Actor-Critic Reinforcement Learning System with Time-Varying Parameters

Masanao Obayashi*, Kosuke Umesako* , Tazusa Oda* , Kunikazu Kobayashi* ,
and Takashi Kuremoto*

*Department of Computer Science and Systems Engineering, Yamaguchi University, Yamaguchi, Japan
(Tel : +81-836-85-9518; E-mail: m.obayas@yamaguchi-u.ac.jp)

Abstract: Recently reinforcement learning has attracted attention of many researchers because of its simple and flexible learning ability for any environments. And so far many reinforcement learning methods have been proposed such as Q-learning, actor-critic, stochastic gradient ascent method and so on. The reinforcement learning system is able to adapt to changes of the environment because of the mutual action with it. However when the environment changes periodically, it is not able to adapt to its change well. In this paper we propose the reinforcement learning system that is able to adapt to periodical changes of the environment by introducing the time-varying parameters to be adjusted. It is shown that the proposed method works well through the simulation study of the maze problem with aisle that opens and closes periodically, although the conventional method with constant parameters to be adjusted does not works well in such environment.

Keywords: reinforcement learning, neural network, time-varying parameter, periodical change

1. INTRODUCTION

The reinforcement learning, which learns the optimal action through rewards/penalties for its own action without knowledge of the environment, has attracted many researchers because of its simple and flexible learning ability. And so far many reinforcement learning methods have been proposed such as Q-learning, actor-critic, stochastic gradient ascent method and so on[1][2] . They would be able to adapt the change of the environment because of its adaptability. However, because that these systems already proposed have constant parameters, they don't work well for periodic changes of environments. In this paper we propose the reinforcement learning method which would be able to adapt the periodical change of the environment introducing the time-varying parameters, that would take values of the range [0,1], assigning to each actions. It is shown that the proposed method works well through the simulation study of the maze problem with aisle that opens and closes periodically, although the conventional method with constant parameters to be adjusted does not works well in such environment.

2. ACTOR-CRITIC REINFORCEMENT LEARNING SYSTEM

In this section actor-critic reinforcement learning system used in this paper is explained briefly. The advantages of actor-critic method are as follows;

1) it would be able to deal with the case of the continuous action space,
2) it works well in the Non-Markov Decision Process environment considering the stochastic action selection.

2.1 Construction of the actor-critic reinforcement learning system

The construction of the reinforcement learning system with actor-critic is shown in Fig.1. In Fig.1 the critic outputs P(t) : the prediction value of sum of the discounted rewards that will be gotten over the future. The prediction error $\hat{r}(t)$ is calculated. In the critic the parameters are adjusted in order to reduce the error $\hat{r}(t)$.

The actor outputs $a(t)$: the input to the environment. In the actor the parameters are adjusted to maximize the prediction value P(t) . Here $X(t)=[x_1, x_2, \cdots, x_n]$ is the state vector of the environment. n(t) is noise used for searching the optimal action, which would make the maximization of the expected discounted return.

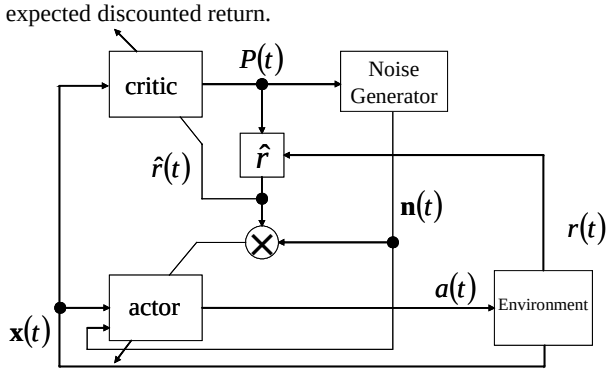

Fig. 1 The construction of the reinforcement learning system

2.1.1 Construction and function of the actor

Fig.2 shows the construction of the actor. The actor is basically consisted of Radial Basis Function Network. The jth basis function of the middle layer node is as follows;

$$y_j = \exp\left[-\sum_{i=1}^{n}(x_i - m_{ij})^2 \Big/ \sigma_{ij}^2\right] \quad (1)$$

Here y_j : jth output of the middle layer, m_{ij} , σ_{ij}^2 : center, dispersion for ith input of jth basis function respectively.

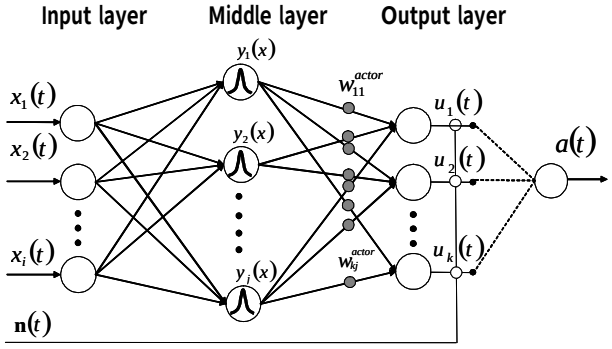

Fig.2 Construction of the actor

The action a_b on time t is selected stochastically using Gibbs distribution Eq.(2).

$$P(a_b \mid \mathbf{x}(t)) = \frac{\exp(u_b(t)/T)}{\sum_{k=1}^{K} \exp(u_k(t)/T)} \quad (2)$$

$$u_k(t) = \sum_{j=1}^{J} w_{kj} y_j(t) + n_k(t) \quad (3)$$

Here $P(a_b \mid \mathbf{x}(t))$: bth action, a_b, selection probability, T : positive constant called temperature constant. K: number of the actions, n_k : additive noise to kth output, u_k : representative value of kth action, w_{kj} : connection weight from jth node of the middle layer to kth output.

2.1.1 Construction and function of the critic

Function of the critic is calculation of P(t) : the prediction value of sum of the discounted rewards that will be gotten over the future and of its prediction error. These are shortly explained as follows;

The sum of the discounted rewards that will be gotten over the future is defined as $V(t)$.

$$V(t) \equiv \sum_{n=0}^{\infty} \gamma^n \cdot r(t+n), \quad (4)$$

where γ ($0 \le \gamma < 1$) is constant called discount rate.
Eq. (4) is rewritten as

$$V(t) = r(t) + \gamma V(t+1). \quad (5)$$

Here the prediction value of $V(t)$ is defined as $P(t)$.
The prediction error $\hat{r}(t)$ is expressed as follows;

$$\hat{r}(t) = r(t) + \gamma P(t+1) - P(t). \quad (6)$$

The parameters of the critic are adjusted to reduce this p rediction error $\hat{r}(t)$. The prediction error $P(t)$ is calculated as follows;

$$P(t) = \sum_{j=0}^{J} w_j y_j(t) \quad (7)$$

Here J : number of nodes in the middle layer of the critic, w_j : weight of the jth output, y_j : jth output of the middle layer of the critic. The construction of the critic is also consisted of the RBFN as shown in Fig.3.

2.1.3 Noise generator

Noise generator let the output of the actor have the diversity by adding the noise to it. It comes to realize the learning of the trial and error. Calculation of the noise $n(t)$ is as follows;

$$n(t) = noise_t \cdot \min(1, \exp(-P(t))), \quad (8)$$

where $noise_t$ is uniformly random number of $[-1,1]$. As the $P(t)$ will be bigger, the noise will be smaller. This leads to the stable learning of the actor.

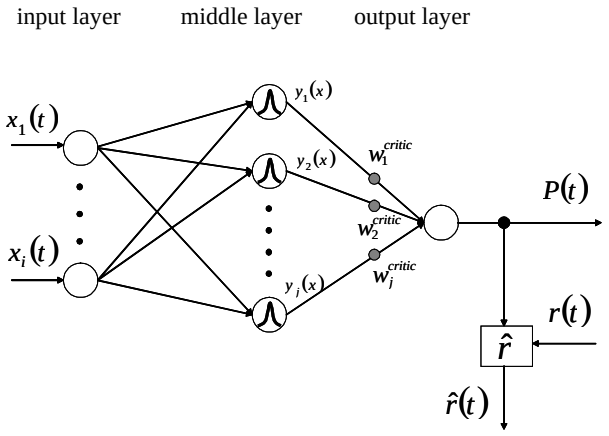

Fig.3 Construction of the critic

3. PROPOSED SYSTEM

In this section the proposed reinforcement learning system whose function varies periodically in order to adapt the periodical changes of the environment. The actor has above function in modules as shown in Fig.4..

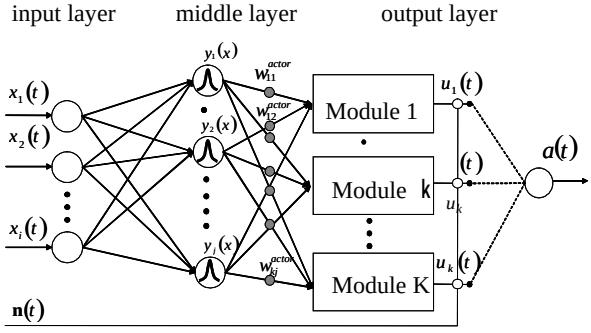

Fig.4 Construction of the actor with time-varying modules

In Fig.4 module k outputs u_k expressed as Eq.(9).

$$u_k(t) = \sum_{l=1}^{N} \frac{\lambda_l(t)}{N} \cdot u_{l,k}(t) + \left(1 - \sum_{l=1}^{N} \frac{\lambda_l(t)}{N}\right) \cdot u_{N+1,k}(t) + n_k(t), \quad (9)$$

where $u_{l,k}(t)$ is expressed such as Eq.(10) using the out puts of the middle layer $y_j(t)$.

$$u_{l,k}(t) = \sum_{j=1}^{J} w_{l,kj} y_j(t) \quad , \quad (10)$$

where $w_{l,kj}(t)$ is the weight from jth node in middle layer to lth output constructing kth output u_k . $\lambda_l(t)$ in Eq.(9) is lth time-varying parameter to realize the adaptation fo r the periodical change of the environment.

$$\lambda_l(t) = \frac{1}{1 + \exp\{-\alpha \cdot \sin(s_l \pi t + \theta_l)\}} \quad , \quad (11)$$

where α is constant, s_l , θ_l are learning parameters. These parameters are adjusted to adapt the change of the environment.

4. LEARNING ALGORITHM

In the learning algorithm, learning is executed by Back Propagation(B.P.) In the critic the parameters are adjusted in order to reduce the error $\hat{r}(t)$ and in the actor the p arameters are adjusted to maximize the prediction value P (t). The renewal parts of the parameters are as follows,

$$\Delta w_{l,kj}^{actor}(t) = \eta \cdot \frac{\lambda_l(t-1)}{N} \cdot y_j(t-1) \cdot \hat{r}(t-1) \cdot n_k(t-1) \quad , \tag{12}$$

$$\Delta s_l = \eta \cdot \frac{u_{l,k}(t-1) - u_{N+1,k}(t-1)}{N} \cdot \frac{\alpha \cdot \exp\{-\alpha \cdot \sin(s_l \pi(t-1) + \theta_l)\} \cdot \cos(s_l \pi(t-1) + \theta_l) \cdot \pi(t-1)}{[1 + \exp\{-\alpha \cdot \sin(s_l \pi(t-1) + \theta_l)\}]^2} \cdot \hat{r}(t-1) \cdot n_k(t-1) \quad , \tag{13}$$

$$\Delta \theta_l = \eta \cdot \frac{u_{l,k}(t-1) - u_{N+1}(t-1)}{N} \cdot \frac{\alpha \cdot \exp\{-\alpha \cdot \sin(s_l \pi(t-1) + \theta_l)\} \cdot \cos(s_l \pi(t-1) + \theta_l)}{[1 + \exp\{-\alpha \cdot \sin(s_l \pi(t-1) + \theta_l)\}]^2} \cdot \hat{r}(t-1) \cdot n_k(t-1) \tag{14}$$

$$\Delta w_j^{critic}(t) = \eta \cdot y_j(t-1) \cdot \hat{r}(t-1) \quad , \tag{15}$$

where η is learning rate. Actor executes B.P. learning co nsidering $\hat{r} \cdot n_k$ as error signal. This is looking forward to convergence of the learning, as learning progresses, $\hat{r}(t)$ and $n_k(t)$ will be smaller.

5. COMPUTER SIMULATIONS

5.1 Simulation Environment

In this section performance of the proposed method is c ompared with that of the conventional method without ti me-varying parameters using simple maze problem with t he aisle, colored gray, which opens and closes periodicall y as shown in Fig.5..

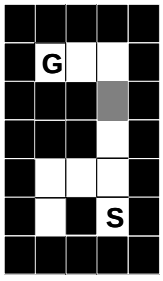

Fig.5 Periodic changeable maze environment

In Fig.5 S block means start point, G block means goal point, white blocks mean aisle, black blocks mean walls, and gray block means periodical open/close block. Here, when gray block closes, if agent is in the gray block, it keeps open. Agent moves after gray block opens/closes.

Time chart of the gray block open/close is shown in Fi g.6.

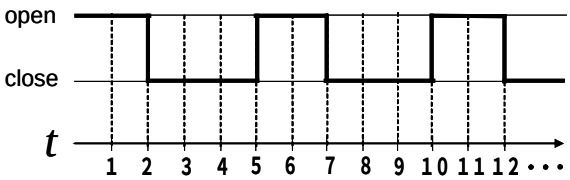

Fig.6 Aisle open/close time chart

5.1 Observations

Agent observes the status of left block(x_1), left-front (x_2), front(x_3), right-front(x_4), right(x_5) as shown in Fig.7. If observation block is wall, agent observes 1. If it is aisle, agent observes 0. Therefore the observation stat e has 5 dimensions. For example, in the case of Fig.7, a gent gets the observation state $x = [0,1,0,0,1]$. Agent mov es 1 block to the left block, or to the front, or to the ri ght, or to the back, keeping the front of the agent to up side. Therefore K, number of modules in Fig.4, is equal to 4.

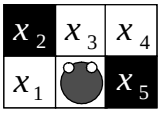

Fig.7 Observation of the agent

[**Simulation 1 : the case of one time-varying paramet er**]

In this simulation, the used values of the parameters ar e that $\eta = 0.2, T = 0.4, \alpha = 1.0$. $N = 1$ in Eq.(9). The li mit time step in 1 learning trial is 50. We call 1 learnin g trial when agent gets goal or doesn't get goal within li mit time step. 30000 learning trials are executed. We call these learning trials 1 trial. The rewards is 1 when agen t gets the goal block, the rewards is –1 when agent colli des the wall blocks and agent doesn't get the goal within limit step, the rewards is 0 when other cases are. The a verage of 100 trials are shown in Fig.8.

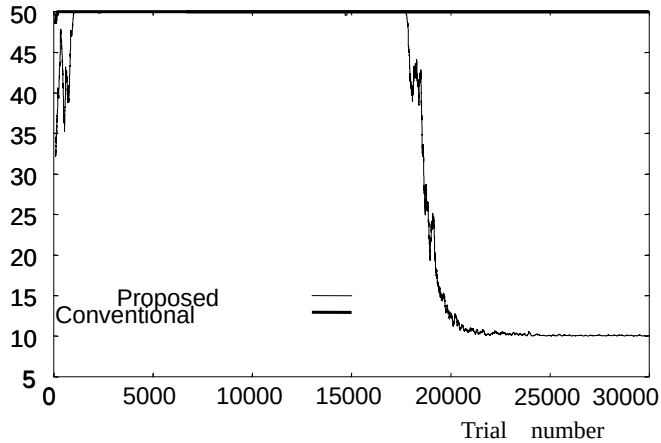

Fig.8 Average step numbers to get goal in the case of o ne time-varying parameter

The results shows that the proposed method succeeded in within about 17000 average trial numbers, but the conventional didn't succeed in within 30000 trials. The example of the success of the trial with ten steps of the proposed method is shown in Fig.9.

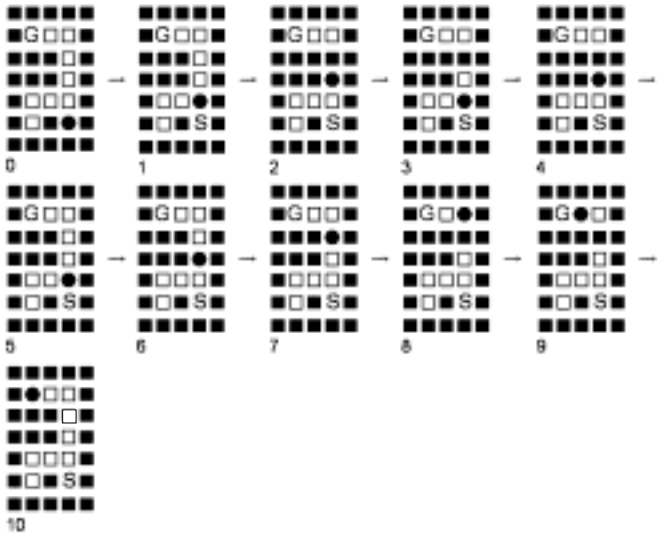

Fig.9 Agent movement from start to goal with 10 steps

From Fig.9 it is found that after aisle opened agent got into the gray aisle at 7 step and into the goal at 10 step. Transition of values of $\lambda, 1-\lambda$ are shown in Fig.10.

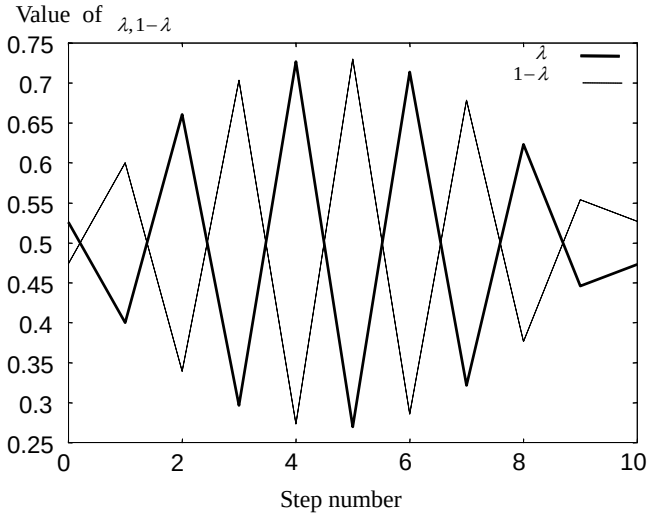

Fig.10 Transition of values of time-varying parameter λ

[**Simulation 2 : the case of two time-varying parameters**]

In this simulation, the used values of the parameters are that $\eta = 0.2, T = 0.4, \gamma = 0.5, \alpha = 1.0$. $N = 2$ in Eq.(9). The limit time step in 1 trial is 50. 10000 trials are executed. The results of 100 learning trials are shown in Fig.11. In comparing the performance of one time-varying parameter case: Fig.9 with that of two time-varying parameters case: Fig.11,

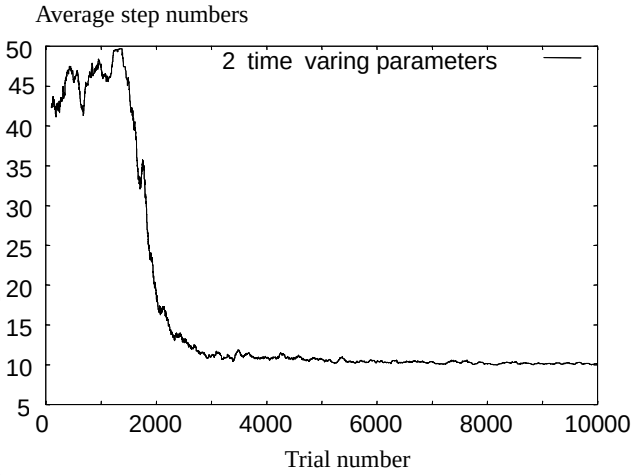

Fig.11 Average step numbers to get goal in case of two Time-varying parameters

the performance of the Fig.11 case is much better than that of the Fig.9 case. The Fig.11 case converges in about 10 % of trials of the Fig.9 case. Transition of values of two time-varying parameters are shown in Fig.12.

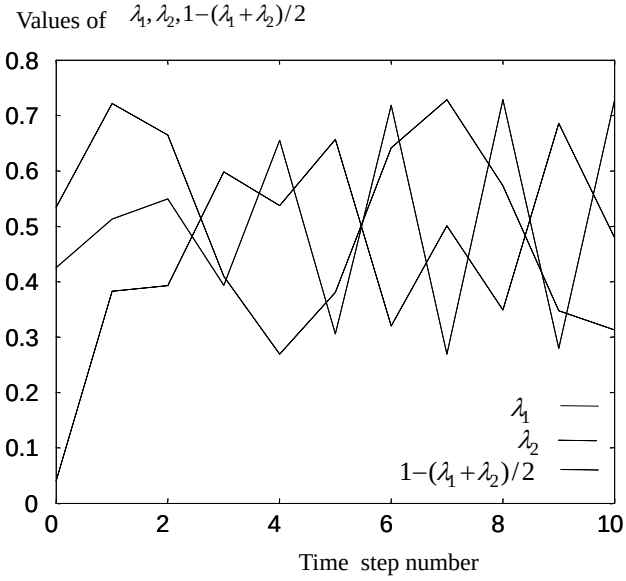

Fig.12 Transition of values of two time-varying parameters λ_1, λ_2

6. CONCLUSIONS

We proposed the reinforcement learning system, which is able to adapt the changeable environment introducing the time varying parameters to the actor and showed its effectiveness. As the future work improvement of the speed of learning is important to apply our method for the real system and for online use.

REFERENCES

[1] R.S.Sutton and A.G.Barto, "Reinforcement Learning, An Introduction " *MIT Press*,1998

[2] J.H.Lee and S.Y. Oh and D.H. Choi, "TD based Reinforcement Learning Using Neural Networks in Control Problems with Continuous," *Proc. of IJCNN*, pp. 2028-2033, 1998