논문 2010-4-23

카트-폴 균형 문제를 위한 실시간 강화 학습

On-line Reinforcement Learning for Cart-pole Balancing Problem

김병천*, 이창훈**

Byung-Chun Kim*, Chang-Hoon Lee**

요 약 Cart-pole 균형 문제는 유전자 알고리즘, 인공신경망, 강화학습 등을 이용한 제어 전략 분야의 표준 문제이다. 본 논문에서는 cart-pole 균형문제를 해결하기 위해 실시간 강화 학습을 이용한 접근 방법을 제안하였다. 본 논문의 목적은 cart-pole 균형 문제에서 OREL 학습 시스템의 학습 방법을 분석하는데 있다. 실험을 통해, 본 논문에서 제안한 OREL 학습 방법은 Q-학습보다 최적 값 함수에 더 빠르게 접근함을 알 수 있었다.

Abstract The cart-pole balancing problem is a pseudo-standard benchmark problem from the field of control methods including genetic algorithms, artificial neural networks, and reinforcement learning. In this paper, we propose a novel approach by using online reinforcement learning(OREL) to solve this cart-pole balancing problem. The objective is to analyze the learning method of the OREL learning system in the cart-pole balancing problem. Through experiment, we can see that approximate faster the optimal value-function than Q-learning.

Key Words : Reinforcement Learning, Q-learning, Cart-pole Balaning, Optimal value function

Ⅰ. 서 론

학습(learning)이란 과거의 경험을 이용하여 현재의 문제를 해결하기 위한 지식(knowledge)이나 기술(skill)을 의미하며[1], M.L. Minsky에 의해 소개된 강화 학습(reinforcement learning)은 동물의 학습을 연구하는 과정에서 기원하였다[2].

강화 학습은 그림1과 같이 학습을 수행하는 에이전트(agent)와 에이전트 외부에 존재하는 환경과 시행-착오(trial-and- error)를 통해 상호 작용(interaction)하면서 학습한다. 즉, 에이전트는 학습을 수행하는 동안 주어진 환경에서 취할 수 있는 행동(action, a_t)을 시도(trial)하며, 에이전트의 외부 환경으로부터 에이전트가 선택한 행동에 대한 평가로서 스칼라(scalar) 형의 강화-값(γ_t : reinforcement value)을 받아 강화된다[3].

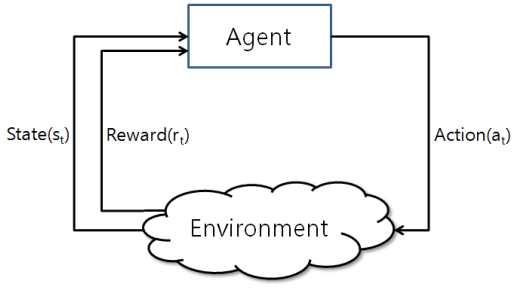

그림 1. 강화학습 모델
Fig. 1. Reinforcement Model

이와 같은 형태의 강화 학습은 동적 환경에서 효율적으로 학습을 수행할 수 있기 때문에 cart-pole 균형 문제[4,5], 교통신호 제어(traffic light control)[6], 엘리베이터 최

*정회원, 한경대학교 웹정보공학과
**정회원, 한경대학교 컴퓨터공학과(교신저자)
접수일자 2010.6.13, 수정일자 2010.7.20
게재확정일자 2010.8.13

적운행[7] 그리고 로봇 이동[8] 등과 같은 동적 환경에서 학습을 위해 널리 이용되고 있다.

본 논문에서는 cart-pole 제어 문제와 같은 동적 환경에서 효율적으로 학습을 수행할 수 있는 실시간 강화 학습(OREL : Online REinforcement Learning system)을 제안한다.

OREL은 cart-pole 균형문제를 효율적으로 학습하기 위해 cart와 pole의 상호 관계를 적용한 강화-값을 이용한다. 일반적으로 cart-pole 시스템을 제어하기 위한 강화 학습의 성능평가는 학습 시스템이 몇 회의 시도 만에 cart가 트랙의 범위를 벗어나지 않고, pole이 쓰러지지 않도록 균형을 유지 할 수 있는가를 평가 기준으로 한다.

본 논문에서 제안한 OREL은 cart-pole 제어 환경에서 강화 학습으로 가장 널리 알려진 Q-학습(Q-learning)과 [9] 비교한 결과 학습의 성능을 결정하는 최적 값 함수(optimal value-function)에 빠르게 수렴하는 것을 실험을 통해 알 수 있었다.

본 논문의 구성은 다음과 같다. 2장에서는 cart-pole 시스템에 대해 설명 하였고, 3장에서는 본 논문에서 제안된 OREL의 학습 알고리즘에 대하여 설명하였다. 그리고 4장에서는 제안된 알고리즘과 가장 널리 이용되고 있는 Q-학습과 cart-pole 환경에서 수렴 속도를 비교 분석하였다. 5장에서는 결론과 함께 향후 연구 방향을 제시하였다.

II. 관련연구

1. 강화학습

cart-pole과 같은 동적 환경에서 학습을 수행하개 위해 Q-학습이 널리 이용되고 있다. Q-학습은 통계적 프로그래밍(stochastic dynamic programming)에 근거한 학습 방법으로서 학습을 수행하는 에이전트가 현재상태(s_t)에서 임의의 행동(a_t)을 수행하였을 때 외부 환경으로부터 강화 값에 대한 근사 값(approximation value)을 (상태-행동) 쌍에 대한 Q-함수, ($Q(s_t,a_t)$)에 할당한다. 그리고 나서 다음 상태(s_{t+1})의 (상태-행동) 쌍에 대한 Q-함수가 최대가 되는 행동(a_{t_1})을 선택하여 현재 상태의 (상태-행동) 쌍에 대한 Q-값을 식1과 같이 갱신하면서 학습을 수행한다.

$$Q(s_t,a_t) = (1-\alpha) \cdot Q(s_t,a_t) + \alpha \cdot \delta_t \quad (1)$$

식1에서 α는 학습률(learning rate), δ_t는 현재 상태에서 선택한 (상태-행동) 쌍에 대한 TD-오류(error)로서 식2와 같이 계산된다.

$$\delta_t = r_{t+1} + \gamma\{\max Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)\} \quad (2)$$

현재 상태에서 Q-함수에 의해 계산된 Q-값들 중에서 가장 적당한 행동을 선택하는 일반적인 방법은 식3과 같은 볼쯔만(Boltzmann) 확률 분포에 따라 행동을 선택하는 방법이 널리 이용되고 있다.

$$p(\bar{a} \mid s) = \frac{e^{\frac{Q(s_t,\bar{a})}{T}}}{\sum_{a \in A(s_t)} e^{\frac{Q(s_t,a)}{T}}} \quad (3)$$

식3에서 T는 행동 선택의 임의성(randomness) 정도를 제어하는 온도변수(temperature variable)이다. 볼쯔만 확률 분포에 의한 행동 선택은 현재 상태 s_t에서 선택 가능한 모든 행동 $\bar{a}$ 에 대한 확률 값 $p(\bar{a} \mid s_t)$를 계산하고 그 값 중에서 가장 큰 값과 (0,1) 사이의 난수(random number)를 비교하여 행동을 선택한다.

2. Cart-pole 시스템

cart-pole 시스템은 강화 학습의 성능을 평가하기 위해 널리 이용되는 표준 문제이다. cart-pole 제어 문제는 그림 2와 같이 cart와 cart에 수직으로 세워진 pole로 구성되어 있다.

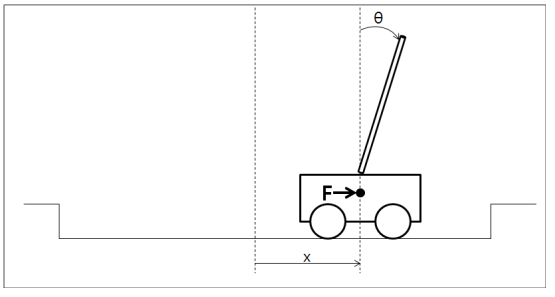

그림 2. cart-pole 시스템
Fig. 2. cart-pole system

그림2에서 cart($1.0kg$)는 일정한 공간의 트랙($|x| \leq \pm 2.4m$)내에서 $10N$의 힘으로 왼쪽 또는 오

른쪽으로 자유롭게 이동할 수 있다. pole은 길이가 $1.0m$이며, cart와 트랙에 수직(vertical) 방향으로 $\pm 12^\circ$ 까지만 움직일 수 있다.

Cart-pole 시스템에서 cart와 pole의 상태는 표1과 같이 4개의 상태 값(x, θ, $\dot{x}$, $\dot{\theta}$)으로 표현된다.

표 1. cart-pole 시스템의 상태 값의 의미
Table1. Status value of cart-pole system

변 수	의미
x	cart의 위치
θ	cart와 pole의 수선이 이루는 각
$\dot{x}$	cart의 속도
$\dot{\theta}$	pole의 각속도

cart-pole 시스템에서 학습 시스템은 매 학습 단계마다 cart에 일정한 힘($F = \pm 10N$)이 적용되며, cart에 힘을 가했을 때 cart와 pole의 동작($\ddot{x}_t$, $\ddot{\theta}_t$)은 식4와 식5와 같이 계산된다.

$$\ddot{x}_t = \frac{F_t + m_p \bullet l(\dot{\theta}_t^2 \bullet \sin\theta_t - \ddot{\theta}_t \bullet \cos\theta_t)}{m} \tag{4}$$

$$\ddot{\theta}_t = \frac{m_g \bullet \sin\theta_t - \cos\theta_t[F_t + m_g \bullet l \bullet \dot{\theta}_t^2 \bullet \sin\theta_t]}{(\frac{4}{3})m \bullet l - m_p \bullet l \bullet \cos^2\theta_t} \tag{5}$$

식4와 식5에서 사용된 변수들의 의미는 표2와 같다.

표 2. cart-pole 시스템 변수들
Tabel 2. cart-pole system variables

변 수	의 미
l	pole의 길이($1.0m$)
m_p	pole의 무게($0.1kg$)
m	cart의 와 pole의 무게($1.1kg$)
F	cart에 가해지는 힘($\pm 10N$)
g	가속도($-9.8m/s^2$)

일정한 힘($\pm 10N$)이 cart에 적용되었을 때 cart와 pole의 다음 상태 값은 오일러(Euler)가 제시한 불연속 방정식(discrete-time equation) $\theta_{(t+1)} = \theta_t + \tau\dot{\theta}_t$을 이용하여 구할 수 있다(이때, $\tau = 0.02$초).

Ⅲ. 실시간 강화학습

본 논문에서는 cart-pole 시스템과 같은 동적 환경을 제어하기 위한 실시간 강화 학습 시스템(OREL)을 제안한다. cart-pole 시스템을 제어하기 위한 ONRELS의 구조는 그림3과 같이 학습기(learner)와 선택기(selector)와를 가지고 있다.

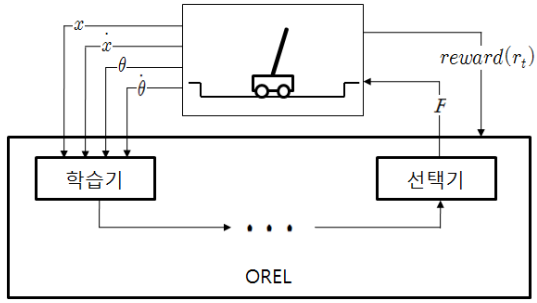

그림 3. OREL 시스템 구조
Fig. 3. OREL system structure

OREL 시스템의 학습기는 cart-pole 시스템의 상태 정보를 입력받아 에이전트가 현재 상태에서 선택 가능한 모든 (상태-행동) 쌍들에 대해 각 (상태-행동) 쌍들을 평가하여 최적의 (상태-행동) 쌍을 선택하는 역할을 수행하고, 선택기는 학습기에 의해 선택된 (상태-행동) 쌍에 대한 최적 값 함수를 갱신하면서 에이전트가 수행해야할 행동을 결정하는 역할을 한다.

1. 학습기

학습기는 식6과 같이 현재 상태에서 선택 가능한 모든 (상태-행동) 쌍에 대한 값을 갱신한다.

$$\begin{aligned} &\text{for} \quad all \quad s_t, a(a \in A(s_t)), \\ &\quad Q(s_t, a_t) = Q(s_t, a_t) + \alpha\delta_t \\ &\delta_t = r_{t+1} + \gamma\{ \max_{a \in A(s_t)} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \} \end{aligned} \tag{6}$$

식6에서 α는 학습율, $\gamma(0<\gamma<1)$는 할인율(discount factor) 그리고 r_{t+1}은 에이전트가 선택한 행동에 대한 평가를 나타내는 강화 값이며, 식7과 같이 계산된다.

$$r_{t+1}(s_t,a_t)=\begin{cases}-1 & \text{if } |x|\geq 2.4m \text{ or } |\theta|\geq 12^o \\ 0 & otherwise\end{cases} \quad (7)$$

식7을 이용하여 현재 상태에서 선택 가능한 모든 (상태-행동) 쌍에 대한 값을 갱신하고 나서, 식8의 Gibbs 확률 분포를 이용하여 현재 상태에서 가장 적합한 (상태-행동) 쌍을 선택한다.

$$p(a\mid x)=\frac{e^{kQ(s_t,a_t)}}{\sum_{a'\in A(s_t)} e^{kQ(s_t,a')}} \quad (8)$$

식8에서, $k(0<k<1)$는 이미 학습된 Q-값을 어느 정도의 가중치를 가지고 참조할 것인가를 결정하는 상수이다. k값이 작을수록 현재의 Q 값을 무시하고 새로운 (상태-행동) 쌍을 선택할 확률이 높아진다. 그러므로 초기 학습 단계에서는 k 값을 낮게 설정하여 에이전트가 새로운 상태를 탐색할 수 있게 하고, 학습을 반복할수록 k값을 증가시켜 큰 값을 갖는 (상태-행동) 쌍을 선택하게 한다. 학습기의 학습 알고리즘은 그림4와 같다.

```
Learner() {
   Initialize Q(s,a) Table and e(s,a) Table
      arbitrarily for all s, a;
   Repeat {
     Sesect s_t as a start state;
     Choose a_t from s_t using policy derived from
        Q(s_t,a_t);
     while(count < 100000) {
     Take an action a_t;
     Observe s_{t+1} , r_{t+1};
     Choose a_{t+1} from s_{t+1};
     Update_Qvalue(s_t, a_t, r_{t+1}, s_{t+1});
     s_t = s_{t+1};
     a_t = a_{t+1}
    } // End of while
  } // End of repeat
}
```

그림 4. 학습 알고리즘
Fig. 4. Learning Algorithm

그림3의 학습 알고리즘에서 *Update_Qvalue* $(s_t,a_t,r_{t+1},s_{t+1})$는 선택 가능한 모든 (상태-행동) 쌍에 대한 평가 값을 갱신하는 함수로서 처리 절차는 그림 5와 같다.

```
Update-Qvalue(s_t) {
   For al s_t,a_t (a∈A(s_t)) {
     if(s_t is observable) {
       r_{t+1} = REWARD;
       Q(s_t,a_t) = (1-α)Q(s_t,a_t) + αδ_t
     }
   else
       Q(s_t,a_t) = -∞ ;
  }
}
```

그림 5. *Update_Qvalue* 함수
Fig 5. Update_Qvalue function

2. 선택기

선택기는 식7에 의해 선택된 (상태-행동) 쌍에 대한 평가를 식8과 같이 갱신한다.

$$Q(s_t,a_t)=(1-\alpha)Q(s_t,a_t)+\alpha\delta_t e(s_t,a_t) \quad (8)$$

식8에서 δ_t는 현재 상태에서 선택된 (상태-행동) 쌍에 대한 Q-값과 현재 상태에서 선택 가능한 다음 상태들 중에서 최대 Q-값을 갖는 (상태-행동) 쌍과의 TD-오류(temporal-difference error)로서 식9와 같다.

$$\delta_t=r_{t+1}+\gamma\{\max_{a\in A(s_t)} Q(s_{t+1},a_{t+1})-Q(s_t,a_t)\} \quad (9)$$

식9에서 $e(s_t,a_t)$는 선택된 (상태-행동) 쌍이 얼마나 적합한가를 나타내는 적합도로서 식10과 같다.

$$e(s_t,a_t)=\begin{cases}1 & \text{if } s=s_t \text{ and } a=a_t \\ \gamma\lambda e(s_{t-1},a_{t-1}) & otherwise\end{cases} \quad (10)$$

식10은 에이전트가 탐색 과정에서 선택한 현재 상태의 (상태-행동) 쌍이 이미 선택된 (상태-행동) 쌍인 경우 현재 상태에서 선택한 (상태-행동) 쌍에 대한 적합도를 1로 하고, 그렇지 않은 경우 적합도는 $\gamma\lambda(0<\gamma<1,\ 0<\lambda<1)$ 만큼씩 감소된다.

Ⅳ. 실험 및 결과

일반적으로 cart-pole 균형 문제를 제어하기 위한 강화 학습 방법들은 얼마나 오랜 시간동안 cart가 트랙의 범위를 벗어나지 않고, pole이 쓰러지지 않고 균형을 유지할 수 있는가를 평가 기준으로 한다.

본 논문에서 제안한 OREL 학습 방법과 강화학습의 가장 대표적인 Q-학습 방법을 cart-pole 시스템에 적용하였다. cart를 100,000번 움직이는 동안 pole이 균형을 유지하기 위해 몇 번 시도하였는가를 평가 기준으로 하였다. 실험 결과 그림5와 같이 Q-학습은 126회 만에 성공하였고 OREL은 44회 만에 성공하였다.

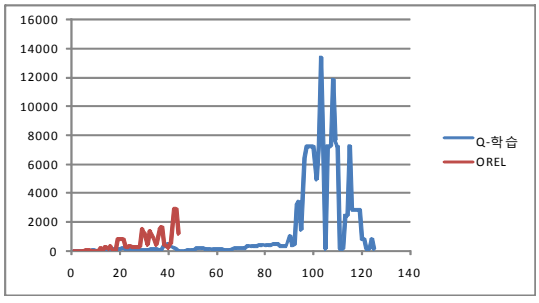

그림 6. Q-학습과 OREL의 학습 회 수 비교
Fig. 6. Learning number of Q-learning and OREL learning

OREL 학습 시스템이 44번 시도하는 동안 cart의 위치는 그림7과 같이 $(-2.4m < x < 2.4m)$ 사이에서 안정적으로 움직였다.

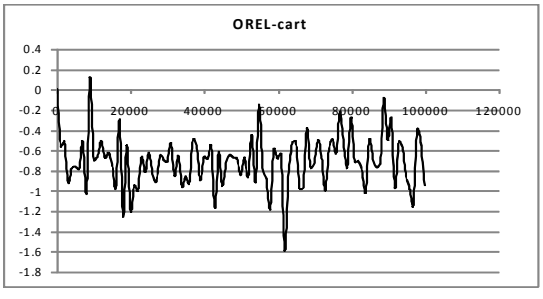

그림 7. cart의 위치
Fig. 7. cart position

또한 OREL 학습 시스템의 cart가 100,000번 이동하는 동안 cart와 pole이 이루는 각(θ)는 그림8과 같이 $(-10.2^{\circ} < \theta < 10.2^{\circ})$사이에서 안정적으로 움직였음을 알 수 있다.

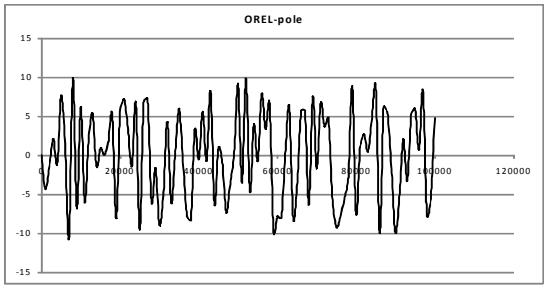

그림 8. pole의 각도
Fig. 8. pole angle

Ⅴ. 결 론

본 논문에서는 cart-pole 시스템과 같은 동적 환경에서 최적값 함수에 빠르게 수렴할 수 있는 강화 학습 방법을 제안하였다. 실험결과 강화학습의 가장 대표적인 Q-학습 방법보다 최적값 함수에 빠르게 수렴함을 알 수 있었다. 그러나 주어진 환경에 대한 모형(model)을 이용한 학습 방법이기 때문에 매우 제한적이라 할 수 있다.

실세계에 대한 정확한 모형을 얻기가 매우 어려우므로 강화 학습의 성능을 더욱 향상시키기 위한 계획(planning)과 학습을 유기적으로 통합한 강화학습 알고리즘에 대한 연구가 필요하다. 또한 cart-pole 시스템은 하나의 에이전트가 학습을 수행하지만, 여러 에이전트가 서로 협력하면서 학습을 수행할 수 있는 학습 알고리즘에 대한 연구가 필요하다.

참 고 문 헌

[1] 김병천, 윤병주, "복수전략학습", *정보과학회지*, 13권, 5호, pp45-52, 1995.

[2] M.L.Minsky *Theory of Neural-Analog Reinforcement Systems and Application to the Brain-Model Problem*, Ph.D. Thesis, Princeton University, Princeton, 1954.

[3] A. G. Barto, D. A. White and D. A. Sofge, "Reinforcement Learning and adaptive critic model", *Handbook of Intelligent Control*, pp.

469-491,1992.

[4] C. W. Anderson, "Learning to control an inverted pendulum using neural networks", *IEEE Control Systems Magazine*, pp.31-37, 1989.

[5] O. Pinngern and T. H. Nguyen, "International Symposium on Electrical & Electronics Engineering", HCM City, Vietnam, 2007.

[6] As'ad Salkham, Raymond Cunningham, Anurag Garg, and Vinny Cahill, "A Collaborative Reinforcement Learning Approach to Urban Traffic Control", *IEEE/WIC/ACM International Conference*, Vol. 2 (2008), pp. 560-566.

[7] T. Walczak and P. Cichosz. "A distributed learning control system for elevator groups", *Artificial Intelligence and Soft Computing (ICAISC-06), volume 4029 of Lecture Notes in Computer Science*, pp.1223 - 232. Springer, 2006.

[8] K Conn and R A Peters, ""Reinforcement Learning with a Supervisor for a Mobile Robot in a Real world Environment", *Computational Intelligence in Robotics and Automation*, pp. 73-78, 2007

[9] G. Cybenko, R. Gray, and K. Moizumi, "Q-learning : A Tutorial and Extensions", *Mathematics of Artificial Neural Networks*, Oxford University, July, 1995.

저자 소개

김 병 천(정회원)

- 1988년 한남대학교 컴퓨터공학과 학사
- 1991년 숭실대학교 컴퓨터공학과 석사
- 1999년 명지대학교 컴퓨터공학과 박사
- 경력 1991년 ~ 현재 한경대학교 웹정보공학과 교수

<주관심분야 : Machine Learning, Computer Vision, Agent System 등>

이 창 훈(정회원)

- 1998년 중앙대학교 대학원 컴퓨터공학과(공학박사)
- 2002년 ~ 현재 한경대학교 컴퓨터공학과 교수

<주관심분야 : 객체지향, 정형화방법, 컴포넌트, 영상처리 등>