Couple Particles in Action Space for Reinforcement Learning

Akira Notsu, Katsuhiro Honda and Hidetomo Ichihashi[†],

Osaka Prefecture University, 1-1 Gakuencho, Nakaku, Sakai, Osaka 599-8531, JAPAN

Summary
In this paper, we propose a novel action-search particle-filtering algorithm for learning processes. This algorithm is designed to perform search domain reduction and heuristic space segmentation. In this method, each action space is divided into new two segments using two particles. Appropriate search domain reduction can minimize learning time and enable the recognition of the evolutionary process of learning. In a numerical experiment, the proposed filtering method is applied to a single pendulum simulation in order to demonstrate the adaptability of this simulation model.

Key words:
Reinforcement Learning, TD-Learning, Particle Filter, Single Pendulum Simulation.

1. Introduction

This paper proposes a novel particle-filtering algorithm for reinforcement learning that periodically divides the agent environment and action search space during the learning process. The standard reinforcement learning algorithm [1], [2] requires well-suited segmentation before learning. In the proposed method, each state is searched by the couple particle method, and distance between states are fixed according to the importance of each state. Herein, human-like space segmentation in actor–critic learning, which we investigated previously [3], [4], is further refined.
In post-supervised design, clustering, which is done after many trial and error iterations, is a probabilistic method. A genetic algorithm representative of these schemes is called a learning classifier system [5], [6]. In this system, inputted data are encoded as binary rules on which a genetic algorithm alters and selects the best rules. However, inputted data can be encoded during the learning, as in previously investigated method. Our proposed algorithm is applicable not only for reinforcement learning but also for social simulation, which requires each agent to behave rationally during the trial and error iterations. Such an approach provides clues for understanding culture and language.

2. Reinforcement learning

Reinforcement learning assumes that agents can visit only a finite number of states, and that when visiting a state, they will collect a numerical reward (which can be interpreted as a punishment in the case of a negative reward [7]). Each state has a different value associated with it. Given an initial state, subsequent states are reached according to actions. The value of a given state is defined as the average reward that can be collected by selecting actions available in that state.
The actor–critic method includes a critic, which maintains the state value estimate V, and the actor, which is responsible for choosing the appropriate action at each state.
Value-function approaches attempt to find a policy that maximizes the return for taken actions by maintaining a set of estimates of expected rewards based on a policy π. In this method, an agent attempts to estimate the expected reward

$$V(s) = E[R|s,t,\pi] = Rt + \gamma E[R|s,t+1,\pi], \quad (1)$$

where an agent following policy π takes an action in state s.
In the present paper, we use a normal distribution for the stochastic policy π. Updating stochastic policy is executed by adjusting output mean a, rate α and variance v

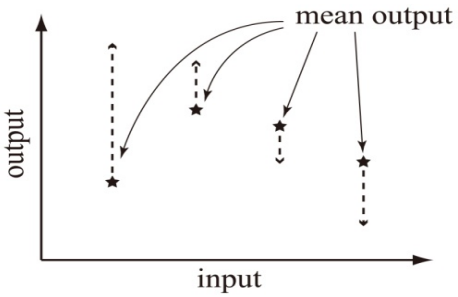


Fig. 1 Actor–Critic method.

Fig. 2 Updating the output by TD-error.

Manuscript received December 5, 2010
Manuscript revised December 20, 2010

3. Couple particles in each action space

We propose couple particles in the action space. The proposed method using couple particles can search for the global solution (best output) and a local solution, which has the potential of being the global solution.
In each state, two particles have expected rewards according to $V(s)$. If inputted data have a positive TD-error, the nearer particle is updated (shown in fig. 3). This method decreases the probability of the local solution.

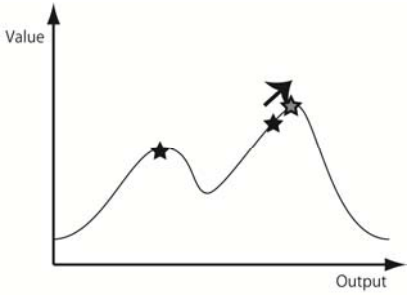


Fig. 3 Updating the output by TD-error.

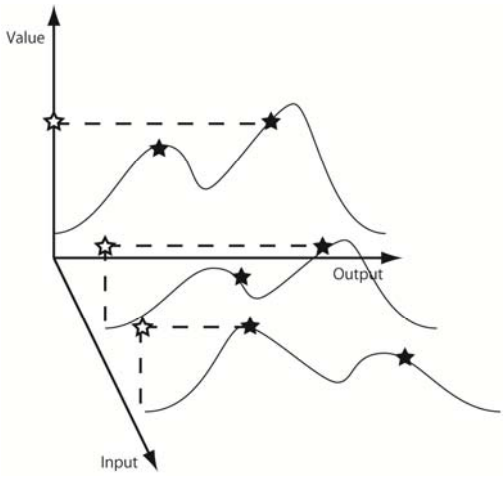


Fig. 4 Couple particles in the space.

In addition, because an evenly segmented state is not always good for learning, we also propose adjusting the distance between state prototypes by setting the distance between successive best particles in the output–input space (in fig. 5). This is based on the fact that if the best output is quite different, near states should be searched closely, and otherwise do not need to be. In the process, the updated particle entrains the state particle.
Therefore, our proposed method requires that the input and output spaces be normalized.

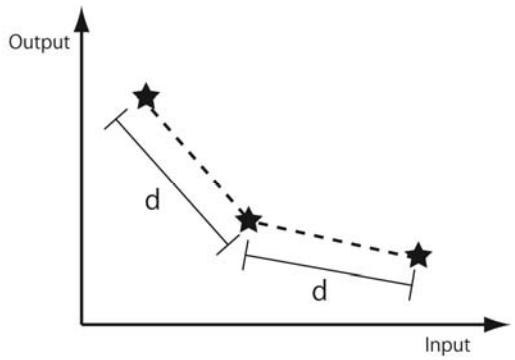


Fig. 5 Updating the input space distribution by distance

4. Single pendulum simulation

We simulated the single pendulum standing task (fig. 6). This physical calculation is performed by Runge–Kutta method.

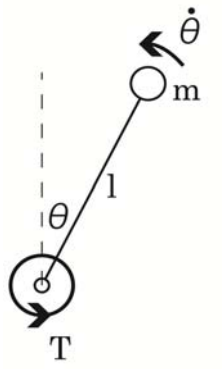


Fig. 6 Single pendulum

The following motion equation is simulated in increments of 0.05 seconds:

$$m\ddot{\theta} = -mgl\sin(\pi - \theta) + T + (-k\dot{\theta})$$

where the parameters and variables are define as

- ✓ m = 1 kg
- ✓ l = 1 m
- ✓ k = 0
- ✓ g = 9.8 m/s^2
- ✓ $-\pi < \theta \leq \pi$
- ✓ $-3\pi < \dot{\theta} \leq 3\pi$
- ✓ $-20 < T \leq 20$
- ✓ 1 episode = 5 seconds, 1 learning cycle = 0.2 seconds
- ✓ $\theta = \pi, \dot{\theta} = \ddot{\theta} = 0$, in the initial state
- ✓ $R = \frac{(\pi - |\theta|)}{\pi} - T/40$
- ✓ $\alpha = 0.1, \gamma = 0.9$
- ✓ variance $V_0 = \frac{\pi}{2}$, $V_{1000} = V_0 \cdot \beta^{1000} = 0.0001 \cdot V_0$ for constant β.

Figures 7 to 9 show the results of several simulations. The vertical axes in these figures show reward and the horizontal axes show episode times. In each figure, the left panel shows the 100-trial mean of the reward and the right panel shows the best reward. The shown results are for the normal method (blue), our proposed method in which couple particles are applied (green), and our proposed method along with adjusting the distance between state prototypes (red). The agent has 100 states in the learning process.

In the simulations shown in fig. 7, each state is segmented evenly. As shown, the proposed method is not always good for learning; however, it does not inhibit the learning process.

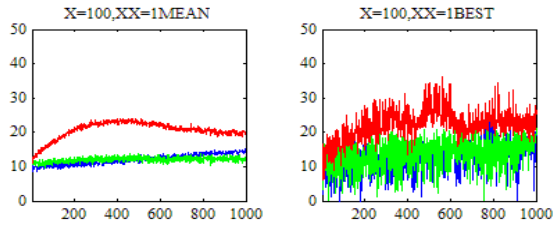


Fig. 7 Results in better segmentation by the particles

In figs. 8 (over segmented θ) and 9 (over segmented δ), the state is not segmented appropriately. As shown, the proposed method leads to be better reward history in terms of speed and value.

Our simple segmentation method, especially in the state space, worked very well. Further adjustments in the parameters of the proposed method have the potential of further improving the results.

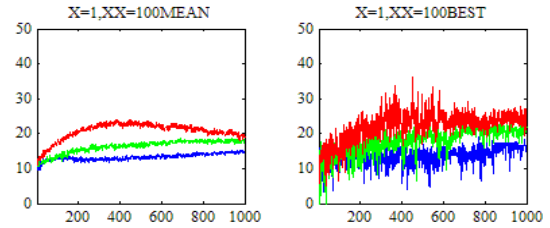


Fig. 8 Results in worse segmentation by the particles (over segmented θ)

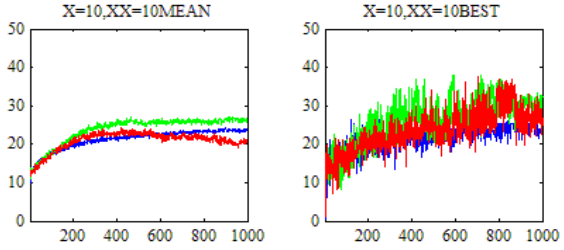


Fig. 9 Results in worse segmentation by the particles (over segmented δ)

5. Conclusion

We proposed a novel particle filtering approach to create a novel reinforcement learning algorithm that segments the agent's environment and action. The proposed filtering method is applied to a single pendulum simulation in order to demonstrate the adaptation ability of the model.

Future work

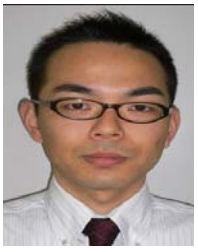
We are implementing this method for a double pendulum simulation.

Acknowledgments

This work was supported in part by the Ministry of Education, Culture, Science and Technology, Japan under Grant-in-Aid for Scientific Research No. 21700242 and funded by the Ministry of Internal Affairs and Communications, Strategic Information and Communications R&D Promotion Programme.

References

[1] T. Jaakkola, M. Jordan, and S. P. Singh, On the Convergence of Stochastic Iterative Dynamic Programming Algorithms, Neural Computation, vol. 6, pp. 341--362, 1992.
[2] C. J. C. H. Watkins, and P. Dayan, Technical Note: Q-Learning, Machine Learning, vol. 8, pp. 56--68, 1992.
[3] A. Notsu, H. Ichihashi, K. Honda, State and Action Space Segmentation Algorithm in Q-learning, Proc. of 2008 International Joint Conference on Neural Networks, 2385-2390, 2008.
[4] A. Notsu, H. Wada, K. Honda, H. Ichihashi, Cell Division Approach for Search Space in Reinforcement Learning, International Journal of Computer Science and Network Security , 8, 6, 18-21, 2008.
[5] J. H. Holland, and J. Reitman, Cognitive Systems Based on Adaptive Algorithms, in D. A. Waterman and F. Hayes-Roth (Eds.), Pattern Directed Inference Systems, Academic Press, 1978.
[6] J. H. Holland, K. J. Holyoak, R. E. Nisbett, and P. R. Thagard, Induction, The MIT Press, 1986.
[7] R. S. Sutton, Learning to Predict by Method of Temporal Differences, Machine Learning, vol. 3, 1, pp. 9-44, 1988.



Akira Notsu received the B.E., M.I. and D. Informatics degrees from Kyoto University in 2000, 2002, and 2005, respectively.
He is currently an Assistant Professor, Department of Computer Science and Intelligent Systems, Osaka Prefecture University. His research interests include agent-based social simulation, communication networks, game theory, human–machine interface, and cognitive engineering.

Katsuhiro Honda (M'01) received the B.E., M.E. and D.Eng. degrees in industrial engineering from Osaka Prefecture University, Osaka, Japan, in 1997, 1999, and 2004, respectively.
He is currently an Assistant Professor, Department of Computer Science and Intelligent Systems, Osaka Prefecture University. His research interests include hybrid techniques of fuzzy clustering and multivariate analysis, data mining with fuzzy data analysis, and neural networks. He received a paper award and a young investigator award from Japan Society for Fuzzy Theory and Intelligent Informatics (SOFT) in 2002 and 2005, respectively, and gave a tutorial titled "Introduction to Clustering Techniques" at the 2004 IEEE Int. Conf. Fuzzy Systems (FUZZ-IEEE 2004).

Hidetomo Ichihashi (M'94) received the B.E. and D.Eng. degrees in industrial engineering from Osaka Prefecture University, Osaka, Japan, in 1971 and 1986, respectively.
From 1971 to 1981, he was with the Information System Center of Matsushita Electric Industrial Co., Ltd., Tokyo, Japan. From 1981 to 1993, he was a Research Associate, Assistant Professor, and Associate Professor at Osaka Prefecture University, where he is currently a Professor in the Department of Computer Sciences and Intelligent Systems. His fields of interest are adaptive modeling of GMDH-type neural networks, fuzzy C -means clustering and classifier, data mining with fuzzy data analysis, human–machine interface, and cognitive engineering.