A.S.Kamal and J.Murata (2004), Reinforcement Learning for High Dimensional Problem with Symmetrical Actions. Proc. of 2004 IEEE Int. on Systems, Man, and Cybernetic: 6192-6197.

[2] T. Tateyama, S. Kawata, T. Oguchi (2002), Automatic Generation of Macro-Actions using Genetic Algorithm for Reinforcement Learning, SICE2002 AUG5-7, Osaka.

[3] Martin Stolle and Doina Precup (2002), Learning Options in Reinforcement Learning, SARA 2002, LNAI 2371, 213-223.

[4] R. Mattew Kretchmar, Todd Feli, Rohit Bansal (2003), Improved Automatic Discovery of Subgoals for Options in Hierarchical Reinforcement Learning, Jornal of Computer Science & Technology, Vol.3-No.2.

[5] Özgür Şimşek, Alicia P. Wolfe, Andrew G. Barto (2005), Identifying Useful Subgoals in Reinforcement Learning by Local Graph Partitioning, the Twenty-second International Conference on Machine Learning, 816-823.

[6] Shie Mannor, Ishai Menache, Amit Hoze, Uri Klein (2004), Dynamic Abstraction in Reinforcement Learning via Clustering, twenty-first International Conference on Machine Learning, 71-77.