Artif Life Robotics (2013) 18:104–108
DOI 10.1007/s10015-013-0106-0

ORIGINAL ARTICLE

Reinforcement learning for dynamic environment: a classification of dynamic environments and a detection method of environmental changes

Masato Nagayoshi · Hajime Murao · H. Tamaki

Received: 11 March 2013 / Accepted: 30 August 2013 / Published online: 19 September 2013

Abstract Engineers and researchers are paying more attention to reinforcement learning (RL) as a key technique for realizing computational intelligence such as adaptive and autonomous decentralized systems. In general, it is not easy to put RL into practical use. In prior research our approach mainly dealt with the problem of designing state and action spaces and we have proposed an adaptive co-construction method of state and action spaces. However, it is more difficult to design state and action spaces in dynamic environments than in static ones. Therefore, it is even more effective to use an adaptive co-construction method of state and action spaces in dynamic environments. In this paper, our approach mainly deals with a problem of adaptation in dynamic environments. First, we classify tasks of dynamic environments and propose a detection method of environmental changes to adapt to dynamic environments. Next, we conducted computational experiments using a so-called "path planning problem" with a slowly changing environment where the aging of the system is assumed. The performances of a conventional RL method and the proposed detection method were confirmed.

Keywords Reinforcement learning · Dynamic environment · Slowly changing environment · Detection of environmental changes · Entropy

1 Introduction

In recent years, artificial systems have become more complicated and large-scaled. The conventional way, in which systems are controlled in a top-down manner mainly by humans, is facing up to the difficulties of not only optimality but also adaptability and flexibility. One of the solutions to this issue is to develop an autonomously adaptive system.

Engineers and researchers are paying more attention to reinforcement learning(RL) [1] as a key technique of realizing autonomous systems. In general, however, it is not easy to put RL into practical use. Such issues as satisfying the requirement of learning speed, resolving the perceptual aliasing problem, designing reasonable state and action spaces of an agent and adapting dynamic environments must be resolved. In prior research, our approach mainly dealt with the problem of designing state and action spaces and we have proposed a co-construction method of state and action spaces [2]. However, it is more difficult to design state and action spaces in dynamic environments than in static ones. Thus, it may be even more effective to use an adaptive co-construction method of state and action spaces in dynamic environments.

In this paper, our approach deals with problems of adaptation in dynamic environments. Previously, many methods for dynamic environments have been proposed.

This work was presentedin part at the 18th International Symposium on Artificial Life and Robotics, Daejeon, Korea, January 30–February 1, 2013.

M. Nagayoshi (✉)
Niigata College of Nursing, 240 Shinnan, Joetsu 943-0147, Japan
e-mail: nagayosi@niigata-cn.ac.jp

H. Murao
Faculty of Cross-Cultural Studies, Kobe University, 1-2-1 Tsurukabuto, Nada-ku, Kobe 657-8501, Japan

H. Tamaki
Graduate School of Engineering, Kobe University, Rokko-dai, Nada-ku, Kobe 657-8501, Japan

However the researchers have only referred to the dynamic environment, without focusing on the specific problems that are inherent in a dynamic environment. We make a classification of dynamic environments and propose a detection method of environmental changes to adapt a dynamic environment. In addition, computational experiments are conducted by using a so-called “path planning problem” with a slowly changing environment where the aging of the system is assumed. The performances of a conventional RL method and the proposed detection method are confirmed.

2 Q-learning

Q-learning works by calculating the quality of a state-action combination, namely the Q-value, that gives the expected utility of performing a given action in a given state. By performing an action $a \in \mathcal{A}_{\mathrm{Q}}$, where $\mathcal{A}_{\mathrm{Q}} \subset \mathcal{A}$ is the set of available actions in Q-learning and $\mathcal{A}$ is the action space of the agent. The agent can move from state to state. Each state provides the agent with a reward r. The goal of the agent is to maximize its total reward.

The Q-value is updated according to the following formula, when the agent is provided with the reward:

$$Q(s(t-1), a(t-1)) \leftarrow Q(s(t-1), a(t-1)) + \alpha_{\mathrm{Q}}\{r(t-1) + \gamma \max_{b \in \mathcal{A}_{\mathrm{Q}}} Q(s(t), b) - Q(s(t-1), a(t-1))\} \tag{1}$$

where $Q(s\ (t-1), a(t-1))$ is the Q-value for the state and the action at the time step $t-1, \alpha_{\mathrm{Q}} \in [0,1]$ is the learning rate of Q-learning, $\gamma\ \in [0,1]$ is the discount factor.

The agent selects an action according to the stochastic policy $\pi\ (a|s)$, which is based on the Q-value. $\pi\ (a|s)$ specifies the probabilities of taking each action a in each state s. Boltzmann selection, which is one of the typical action selection methods, is used in this research. Therefore, the policy $\pi(a|s)$ is calculated as

$$\pi(a|s) = \frac{\exp(Q(s,a)/\tau)}{\sum_{b \in \mathcal{A}} \exp(Q(s,b)/\tau)} \tag{2}$$

where τ is a positive parameter labeled temperature.

3 Dynamic environments

Dynamic environments are time-varying environments, i.e. when the state transition probability $T\ (s(t), a\ (t), s(t+1))$, which is the probability of transition from $s(t)$ to $s(t+1)$ under $a(t)$, changes, or the return function $R\ (s, a)$, which is the reward at s under a, changes, an “environmental change” has occurred and the environment is a dynamic environment. Hereafter on the premise of time-variance, the state transition probability and the reward function including time t are shown $T_t\ (s(t), a\ (t), s\ (t+1))$ and $R_t\ (s, a)$ respectively.

The learning becomes difficult when environmental changes occur as in the following formula:

$$\arg\max_{a \in \mathcal{A}} R_t(s(t), a) \neq \arg\max_{a \in \mathcal{A}} R_{\mathbf{t+1}}(\mathbf{s}(\mathbf{t}), \mathbf{a}) \tag{3}$$

where the best action changes over time in the states where the agent transits. In particular, in the course of making the action selection probability of the best action larger, environmental changes occur. After such an occurrence, the agent first needs to make the action selection probability of the action smaller, for example if the learning module is Q-learning, in the course of making the entropy of action selection probability $H\ (s)$:

$$H(s) = -(1/\log|\mathcal{A}_{\mathrm{Q}}|) \sum_{a \in \mathcal{A}_{\mathrm{Q}}} p(a|s) \log p(a|s), \tag{4}$$

the agent needs a process for finding the best action again. This process is the reason for learning difficulties in dynamic environments.

4 Classification of dynamic environments

Until now, Simada et.al. [3] divide ‘share states’ into 2 types: if the following equation is satisfied, then the state is the ‘share state’ of type1, and if not, the state is the ‘share state’ of type 2.

$$\arg\max_{a \in \mathcal{A}} R_i(s, a) = \arg\max_{a \in \mathcal{A}} R_j(s, a) \tag{5}$$

Also, they indicate that determining adaptability to dynamic environments is dependent on the abilities of the “detection of the states of type 1 and type 2” and “reusing and re-studying of learning results”.

In this section, previous works are organized by classifying this tasks of dynamic environments more finely.

For the sake of ease, we limit to episodic tasks such as acquiring a series of actions from start states to goal states. In our classification of dynamic environments we use the Boltzmann selection method [Eq. (2)], that is, the agent selects an action with a larger value based on a higher probability.

Here we assume that each differential value of the entropy of the action selection probability $\dot{H}_{\mathrm{C}}(T, M, t_{\mathrm{E}})$ is given and is caused by environmental changes, where t_{E} is the episode number, i.e. 1 episode is defined as the period from when the agent is located at a start state to when the agent arrives at a goal state, M is the method used, and T is the task.

1. The (a) presence or (b) absence of the influence of environmental changes: If the following formula is satisfied, then the agent can adapt to environmental changes by the learning performance of the method M even if environmental changes occurred.

$$\forall t_{\mathrm{E}}\,(\dot{H}_{\mathrm{C}}(T, M, t_{\mathrm{E}}) \leq 0) \tag{6}$$

2. The process of environmental changes: In almost all previous works, it is assumed that the cases of (1) $\exists_1 t_{\mathrm{E}}\,(\dot{H}_{\mathrm{C}}\,(T, M, t_{\mathrm{E}}) > 0)$ or $\dot{H}_{\mathrm{C}}\,(T, M, t_{\mathrm{E}})$ is a repetition of a shape similar to an impulse function. In particular, those tasks in which the characteristic changes momentarily or stages are switched are assumed. On the other hand, (2) $\dot{H}_{\mathrm{C}}(T, M, t_{\mathrm{E}})$ is a continuously small positive value, if the aging of the system or a change of human characteristics is assumed.
3. The timing of the appearance of environmental changes in relation to learning progress: It is known that the entropy of the action selection probability approaches 0 in tasks without environmental changes [4]. Here, if the task has no environmental changes, it could be divided into 3 phases based on the entropy of the action selection probability. Using terms borrowed from biological cell division, we have labeled these 3 phases a prophase $\mathbf{t}_{\mathrm{EB}}$, a metaphase $\mathbf{t}_{\mathrm{EM}}$, and a anaphase of learning $\mathbf{t}_{\mathrm{EL}}$. The task can be classified according to when the appearance of environmental changes, $t_{\mathrm{EC}} = \min\{t_{\mathrm{E}} | \dot{H}_{\mathrm{C}}(T, M, t_{\mathrm{E}}) > 0\}$ occurs: (i) $\mathbf{t}_{\mathrm{EB}}$ (ii) $\mathbf{t}_{\mathrm{EM}}$ or (iii) $\mathbf{t}_{\mathrm{EL}}$. Normally, in the case of the task having $t_{\mathrm{EC}} \in \mathbf{t}_{\mathrm{EB}}$, it is easy to adapt to environmental changes since the entropy of the action selection probability is large. In the case of the task having $t_{\mathrm{EC}} \in \mathbf{t}_{\mathrm{EL}}$, influences of environmental changes become large since the entropy is small.

Except in (b) the absence of influences of environmental changes, it is necessary for researchers to take the remaining $2 \times 3 = 6$ types into consideration. In this paper, we focus on the above six types.

In viewing of previous works organized in consideration of the above classification of tasks of dynamic environments, Takahashi et al. [5] deal with 3 types: (i), (ii), and (iii) in the case of (a) (1) above by introducing an evaluation index of learning progress in a detection method of environmental changes. Other works deal with the types (a) (1) (ii) and (a) (1) (iii) [3] as above. Thus, it is necessary for researchers to consider tasks of dynamic environments regarding the types other than these above.

5 Detection method of environmental changes

The entropy of the action selection probability $H\,(s)$, shown in Eq. 4, becomes smaller in tasks without environmental changes. In contrast, when an environmental change occurs, the entropy becomes larger from the time of the occurrence. Hence, when the entropy $H\,(s)$ begins to increase, the agent is able to detect the environmental change.

However, Preliminary computational experiments indicate that the entropy of the action selection probability in the state s (Eq. 4) shows a range of fluctuations even if in a static environment. In order to decrease false-detections of environmental changes by the influence of the fluctuations, the agent detects the time of the occurrence of the reversal of a downward trend using MACD (moving average convergence/divergence) [6], which is one of the most popular tools in technical analysis trading. The entropy $H_{\mathrm{D}}^{+}(s)$, after updating the Q-value, is used to refine the detection only when the agent selects an action with the maximal Q-value, if the learning module is Q-learning. Then, a short-term $(n = \theta_{\mathrm{EMAS}})$ EMA (exponential moving average) value and a long-term $(n = \theta_{\mathrm{EMAL}})$ EMA value of the entropy are calculated according to the following equation at a rate of once every θ_t update of $H_{\mathrm{D}}^{+}(s)$. If θ_t is set at a large value, then the accuracy is fine but the speed of the detection from the occurrence time of the environmental change is slow. On the other hand, if θ_t is set at a small value, then the speed is fast but the accuracy is disrupted.

$$\mathrm{EMA}^n\,(s(t)) = (1 - \alpha) \times \mathrm{EMA}^n_{\mathrm{old}}(s(t)) + \alpha \times H_{\mathrm{D}}^{+}\,(s(t)) \tag{7}$$

where EMA^n is a n-term EMA value, $\mathrm{EMA}^n_{\mathrm{old}}\,(s(t))$ is the latest known value of EMA in $s\,(t)$, $\alpha = 2/(n + 1)$ and n are constant numbers expressing the smoothing constant and the average amount of time respectively.

A MACD (moving average convergence/divergence) value is calculated according to the following equation, after updating the short-term and the long-term EMA values.

$$\mathrm{MACD}(s) = \mathrm{EMA}^{\theta_{\mathrm{EMAS}}}\,(s) - \mathrm{EMA}^{\theta_{\mathrm{EMAL}}}\,(s) \tag{8}$$

In addition, a 'signal' value is a moving average for the latest series of θ_{MACD} values of $\mathrm{MACD}(s)$. In particular, when $\mathrm{MACD}(s)$ becomes larger than the signal value and $\mathrm{MACD}(s) < 0$ and the following formula is satisfied from the condition of $\mathrm{MACD}(s)$ being smaller than the signal value, the agent detects the environmental change.

$$\mathrm{MACD}\,(s) - \mathrm{MACD}_{\mathrm{old}}\,(s) > \theta_{\mathrm{M}} \tag{9}$$

where, $\mathrm{MACD}_{\mathrm{old}}(s)$ is the latest known value of MACD (s). The differential value θ_{M} of MACD becomes infinitesimally small as the learning progresses, without environmental changes. Here, the above formula has been added in order to detect environmental changes only when a rapidly increasing trend occurs.

Table 1 Usual parameters of MACD

Parameter	Value	Parameter	Value
θ_{EMAS}	12	θ_{EMAL}	26
θ_{MACD}	9		

Usual parameters of MACD shown in Table 1 to detect environmental changes are used in the following experiments.

When the environmental change is detected, all Q-values in the detected state are set to the average value of the Q-values in the detected state.

6 Computational examples

Q-learning (hereafter called "QL") and the proposed detection method (hereafter called "PD") are applied to a so-called "path planning problem" with a slowly changing environment where the aging of the system is assumed in a continuous state and action spaces, as shown in Fig. 1. Here, the agent has a circular shape [diameter 50 (mm)], and the continuous space is 500 × 500 (mm) bounded by the external wall, with internal walls as shown in black. One of internal walls is slowly extended to the right from an episode t_{ES} to an episode t_{EE} as shown in the continuous space pictured on the right of Fig. 1. The agent can observe the center position of itself $(x_{\mathrm{A}}, y_{\mathrm{A}})$ as the input, and decide the direction θ_{A} as the output. The agent moves 25 (mm) in a direction defined by θ_{A} to which gaussian noise has been added.

The positive reinforcement signal $r_t = 10$ (reward) is given to the agent only when the center of the agent arrives in the goal area, and the reinforcement signal is $r_t = 0$ at all other steps. The period from when the agent is located at the starting point to when the agent is given a reward, labeled as 1 episode, is repeated.

After dividing the state space evenly into 20 × 20 spaces, and the action space evenly into 8 spaces, QL and PD based on QL are compared with 3 occurrence times of the environmental change: $\mathbf{t}_{\mathrm{EB}} = \mathbf{50}, \mathbf{t}_{\mathrm{EM}} = \mathbf{500}$, and $\mathbf{t}_{\mathrm{EL}} = \mathbf{1{,}500}$ (episode). The internal wall is extended to $x = 500$ during 175 (episodes) from the occurrence of the environmental change.

Computer experiments have been carried out with the parameters of Q-learning: $\alpha_{\mathrm{Q}} = 0.1, \tau = 0.1$, and $\gamma = 0.9$. In addition, the rating number θ_t of the detection method is set at 5 by trial and error, the differential value θ_{M} of MACD is set at 0.01 by trial and error between greater than 0 and less than 0.1. All initial Q-values are set at 5.0 as the optimistic initial values [1].

The average number of steps required to accomplish the task was observed during learning over 20 simulations with QL and PD, as described in Figs. 2 and 3 respectively. The average number of detections, that is the average number of states where the environmental change is detected, was observed during learning over 20 simulations with PD, as described in Fig. 4.

It can be seen from Figs. 2 and 3 that, (1) when the environmental change occurs later, that is, as the learning progresses, the influence of the environmental change becomes larger, (2) PD has better performances than QL with regard to the influences of the environmental change. It can be seen from Fig. 4 that, (3) PD has better performances to detect the environmental change as the learning progresses, (4) but PD has few false-detections, which are detections before the environmental change. Then, we have confirmed that (5) the average number of detections became gradually larger after episode 3,000.

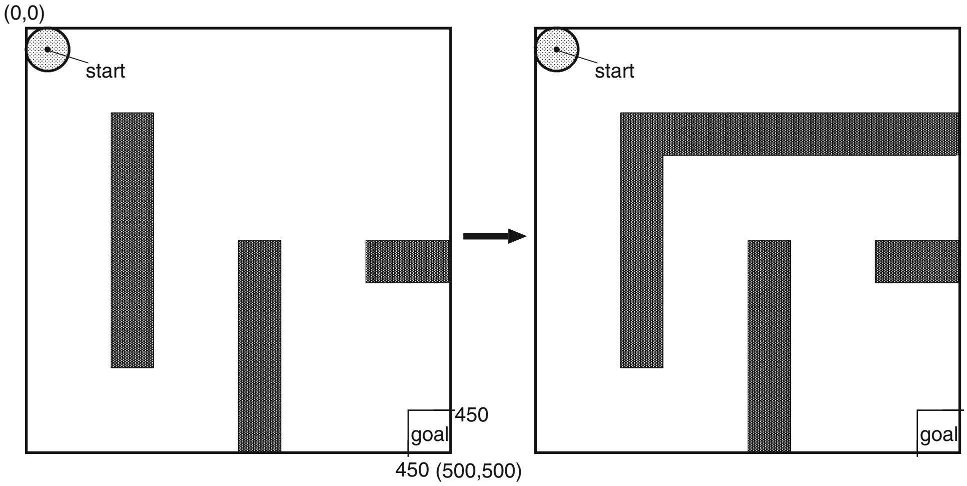

Fig. 1 Dynamic path planning problem

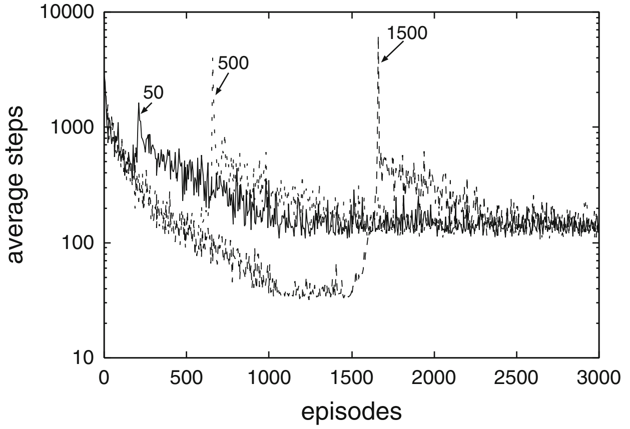

Fig. 2 Required steps of 3 occurrence times of the environmental change by Q-learning

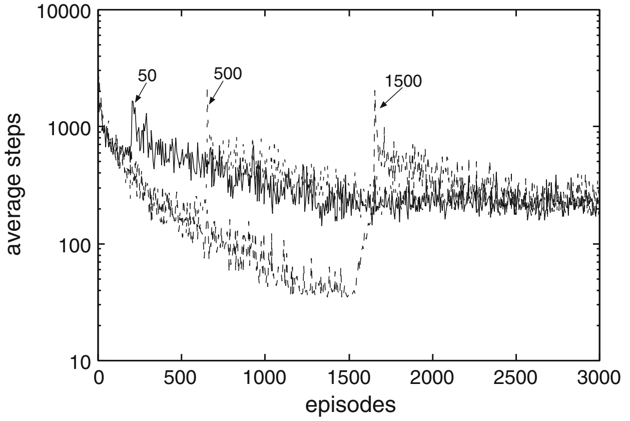

Fig. 3 Required steps of 3 occurrence times of the environmental change by the proposed method

7 Conclusion

We have classified tasks of dynamic environments and proposed the detection method of environmental changes to adapt to dynamic environments. Then, with computational experiments we confirmed that the proposed method has better performances than Q-learning with regard to the influences of the environmental change, and as the learning progresses, detection of the environmental changes improves when the environmental change occurs, but the

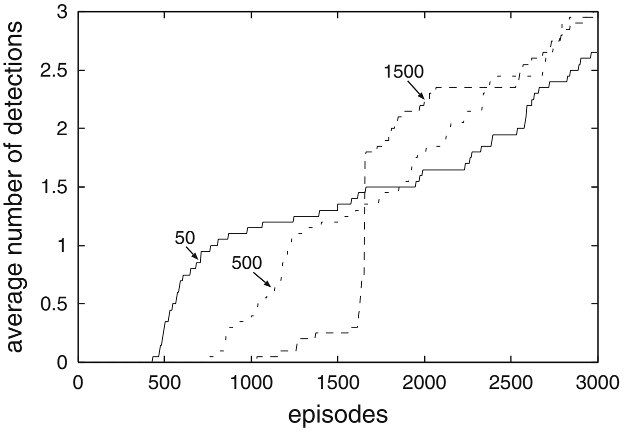

Fig. 4 Number of detections of 3 occurrence times of the environmental change

number of detections becomes larger without environmental changes.

Our future projects include (1) to upgrade the detection method to consider 3 occurrence times of the environmental change, and (2) to apply the adaptive co-construction method of state and action spaces in dynamic environments.

References

1. Sutton RS, Barto AG (1998) Reinforcement learning, a Bradford book. MIT Press, London
2. Nagayoshi M, Murao H, Tamaki H (2012) Developing reinforcement learning for adaptive co-construction of continuous high-dimensional state and action spaces. Artif Life Robot 17(2):204–210
3. Shimada S, Anzai Y (2001) Improving adaptability of reinforcement learning system to dynamic environment by decomposing and reusing macro-operators. J IEICE J84-D-I(7):1076-1088 (in Japanese)
4. Nagayoshi M, Murao H, Tamaki H (2006) A state space filter for reinforcement learning. In: Proceedings of AROB 11th'06, pp 615–618 (GS1-3)
5. Takahashi T, Adachi M (2006) Evaluating progress of reinforcement learning using recurrence plots. In: IEICE Technical Reports, NLP2005-155, pp 25–30 (in Japanese)
6. Appel G (2005) Technical analysis power tools for active investors. Financial Times Prentice Hall, London