Artif Life Robotics (2009) 13:526–530
DOI 10.1007/s10015-008-0608-3

ORIGINAL ARTICLE

M. Obayashi · N. Nakahara · T. Kuremoto · K. Kobayashi

A robust reinforcement learning using the concept of sliding mode control

Received and accepted: July 28, 2008

Abstract In this article, we propose a new control method using reinforcement learning (RL) with the concept of sliding mode control (SMC). Some remarkable characteristics of the SMC method are good robustness and stability for deviations from control conditions. On the other hand, RL may be applicable to complex systems that are difficult to model. However, applying reinforcement learning to a real system has a serious problem, i.e., many trials are required for learning. We intend to develop a new control method with good characteristics for both these methods. To realize it, we employ the actor–critic method, a kind of RL, to unite with the SMC. We are able to verify the effectiveness of the proposed control method through a computer simulation of inverted pendulum control without the use of inverted pendulum dynamics. In particular, it is shown that the proposed method enables the RL to learn in fewer trials than the reinforcement learning method.

Key words Robust · Reinforcement learning · Actor–critic · Sliding mode control · Inverted pendulum

1 Introduction

Recently, a nonlinear system control method has been developed which unites conventional model-based control theory (sliding mode control, H_∞ control, etc.) and model-free control theory (neurocontrol, fuzzy control, reinforcement learning control, etc.), and is very vigorous. However, very little research has combined reinforcement learning with model-based control theory.

One of the few pieces of research is about robust reinforcement learning control[1] which unites reinforcement learning and H_∞ control. Based on the theory of H_infinity, Morimoto and Doya[1] considered a differential game in which a disturbance agent tries to make the worst possible disturbance while a control agent tries to make the best control input. They formulated the problem as finding a min–max of a value function that takes into account the amount of the reward and the norm of the disturbance, and developed a robust reinforcement learning method for both model-based and model-free systems. Reinforcement learning is a method that requires the control input sequences to achieve the control object based on the states of the system, and gives rewards through trial and error inputs to the system. Unnecessary information on the system is very advantageous. On the other hand, a great many trial and error attempts are needed to find the appropriate control input sequences to achieve the control object. In an extreme case, this might possibly destroy the controlled system, and therefore the object should be achieved with as little trial and error as possible.

We used a sliding mode control method to achieve the desirable dynamics of the system by restricting the states of the system on the switching surface of the state space of the system. This is known as a method with superior robustness.

In this article, we consider the model-free case. Under such conditions, we propose a model-free control method combining the concepts of sliding mode control and reinforcement learning. We were able to verify through the computer simulation of an inverted pendulum control problem that the proposed method makes it possible to achieve the control object with fewer trials than conventional reinforcement learning, and also gives a robust performance when the parameters of the system change.

2 Reinforcement learning

2.1 Actor–critic reinforcement learning

The actor–critic method is one of the representative reinforcement learning methods. The actor–critic method

M. Obayashi (✉) · N. Nakahara · T. Kuremoto · K. Kobayashi
Graduate School of Science and Engineering, Yamaguchi University, 2-16-1 Tokiwadai, Ube, Yamaguchi 755-8611, Japan
e-mail: m.obayas@yamaguchi-u.ac.jp

This work was presented in part at the 13th International Symposium on Artificial Life and Robotics, Oita, Japan, January 31–February 2, 2008

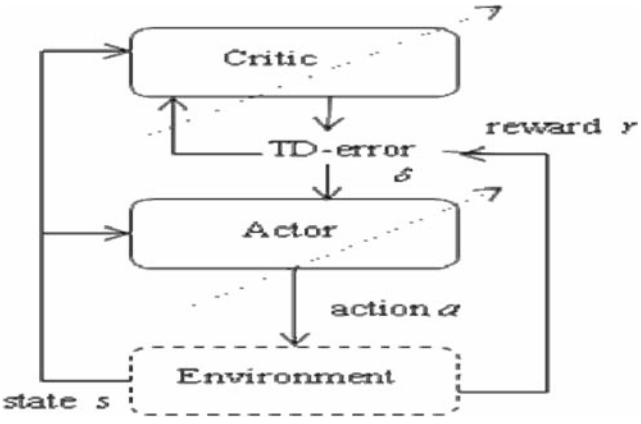

Fig. 1. Actor-critic model

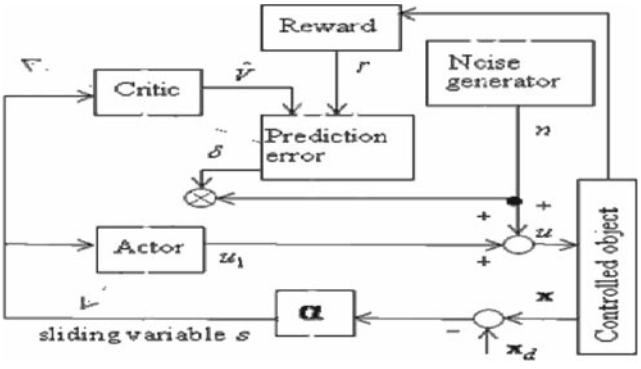

Fig. 2. Structure of the propose d RRL system with SMC

consists of an actor generating a control signal and a critic evaluating it. The actor–critic model is shown in Fig. 1. The critic calculates the predictive state value function $\check{V}$. The prediction error δ is defined by Eq. 7. Here $r_t + \gamma \cdot \check{V}_{t+1}$ in Eq. 7 is the target of $\check{V}_t$, and then δ is called the TD-error. The output of the critic, $\check{V}_t$, is adjusted in order that δ converges to zero. The actor generates a control signal based on δ, i.e., as a result of the previous action. If the TD-error is positive, the choice of the previous action is desirable, and in a similar situation the probability of selecting this action is enhanced. The opposite is also true.[2]

3 Sliding mode control

Sliding mode control is described below. First, it restricts the state of the system to a switching surface set-up in the state space. Then it generates sliding modes (see Eq. 3) on the switching surface, and then stabilizes the state of the system to a specified point in the state space. The main feature of sliding mode control is its robustness, which is such that the control is not susceptible to model uncertainty and disturbance (Fig. 2).

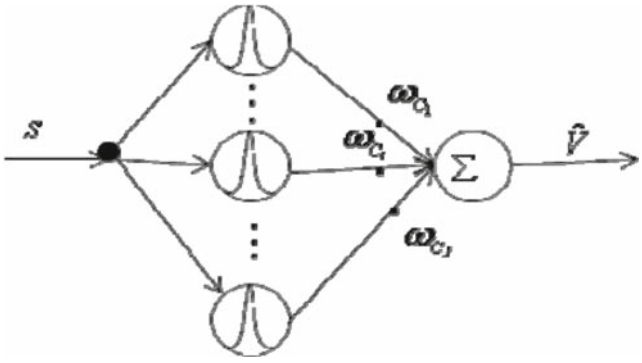

Fig. 3. Structure of the critic

4 Reinforcement learning with the concept of sliding mode control

In this section, the reinforcement learning method with the concept of sliding mode control is explained. The aim of this method is to enhance the robustness, as this cannot be obtained by the conventional reinforcement learning method.

4.1 Controlled system

We now consider the *n*-th order nonlinear differential equation.

$$x^{(n)} = f(\mathbf{x}) + b(\mathbf{x})\,u, \tag{1}$$

where $\mathbf{x} = [x, \dot{x}, \ldots, x^{(n-1)}]^T$ is a state vector of the system. Here, it is assumed that all states are observable. u is the control input, and $f(\mathbf{x})$, $b(\mathbf{x})$ are unknown continuous functions. However, it is known that u is input to the system as in the formulation of Eq. 1.

The object of the system is to decide control input u which leads the states of the system to the target of the states of system, $\mathbf{x}_d$. We define the error vector $\mathbf{e}$ as follows:

$$\begin{aligned}\mathbf{e} &= [e, \dot{e}, \ldots, e^{(n-1)}]^T, \\ &= [x - x_d, \dot{x} - \dot{x}_d, \ldots, x^{(n-1)} - x_d{}^{(n-1)}]^T\end{aligned} \tag{2}$$

Sliding surface H is defined as

$$H : \{\mathbf{e} \mid s(\mathbf{e}) = 0\}, \tag{3}$$

$$s(\mathbf{e}) = \boldsymbol{\alpha}^T \mathbf{e}, \tag{4}$$

where $\alpha = [\alpha_0, \alpha_1, \ldots, \alpha_{n-1}]^T$, $\alpha_{n-1}p^{n-1} + \alpha_{n-2}p^{n-2} + \cdots + \alpha_1 p + \alpha_0$ is a strictly stable polynomial in the meaning of Hurwitz, $p = (d/dt)$ is Laplace operator.

4.2 Constitution of the actor–critic

4.2.1 Constitution of the critic

The critic is constituted from the radial basis function (RBF) network (Fig. 3). The input to the critic is the sliding

variable s in sliding mode control, and its output is predictive state value function, $\check{V}$, as follows:

$$\hat{V} = \sum_{i=1}^{J} \omega_{C_i} \cdot \exp\left\{\frac{-(s - c_{C_i})^2}{\sigma_{C_i}^2}\right\} \tag{5}$$

ω_{C_i} are learning parameters, c_{C_i}, σ_{C_i} is the i-th average and the standard deviation of the i-th RBF without learning here, respectively. Reward r is defined as in Eq. 6, and is given in order to make variable s restricted to zero. The TD-error δ is defined as in Eq. 7, γ is the damping coefficient, and r_d is a positive constant.

$$r_t = \begin{cases} +r_d & (s_{t-1}^2 - s_t^2 > 0) \\ -r_d & (s_{t-1}^2 - s_t^2 < 0) \end{cases} \quad (r_d > 0) \tag{6}$$

$$\delta_t = r_t + \gamma \cdot \hat{V}_{t+1} - \hat{V}_t \quad (0 < \gamma \leq 1). \tag{7}$$

4.2.2 Constitution of an actor

An actor is also constituted of RBFN, as is a critic. The input to the actor is also the sliding variable s, and its output u_1 is part of control signal u. as follows:

$$u_1 = \sum_{i=1}^{N} \omega_{Ai} \cdot \exp\left\{\frac{-(s - c_{Ai})^2}{\sigma_{Ai}^{\ 2}}\right\} \tag{8}$$

ω_{A_i} is a learning parameter,

4.3 Learning

4.3.1 Learning the critic's parameters

The learning of the critic is done by using the well-known back propagation method, which makes the TD-error zero. The updating rules of the parameters are as follows:

$$\Delta\omega_i^c = -\eta_c \cdot \frac{\partial \delta_t^2}{\partial \omega_i^a}, \quad (i = 1, \ldots, J), \tag{9}$$

The parameters of critic RBFN, ω_i^c $(i = 1, \ldots, J)$, are adjusted in order to make $\delta = 0$.

4.3.2 Learning the actor's parameters

The parameters of the actor, ω_i^a $(i = 1, \ldots, N)$, are adjusted by using the output u_1 of the actor and noise n.

$$\Delta\omega_i^a = \eta_a \cdot n_t \cdot \delta_t \cdot \frac{\partial u_1}{\partial \omega_i^a}, \quad (i = 1, \ldots, N), \tag{10}$$

$\eta_a(>0)$ is the learning coefficient. Equation 10 means that $(-n_t \cdot \delta_t)$ is considered as an error, and ω_i^a is adjusted to the opposite sign to $(-n_t \cdot \delta_t)$.

4.4. Noise for exploration

Noise n_t is to maintain diversity when searching for optimal parameters. When n is bigger, the absolute value of the sliding variable s is bigger, and when n is smaller, the absolute value of the sliding variable is also smaller. The noise used here is a uniform random number, and is generated in order not to go over the upper limit $\bar{n}$.

$$n = z \cdot \bar{n} \cdot \exp\left(-\beta \cdot \frac{1}{s^2}\right), \tag{11}$$

z is uniform random number of range [–1, 1]. $\bar{n}$ is the upper limit of the perturbation signal for searching, and β is an adjustment parameter.

5 Computer simulation

5.1 Controlled system

To verify the effectiveness of the proposed method, we carried out the control simulation using the inverted pendulum problem with dynamics, as described in Eq. 12. (Fig. 4).

$$mg\ddot{\theta} = mgl \sin\theta - \mu_V \dot{\theta} + u \tag{12}$$

The parameters in Eq. 10 are described in Table 1. The simulation is carried out using MatX[3].

5.2 Simulation conditions

One trial means that the control starts at $(\theta_0, \dot{\theta}_0 = (\pi/_{18}[rad], 0[rad/s])$ and continues for 20 s, and the sampling time is 0.02 s. The trial ends if $|\theta| \geq \pi/6$. or controlling time is over 20 s. We set an upper limit of for the output u_1 of the actor. Trial success means that θ is in range $[-\pi/360, \pi/360]$ for the last 10 s.

5.3 Parameters of the proposed method

The number of the RBFN, the function approximator consisting of actor–critic, is 15, the averages are C_C,C_A and the standard deviation σ_A, σ_C of them are set as follows:

Fig. 4. An inverted pendulum

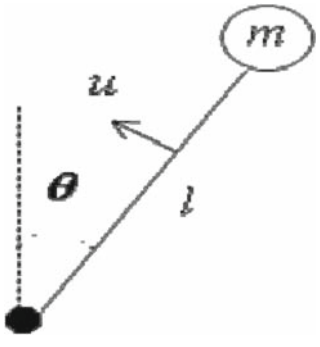

Table 1. Parameters of the system

θ: joint angle
m = 1.0 [kg]
l = 1.0 [m]: lenth of the pendulum
g= 9.8 [m/s^2]
μ_V =0.01: coefficien of friction
u_T: input torque
Observation vector: $\mathbf{x} = [\theta, \dot{\theta}]^T$

Table 2. Success rate of learning

Proposed	Actor–critic	PID
170/300	2/3000	–

$$C_C = [-15,-12,-9,-6,-3,-2,-1,0,1,2,3,6,9,12,15],$$
$$C_A = [-15,-12,-9,-6,-3,-2,-1,0,1,2,3,6,9,12,15],$$
$$\sigma_C = [\,3,3,3,3,2,1,0.5,\ 0.3,\ 0.5,\ 1,2,3,3,3,3],$$
$$\sigma_A = [\,3,3,3,3,2,1,0.5,\ 0.3,\ 0.5,\ 1,2,3,3,3,3].$$

$\alpha = [15,1]^T$. ω_{C_i}, ω_{A_i} are initially set to a small random number and not equal to zero. The upper limit of the control signal u_1 is set to 10 N·m. Other parameters are set to $\gamma = 0.9$, $\bar{n} = 5.0$, $r_d = 10.0$, $\eta_a = 0.1$, $\eta_c = 0.1$, $\beta = 10.0$.

These are decided by trial and error.

5.4 Parameters of the conventional actor–critic method

The predictive state value function, a part of the control signal, and the reward function using in the conventional actor–critic method are, respectively, as follows:

$$\hat{V} = \sum_{i=1}^{J} \omega_{Ci} \exp\left(-\frac{(\theta - c_{Ci_\theta})^2}{\sigma_{Ci_\theta}^2} - \frac{(\dot{\theta} - c_{Ci_{\dot{\theta}}})^2}{\sigma_{Ci_{\dot{\theta}}}^2} \right) \tag{13}$$

$$u_1 = \sum_{i=1}^{J} \omega_{Ai} \exp\left(-\frac{(\theta - c_{Ai_\theta})^2}{\sigma_{Ai_\theta}^2} - \frac{(\dot{\theta} - c_{Ai_{\dot{\theta}}})^2}{\sigma_{Ai_{\dot{\theta}}}^2} \right) \tag{14}$$

$$r_t = 10 \exp\left(-\frac{(\theta_t)^2}{2(\alpha_{R1})^2} - \frac{(\dot{\theta}_t)^2}{2(\alpha_{R2})^2} \right) - 5.0 \tag{15}$$

where, C_{Ci_θ}, $C_{Ci_{\dot{\theta}}}$, σ_{Ci_θ}, $\sigma_{Ci_{\dot{\theta}}}$ in the critic are same as those proposed, and C_{Ai_θ}, $C_{Ai_{\dot{\theta}}}$, σ_{Ai_θ}, $\sigma_{Ai_{\dot{\theta}}}$ in the actor are 1/5 of those proposed. Others are set to $\alpha_{R1} = 0.2$, $\alpha_{R2} = 20$, $\eta_a = 0.1$, $\eta_c = 0.1$, $\beta = 0.1$ by trial and error.

5.5 Parameters of the PID control method

The control signal u_T in PID is defined as follows:

$$u(t) = -K_p e(t) - K_d \dot{e}(t) - K_i \int^t e(t)\, dt \tag{16}$$

where the parameters decided by trial and error are $K_p = 50$, $K_d = 30$, $K_i = 0.5$.

5.6 Simulation results

Table 2 shows the success rate of learning. In Table 2, A/B means the number of successes A per number of trials B. The table shows that learning by the proposed method is more stable than that of the actor–critic, but neither rate is good. The reason may be that (1) the average and standard deviation of RBF were not learned, or (2) the number of the RBF was not adjusted.

The control results of θ, $\dot{\vartheta}$ are shown in Figs. 5 and 6, respectively. Both results by the proposed method oscillate in the first 5 s, but the steady state error of θ is the smallest of the three, and that of $\dot{\vartheta}$ is worse than that by the conventional actor–critic method. Results of the robust performance after a change in m are shown in Table 3. The proposed method is best for smaller changes in m, but if m is made bigger, the proposed method is better than the actor–critic but worse than PID.

Table 3. Robust performance with a change in m

	Proposed	Actor–critic	PID
m-max [kg]	2.202	1.668	3.439
m-min [kg]	0.008	0.021	0.022

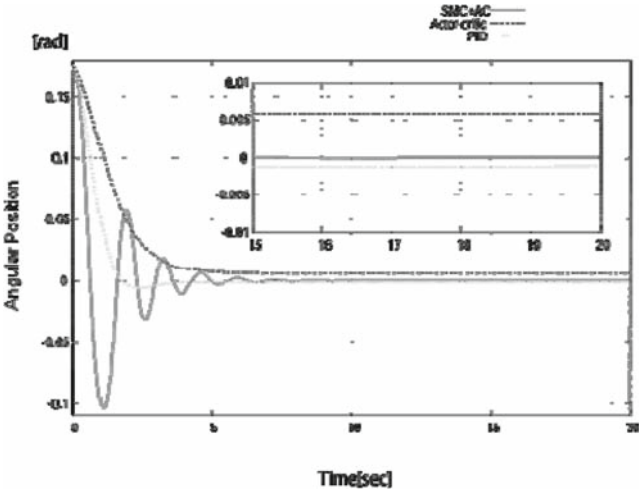

Fig. 5. Result of learning control for θ

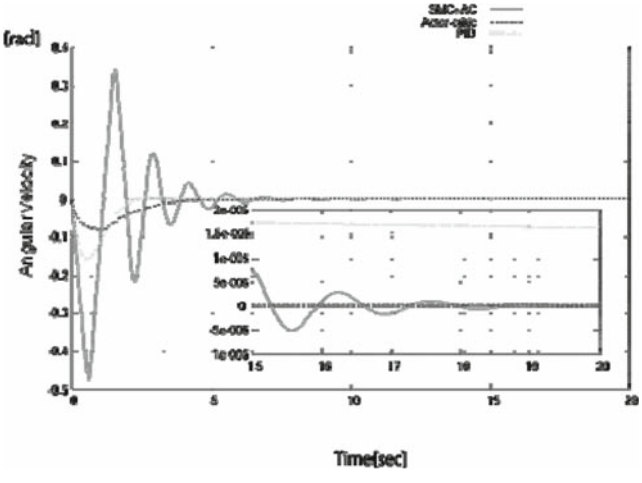

Fig. 6. Result of learning control for $\dot{\theta}$

6 Conclusion

A robust reinforcement learning method the using concept of sliding mode control has been proposed. Through an

inverted pendulum control simulation, it was verified that the proposed method has good robustness, but does not give good control performance, mainly because of its oscillations.

In future work, it will be necessary to improve the control performance and to clarify the limit of robust stability of the proposed method.

References

1. Morimoto J, Doya K (2005) Robust reinforcement learning. Neural Comput 17:335–359
2. Sutton RS, Barto AG (1998) Reinforcement learning. An introduction. MIT Press, Cambridge
3. Koga M (2000) Numerical computation by MATX. Tokyo Denki University Press, Tokyo