Eighth International Conference on Intelligent Systems Design and Applications

Effective integration of imitation learning and reinforcement learning by generating internal reward

Keita Hamahata, Tadahiro Taniguchi, Kazutoshi Sakakibara, Ikuko Nishikawa
Ritsumeikan University
Noji Higashi 1-1-1 Kusatsu, Shiga, JAPAN
{hamahata, taniguchi, sakaki, nishi}@sys.ci.ritsumei.ac.jp

Kazuma Tabuchi,* Tetsuo Sawaragi
Kyoto University
Yoshida honmachi, Sakyo-ku, Kyoto Kyoto, Japan
sawaragi@me.kyoto-u.ac.jp

Abstract

This paper describes an integrative machine learning architecture of imitation learning and reinforcement learning. The learning architecture aims to help integration of the two learning process by generating internal rewards. After observing superiors, human learners usually start practicing through trial and error. Humans usually learn tasks through both imitation learning and reinforcement learning. Imitation learning and reinforcement learning should be harmonized as an effective and integrative learning system. A simple reinforcement learning requires a huge amount of trials and errors in an agent's learning phase. However, imitation learning can reduce the amount of time. Based on this idea, the composition of reinforcement learning and imitation learning is proposed as an integrative machine learning architecture. In this paper, an additional internal reward system, which is generated by the learner agent, is introduced to achieve this goal. The learning architecture is evaluated through an experiment and the effectiveness of the integration is examined.

1. Introduction

Reinforcement learning and imitation learning are both important learning capability of human beings. They are usually studied separately as different learning architectures in the context of robotic machine learning researches. However, the two learning architectures have to work collaboratively when people practice playing sports, acquire engineering skills, and play music instruments in our daily life. When we try a new task, we usually observe other person's action and imitate it to perform better. However, we also know that an imitation learning consisting of a simple observation cannot give us the sophisticated skill. This difficulty owes to several shortcomings of imitation learning. First, the observed agent's behavior is not always optimal. Therefore, the behavior obtained by a learner is not guaranteed to be optimal even if the learner imitates the target behavior completely. Second, an imitator cannot observe all of the imitatee's[1] state variables. It means that the imitator cannot obtain all information about the demonstrator's skill. Third, the physical dynamics of an imitator and an imitatee are usually different. If the imitatee has different body from the imitator, the physical dynamics are different. In such a case, the task and the optimal behavior become different for each agent. Fourth, an imitator usually cannot require an imitatee to exhibit enough amount of demonstration. An imitator has to utilize the limited samples to improve his/her behavior. These problems prevent learning agents from acquiring sufficiently sophisticated skills only through imitation learning scheme. Therefore, reinforcement learning, which bases on trial-and-error, is indispensable in acquisition of physical skills. However, the reinforcement learning also has critical problems. Reinforcement learning utilizes only a scalar performance index, i.e. reward, as feedback information. Therefore, it takes much time for an agent to acquire a satisfactory behavior. To overcome this problem, a large number of researches have been made. However, the fact that the learning speed of reinforcement learning is fundamentally slower than supervised learning cannot be denied because it can use less information than direct supervised learning. In contrast, by utilizing reinforcement

[*] Currently, he is working at Yasukawa electric engineering co.

[1] An imitatee is a person who is imitated by a imitator.

978-0-7695-3382-7/08 $25.00 © 2008 IEEE
DOI 10.1109/ISDA.2008.325

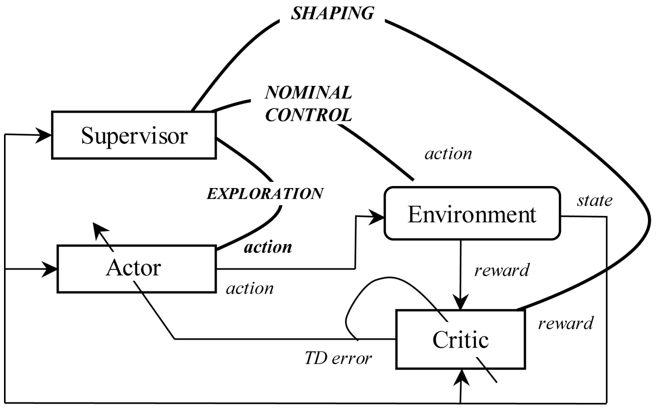

Figure 1. Three pathways for supervisor information [1]

learning, an agent can acquire an optimal behavior, which is suitable to its body through trial-and-errors. These facts tell us that reinforcement learning and imitation learning can back up each other.

In this paper, the integrative learning architecture of imitation learning and reinforcement learning is investigated. We describe the two approaches towards the integration of reinforcement learning and imitation learning. The two approaches and their integrative approach are evaluated based on a simple simulation experiment.

2. Algorithms

Barto et al.[1] addressed that there are three pathways through which a learner can utilize a supervisor's information in reinforcement learning. They are "shaping", "nominal control" and "exploration"(Figure 1).

In nominal control and exploration, a supervisor directly exhibits its motor output to the learner. In their research, they focused on the latter two pathways, and proposed the supervised actor-critic architecture. In contrast, we focus on the "shaping" and "supervised learning" in this paper. Along the "shaping" pathway an additional source of evaluative feedback, i.e., additional reward, is supplied. An imitator itself who observes the supervisor's action can generate the additional reward, and learn a skill through reinforcement learning by referring to the additional reward in addition to the original reward. On the contrary, "supervised learning" is not based on reinforcement learning. An amateur gets to play better by mimicking the supervisor's behavior. Based on supervised learning, an imitator can obtain a certain controller before it starts reinforcement-learning task.

2.1 Reinforcement learning

We assume that the target system is a nonlinear dynamical system, which can be described as

$$x_{t+1} \quad = \quad F(x_t, u_t) + n_t \tag{1}$$

where n_t is a noise term whose probability distribution is $N(0, \sigma^2)$. Actor-critic reinforcement learning architecture has a policy function π (i.e., actor) and a state value function V (i.e., critic), independently. The b is a column vector of basis functions. The b in eqs. (2), (3), and (19) do not have to share the same basis functions. Here, we use the same notation only for the simplicity.

$$V(x) \quad = \quad w^T b(x) \tag{2}$$

$$\mu(x) \quad = \quad \Theta b(x) \tag{3}$$

$$\pi(x, u) \quad = \quad P(u|x) \approx N(\mu(x), \Sigma) \tag{4}$$

where $\mu(x)$ is corresponding to a controller of an agent, and Θ is its parameter matrix. After the agent observes a state x_t and outputs action command u_t, it obtains a reward r_t and observes the next state x_{t+1}. In such case, TD error δ_t is calculated as followings.

$$\delta_t \quad = \quad r_t + \gamma V(x_{t+1}) - V(x_t) \tag{5}$$

Kimura et al.[3] introduced eligibility trace to the actor critic reinforcement learning architecture. Based on the formulation, actor's eligibility trace becomes

$$e_\theta \leftarrow e_\theta + \lambda \frac{\partial}{\partial \theta} ln(\pi(x, u)). \tag{6}$$

By using the eligibility trace the updating rule is described as

$$\triangle \theta \quad \propto \quad \delta_t e_\theta \tag{7}$$

$$\triangle w \quad \propto \quad \delta_t b(x_t) \tag{8}$$

where θ is a parameter of the actor.

2.2 Imitation learning

We assume that a learner observes the series of demonstrator's motions $\{x_t^d\}$. In a occupational sense, "imitation learning" means that a learner utilizes the time series x_t^d and sophisticates his/her controller. In this paper, we introduce two types of methods utilizing the observed time series.

2.2.1 Supervised learning

First, an imitator can directly use the time series of demonstrator's motion for a supervised learning. This is often simply called "imitation learning". However, it's only a model of imitation learning.

Even if an imitator observes $\{x_t^d\}$, an agent doesn't know the imitatee's true action commands $\{u_t^d\}$ without any additional information. The action command is usually hidden states. However, if an imitator has obtained an inverse model $\hat{I}$ of the target system, the imitator can estimate the motor commands.

$$\hat{u}_t^d = \hat{I}(x_t^d, x_{t+1}^d) \tag{9}$$

In this paper, we do not take care of the acquisition of the inverse model. We assume that the agent can obtain the estimated time series of action command $\{\hat{u}_t^d\}$. After obtaining such time series, an agent can acquire the optimal controller which minimizes squared error between the estimated supervisor's motor command and the imitator's motor command based on the least square method or the ridge regression.

$$\hat{U} = [\hat{u}_1^d, \hat{u}_2^d, \cdots, \hat{u}_T^d] \tag{10}$$

$$X = [x_1^d, x_2^d, \cdots, x_T^d] \tag{11}$$

$$E_{ridge} = \sum_{t=1}^{T} ||\hat{u}_t^d - \Theta b(x_t)||^2 + \epsilon \, \mathrm{tr}(\Theta^T \Theta) \tag{12}$$

The eq. 12 is the objective function of the ridge regression. By minimizing the E_{ridge}, the optimal Θ is acquired as follows.

$$\Theta = \hat{U} X^T (XX^T + \epsilon I)^{-1} \tag{13}$$

where I is an identity matrix. Supervised learning method is the simplest way of expressing imitation learning. This operation is corresponding to set an initial parameter of a policy in reinforcement learning.

2.2.2 Shaping rewards

The second way of imitation learning is to shape its reward function based on the observation of supervisor's behaviors. This is more implicit utilization of a supervisor's behaviors. To utilize the observed supervisor's time series, the imitator has to create additional reward function by observing the supervisor's behavior. This is the next question.

If a reward function changes in a reinforcement-learning task, the optimal policy usually changes because a purpose of the task changes. However, Ng and Harada et al.[5] found a class of additional rewards, which keep the optimal policy invariant. If additional reward r^{sub} is designed as follows, the optimal policy which the learner should acquire does not change.

$$r_{t+1}^{sub} = \gamma \Phi(x_{t+1}) - \Phi(x_t) \tag{14}$$

where Φ is a potential function for the additonal reward. A new modulated reward $\bar{r}_t$ is a sum of the original reward r_t and an additional internal reward r_t^{sub},

$$\bar{r_t} = r_t + r_t^{sub}. \tag{15}$$

The potential function Φ uniquely determines the additional reward r^{sub}. Therefore, the design of the potential function is the next problem. Ng and Harada et al.[5]suggests that the potential function gives the most effective shaping when $\Phi = V^*$ where V^* is the optimal value function of the original. Marthi proposed the method of automatic shaping[4].

Therefore, we additionally think of a learning agent to estimate the imitatee's value function by assuming the observed behavior is produced according to the optimal policy. A state value function is

$$V(x_t) = E[\sum_{k=0}^{\infty} \gamma^k r_{t+1+k}] \tag{16}$$

by definition. Therefore, we calculate the estimated state-value function $\hat{V}$ for each state x_t^d. When the observed time series of the imitatee's states are $\{x_0, x_1, \cdots, x_T\}$ and the reward on each state is $\{r_0, r_1, \cdots, r_T\}$, a sample of $\hat{V}$ at time t is calculated to be

$$\hat{V}_t = \sum_{k=0}^{\infty} \gamma^k r_{t+1+k}. \tag{17}$$

After obtaining the state-value pair $(x_t, \hat{V}_t)$ for each time step, a potential function is calculated by minimizing

$$E = \sum_k \sum_t ||\hat{V}_t^k - \Phi(x_t^k)||^2. \tag{18}$$

$$\Phi = \phi^T b(x) \tag{19}$$

The superscript k represents the index of the observed time series. The Φ is also determined by using ridge regression.

By shaping such an internal reward function, the learning agent is expected to accelerate its learning speed in the reinforcement learning task.

3 Experiment

We made a simple experiment to compare several schemes of integration of imitation learning and reinforcement learning.

3.1 Conditions

An agent is placed on a 2D toroidal field (Figure 2). The agent is put on $(x, y) = (-1, -1) + \boldsymbol{n}$ at time $t = 0$ in every trial. $\boldsymbol{n}$ is a gaussian noise whose standard deviation is 0.01. The state vector and action command are defined as $\boldsymbol{x} = (x, y, \dot{x}, \dot{y})$ and $\boldsymbol{u} = (f_x, f_y)$. The physical dynamics is simply described as follows.

$$\ddot{x} = f_x \tag{20}$$

$$\ddot{y} = f_y \tag{21}$$

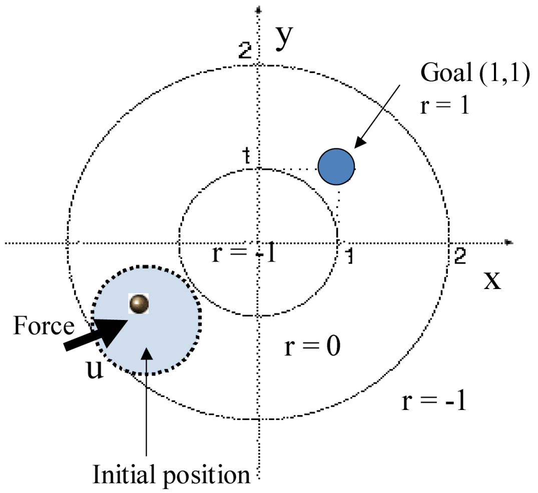

Figure 2. Environment of an experiment

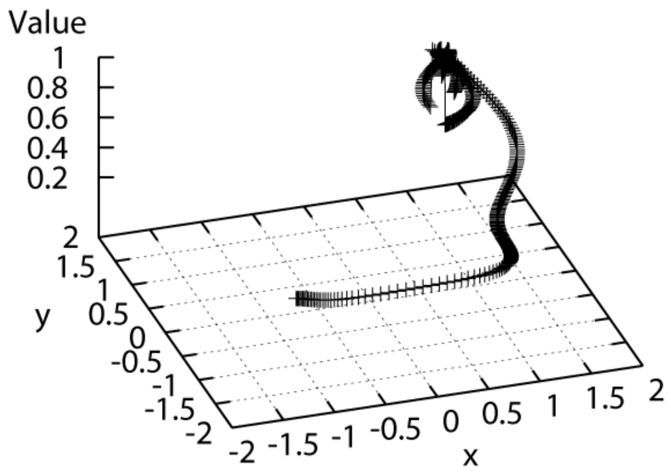

Figure 3. Observed movement of the demonstrater and estimated values on the path

To simulate this system, we employed Euler method. The simulation time step is $0.05[s]$. An agent has to reach the goal whose center is at $(x, y) = (1, 1)$ after leaving around $(x, y) = (-1, -1)$. When the agent arrives within 0.2 of the goal, the agent obtains $r = 1$ and the task is accomplished. If the agent gets into the inner circle whose radius is 1 or the agent goes out of the outer circle whose radius is 2, the task is ended up and a penalty $r = -1$ is given to the agent.

We employed a normalized gaussian network (NGnet) as basis functions of actor and critic in this experiment[9] (eq. 2, 3). $(3, 12, 5, 5)$ basis functions are prepared for radial axis, circumferential direction, $\dot{x}$ axis and $\dot{y}$ axis, i.e., 900 basis functions are distributed.

We gave only one demonstration as a target of imitation to the learning agent. The movement and estimated value function is shown in Figure 3. The value function is estimated by utilizing eq. 23. In supervised learning (section 2.2.1), motor commands u are estimated based on the time series of the demonstrated movement and an inverse model, and the demonstrator's controller is estimated and approximately acquired by the imitator. In reward shaping (section 2.2.2), a value v of each state on the movement is estimated and the demonstrator's value function is partially estimated. Based on the estimated value function, the learner constructs a potential function Φ, internally. However, the number of states that can be observed from a demonstrator's movement is too small to construct a fine value function. Therefore, the degree of generalization is very important in the formation of the potential function. To achieve the adequate generalization, we utilized two methods, i.e., vector quantization using k-means and NGnet with Cauchy distribution functions. After the acquisition of the time series of displayed movement$\{x_t\}$, K-means is applied to the time series. After the quantization, the centers of the normalized gaussian functions are allocated on the centers of the K clusters. If we use the normalized gaussian network to estimate the imitatee's value function, the samples are generalized too much. Therefore, we use normalized gaussian network Cauchy distribution functions whose basis function b_i^{Cauchy} is described as follows.

$$b_i^{Cauchy}(x) = \frac{1}{1 + C\frac{1}{\beta^2}(x - m_i)^2} n_i(x) \tag{22}$$

$$n_i(x) = G(x; m_i, \Sigma_i) / \sum_j G(x; m_i, \Sigma_i) \tag{23}$$

where n_i is a basis function of an NGnet, $G(x; m, \Sigma)$ is a gaussian function whose center is m and variance-covariance matrix is Σ. C and β are parameters of the basis function. In this experiment, the standard deviation of each cluster is used for β.

Potential function is acquired as shown in Figure 4. Figure 4 shows the potential function's section of $(\dot{x}, \dot{y}) = (0, 0)$.

We experimented 4 learning schemes. One is simple actor-critic architecture, the second is an actor-critic with reward shaping, the third is an actor-critic with supervised learning, and the other is an actor-critic learning architecture with both reward shaping and supervised learning. We call the last one "'integrative learning architecture'.

One trial continues until the agent goes outside, it reaches the goal, or 20 seconds pass.

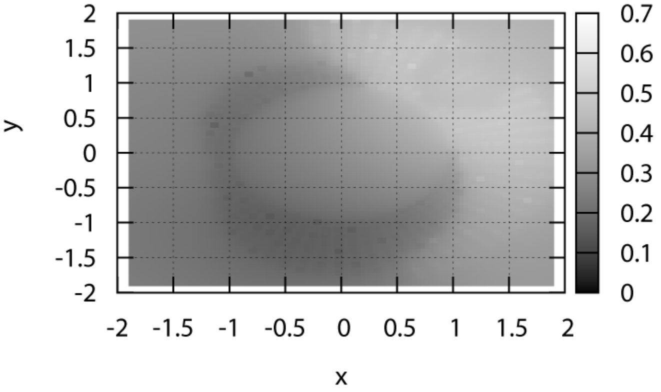

Figure 4. Obtained potential function for reward shaping at a section of $(\dot{x}, \dot{y}) = (0, 0)$

Table 1. Parameters of the Problems

Learning rate α	0.2
Discout factor γ	0.99
Eligibility trace discount factor λ	0.9
Parameter of Cauthy-like function C	1/12

3.2 Result

10 sessions are excuted for each learning scheme. A session consists of 3000 trials. Figure 5 shows the averaged reward for each 100 trials. Here, only the original reward is counted. Therefore, the four schemes are compared under the same condition. This shows that the schemes that utilize the information of demonstrator's movement are superior to the simple actor critic. During first 500 trials, the two schemes, which utilized supervised learning before starting reinforcement learning, is superior to others. However, this is only to be expected. Supervised learning is corresponding to designing an adequate initial parameter to the policy function. Therefore, the initial performance is guaranteed to some extent. However, we can hardly say if it makes the learning process better after starting the reinforcement learning process. On the other hand, the schemes with reward shaping smoothly grew up. Finally, the performance of reward shaping scheme was almost equivalant to that of the supervised learning scheme.

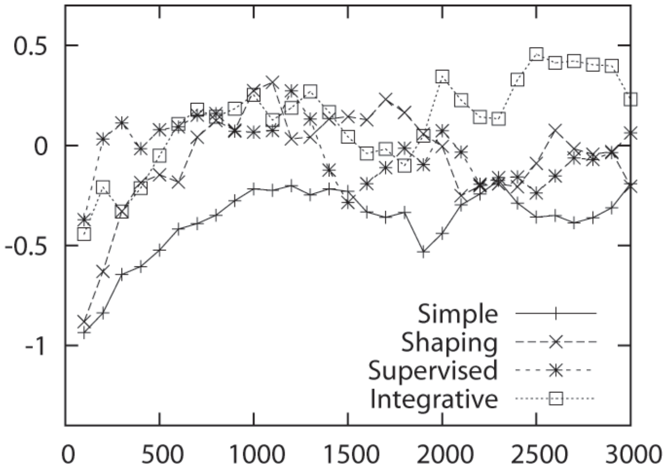

Figure 5. Transition of averaged reward for each learning scheme

Figure 6 shows the averaged original reward through out all the sessions and the trials. This shows that reward shaping and supervised learning schemes have almost same contribution to accelerate reinforcement learning. The integrative learning scheme, which utilizes the both schemes, is slightly better than the two schemes. As we described, supervised learning scheme and reward shaping scheme utilize almost same information about demonstrated time series $\{x_t\}$. Therefore, much information, which can be utilized to acceralate the learning architecture, overlapped each other.

4 Conclusion

In this paper, we described brief investigation about integration of reinforcement learning and imitation learning. Especially, we introduced a method which enables an learning agent to generate additional internal reward function by observing other's movement. A simulation experiment showed that the proposed shaping method accelerates the learning speed.

However, theoretical basis and the effectiveness of the proposed method is not clear enough. In our future works, we have to evaluate the proposed method with more experiments and examine the effectiveness from the mathematical viewpoint.

References

[1] A. Barto and M. Rosentein. Supervised Actor-Critic Reinforcement Learning. *Handbook of Learning and Approximate Dynamic Programming, J. Si, AG Barto, WB Powell, and D. Wunsch (eds.)*, pages 359–380.

[2] T. Inamura, I. Toshima, H. Tanie, and Y. Nakamura. Embodied symbol emergence based on mimesis theory. *International Journal of Robotics Research*, 23(4):363–377, 2004.

[3] H. KIMURA and S. KOBAYASHI. An Analysis of Actor-Critic Algorithms Using Eligibility Traces. Reinforcement Learning with Imperfect Value Functions. *Journal of*

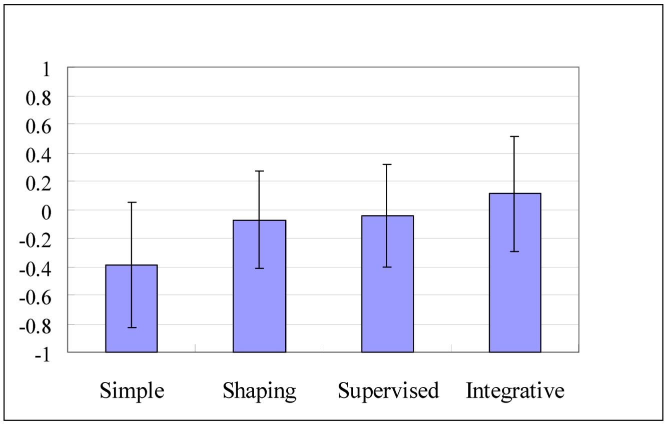

Figure 6. Averaged reward obtained in a traial for each learning schemes

Japanese Society for Artificial Intelligence, 15(2):267–275, 2000.

[4] B. Marthi. Automatic shaping and decomposition of reward functions. *Proceedings of the 24th international conference on Machine learning*, pages 601–608, 2007.

[5] A. Ng, D. Harada, and S. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 278–287, 1999.

[6] R. Sutton and A. Barto. Reinforcement Learning: An Introduction. 1998.

[7] R. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems*, 12(22), 2000.

[8] K. Tabuchi, T. Taniguchi, and T. Sawaragi. Efficient acquisition of behaviors by harmonizing reinforcement learning with imitation learning. In *The 20th Annual Conference of the Japanese Society for Artificial Inteligence*, 2006. (in Japanese).

[9] T. Taniguchi and T. Sawaragi. Incremental acquisition of multiple nonlinear forward models based on differentiation process of schema model. *Neural Networks*, 21(1):13–27, 2008.

[10] M. Tomasello. *The Cultural Origins of Human Cognition*. Harvard Univ Pr, reprint edition, 3 2001.