SGA 기반 강화학습 알고리즘을 이용한 로봇 제어

Robot Control via SGA-based Reinforcement Learning Algorithms

박주영*, 김종호*, 신호근**
*고려대학교 제어계측공학과, **고려대학교 컴퓨터학과
Jooyoung Park*, Jongho Kim*, and Hokuen Shin**
*Dept. of Control & Instrumentation Engineering, Korea University
**Dept. of Computer Science & Engineering, Korea University
E-mail: parkj@korea.ac.kr, oyeasw@korea.ac.kr, hkshin@image.korea.ac.kr

Abstract

The SGA(stochastic gradient ascent) algorithm is one of the most important tools in the area of reinforcement learning, and has been applied to a wide range of practical problems. In particular, this learning method was successfully applied by Kimura et al. [1] to the control of a simple creeping robot which has finite number of control input choices. In this paper, we considered the application of the SGA algorithm to Kimura's robot control problem for the case that the control input is not confined to a finite set but can be chosen from a infinite subset of the real numbers. We also developed a MATLAB-based robot animation program, which showed the effectiveness of the training algorithms vividly.

키워드 : 강화학습, SGA 알고리즘, Kimura의 로봇

1. 서론

강화 학습은 기계학습(machine learning) 분야의 주요한 도구로써 여러 분야에서 흥미 있는 결과를 계속적으로 제공하여 왔는데, 최근에는 자동제어 관련 분야에서도 흥미 있는 적용 사례가 보고된 바 있다[1]. 강화학습은 상태(state) 및 제어 입력(action 또는 control input) 공간이 이산 집합(discrete set)인 경우에는 확고한 이론적 기초가 확립되어 있다[2]. 그 속 상태 및 연속 제어 입력(continuous state and continuous input)을 다루는 경우에는, 극히 제한된 부류의 문제[3-5]에 대해서만 엄밀한 증명이 제공되고 있지만 응용 사례는 상당히 넓은 분야에서 보고되고 있다[2]. 강화학습의 주요 방법론은 시스템의 모델에 관한 정보를 직접 이용하는 모델-기반 방법(model-based methods)과 모델을 필요로 하지 않는 방법(model-free method)으로 구별되는데, 본 논문에서 고려하는 SGA(stochastic gradient ascent) 기법은 후자에 속하는 방법이다. 이러한 시도는 시스템에 대한 구체적인 모델링 정보를 제어입력선택전략을 설계 단계에서 구체적으로 알지 못하는 가운데에서도 시스템 운전 중 관찰된 보상값(rewards)을 이용한 학습을 통해서 효과적인 제어 입력을 구할 수 있기 때문에 상당히 유용한 방법이 될 수 있다. 본 논문에서는 [1]에서 Kimura 등에 의해 소개된 간단한 기는 로봇[1]을 대상으로 하여, 연속 제어 입력을 갖는 경우를 위한 SGA 알고리즘을 적용해보고자 한다. 따라

1) 이하에서는 Kimura의 로봇으로 표기함

서, 본 논문은 기존의 방법론을 특정한 응용 예에 적용해보는 문제를 다룬 논문으로써, SGA 기법을 위한 일종의 튜토리얼 논문이라 할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는, 본 논문의 주요 소재가 되는 Kimura의 로봇[1]에 대하여 간단히 설명한 후, 이 로봇을 위해 제어 입력 공간이 유한 집합인 경우를 위한 SGA 알고리즘을 적용한 예가 소개된다. 그리고, 3장에서는 제어 입력이 한정된 이산 값을 취하는 것이 아니라 실수 범위에서 연속적인 값(continuous value)을 취하는 경우를 위한 로봇 제어 문제에 SGA 학습기법을 응용하는 문제를 다루면서 관련 수식을 유도하고, 실험 결과를 그림으로 정리한다. 마지막으로, 4장에서는 결론과 향후 연구 방향 등을 제시한다.

2. Kimura의 로봇과 학습

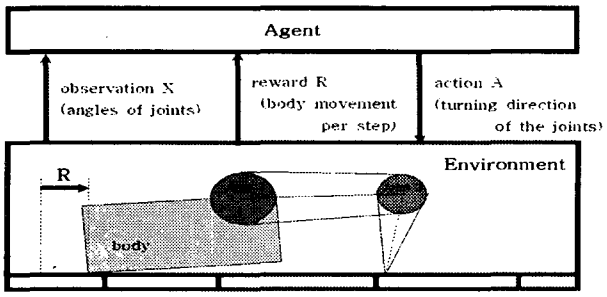

그림 1. Kimura의 로봇[1]

참고문헌 [1]에서 Kimura 등은 강화학습의 효용성을 보이기 위해 간단한 기는 로봇[1]을 응용 문제로 고려하였다. 이 로봇은, 중력이 가해지는 환경 아래에서 두 개의 링크를 가지고 기는 동작을 수행하는 평면형 머니퓰레이터(planar manipulator)로써 그림 1의 구조를 갖는다. 이 로봇에 부과된 임무는 최대한 빨리 전진하는 것인데, 에이전트(agent)는 로봇 및 환경에 대한 구체적인 모델 또는 정보가 주어지지 않은 상태에서 직접적인 경험을 통해 관찰된 보상값(rewards) r 만을 가지고 효과적인 제어 규칙을 발견해내야 한다. 각 시간 스텝 때마다 에이전트는 조인트의 각도를 읽어 들이고 확률적 제어입력 선택 전략에 따라 조인트에 연결된 모터의 회전 방향을 결정한다. 그리고, 학습 과정에서 이용되는 보상값 r 을 위해서는 해당 시간 스텝 동안 전진한 거리가 사용된다. 만일 로봇이 후진하는 경우에는 후진한 거리만큼의 음의 보상값(negative reward)이 생성됨은 물론이다. 직관적으로 생각할 때에, 위의 로봇이 최대한 빨리 전진하기 위해서는 기면서 앞으로 나아가는 패턴을 신속하게 습득해야 함을 알 수 있다

본 논문에서 고려하는 로봇 관련 데이터는 [1]의 경우와 같다. 따라서, 로봇의 위쪽 팔의 길이는 34 cm이고(이하, 단위 생략), 아래쪽 팔의 길이는 20이다. 그리고, 몸체와 위쪽 팔을 잇는 첫 번째 조인트는 몸체의 좌측하단 코너로부터 수평방향으로 32, 수직방향으로 18 떨어진 곳에 위치한다. 몸체와 위쪽 팔을 잇는 조인트의 움직임은 몸체와 수평인 방향에서 [-4, 35] 도 범위에서만 가능하고, 위쪽 팔과 아래쪽 팔을 잇는 두 번째 조인트의 움직임은 위쪽 팔과 수평인 방향에서 [-120, 10] 도 범위에서만 가능하다. 그리고, 아래쪽 팔의 뾰족한 끝부분이 지면에 닿아 있을 때에는, 뾰족한 끝부분은 미끄러지지 않고 몸체만 미끄러짐을 가정한다.

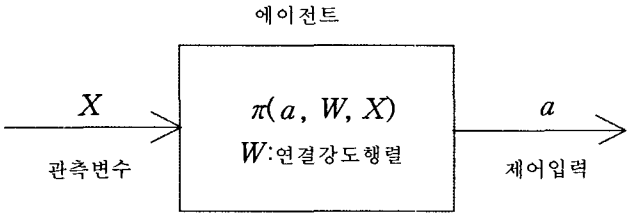

그림 2. 확률적 제어입력선택규칙: 제어입력 a가 발생될 확률은 $\pi(a, W, X)$ 임

에이전트가 생성하는 제어입력은 그림 2와 같이 분포 π에 따라 확률적으로 생성된다. 이 그림에서 관측변수 X는 첫 번째 및 두 번째 조인트의 각도를 정규화 시킨 θ_1과 θ_2로 구성된다. 그리고, 연결강도행렬 W는 w_{ij}들로 이루어지는데, 여기에서 w_{ij}는 i번째 입력과 j번째 시그모이드 함수 사이의 연결강도이다. 이 연결강도와 관련된 적격성(eligibility)은 e_{ij}로 표기하고, 감쇠 평균 적격성(discounted running average eligibility) 또는 적격성 트레이스(eligibility trace) 값은 D_{ij}로 표기한다. 출력 y_i는, 0 또는 1을 확률 f_i에 따라 랜덤하게 출력한다. 여기에서 f_i는 출력이 $y_i = 1$ 일 확률이며, 다음 식에 따라 정해진다:

$$f_j = 1 / (1 + e^{-(\theta_1 w_{1j} + \theta_2 w_{2j} + w_{3j})})$$

에이전트는 그림 2와 같은 확률적 제어입력선택 전략에 따라서 제어입력 a 를 생성한다. 그리고, 각 연결강도 w_{ij}가 갱신규칙[2]

2) 이 식에서 $\alpha > 0$는 학습율이고 b는 적절하게 선택된 reinforcement baseline임

$$\Delta w_{ij} = \alpha(r-b)\frac{\partial}{\partial w_{ij}}\ln(\pi(a, W, X))$$

에 따라 변할 경우에

$$E(\Delta w_{ij}|W) = \alpha\partial E(r|W)/\partial w_{ij}$$

가 성립함[6-7]에 기초한 이산 제어입력을 위한 SGA 알고리즘([1]의 그림 5 참조)을 약 20000회의 시간 스텝 동안 적용하면 로봇의 진행 속도가 꾸준히 증가하는 경향을 나타냄을 관찰할 수 있다. 그림 3는 [1]의 방법론을 따라 작성한 매트랩 프로그램이 제공한 실험 결과로써, 학습이 진행됨에 따라 로봇의 평균 진행 속도가 점차적으로 증가하는 패턴을 보여준다.

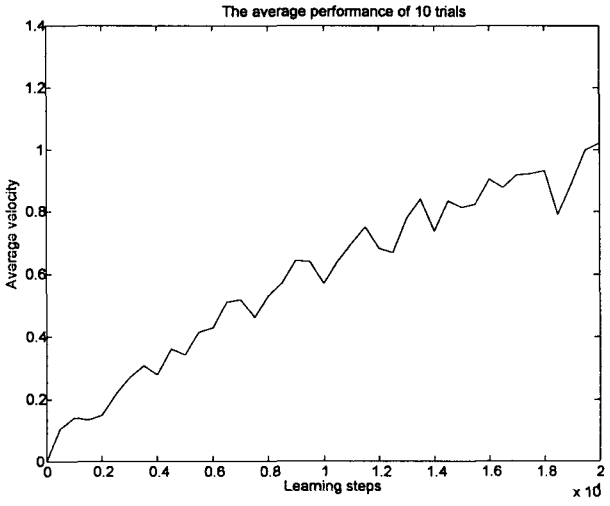

그림 3. 이산제어입력을 갖는 Kimura의 로봇을 SGA 기법으로 학습시킨 결과

3. 연속제어입력을 갖는 경우를 위한 SGA 학습 및 적용

[1]과 [7] 등에서 시도된 이산 제어입력을 갖는 경우를 위한 SGA 기법은, [8]에서 연속 제어입력을 갖는 경우를 위한 SGA 기법으로 확장되고 LQR (linear quadratic regulator) 문제 및 도립진자 제어 문제 등에 적용된 바 있다. 연속 제어입력을 갖는 경우를 위한 SGA 학습 기법은 다음과 같은 절차로 이루어진다 ([8]의 그림2 참조):

(1) 시간스텝 t 때의 관측변수 X_t 를 관찰함

(2) 제어입력 선택을 위한 확률분포 $\pi(a_t, W, X_t)$ 에 따라, 제어입력 a_t 를 샘플링하여 실행함

(3) 보상값 r_t 를 관찰함

(4) 다음과 같이 정의된 적격성[3] $e_i(t)$ 와 적격성 트레이스 $D_i(t)$ 를 계산함:

$$e_i(t) = \frac{\partial}{\partial w_i}\ln(\pi(a_t, W, X_t))$$

$$D_i(t) = e_i(t) + \gamma D_i(t-1)$$

(여기에서, $\gamma \in [0, 1)$ 는 감쇠율이고, w_i는 연결강도행렬 W 의 i 번째 원소임)

(5) 연결강도 갱신[4]: $\Delta w_i(t) = r_t D_i(t)$

(6) 제어입력선택전략 개선: $W \leftarrow \alpha(1-\gamma)\Delta W(t)$

(여기에서, $\alpha > 0$ 는 학습율임)

(7) 시간스텝을 $t+1$ 로 증가시키고, 단계 (1)로 되돌아감

본 논문에서는 [8]에서의 이론 전개를 참고하여,, 각 조인트의 제어입력 선택 전략을 위한 확률분포 π로 다음과 같은 정규분포를 고려하였다:

$$\pi(a, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-(a-\mu)^2}{2\sigma^2}\right)$$

따라서, 평균 μ와 σ의 적격성은 다음과 같아진다[8]:

$$e_\mu = \frac{a_t-\mu}{\sigma^2}, \quad e_\sigma = \frac{(a_t-\mu)^2-\sigma^2}{\sigma^3}$$

여기에서, σ가 0으로 접근하면 각 적격성이 큰 값으로 발산함에 유의하여 학습율을 σ^2에 비례하게 잡아주는 전략[8]을 취하면, 위의 적격성 e_μ와 e_σ는 각각 다음과 같이 조정된다:

$$e_\mu = a_t-\mu, \quad e_\sigma = \frac{(a_t-\mu)^2-\sigma^2}{\sigma}$$

그리고, 각 조인트에 대한 확률적 제어입력선택전략 π의 평균 μ와 σ를 각각

$$\mu = w_1\theta_1 + w_2\theta_2 + w_3, \quad \sigma = 0.1 + \frac{1}{1+\exp(-w_4)}$$

로 잡아주면 함수의 미분 및 연쇄법칙(chain rule)을 이용하여 첫 번째 조인트에 가해지는 입력과 관련된 연결강도의 적격성을 다음과 같이 구할 수 있다:

$$e_1 = e_\mu\frac{\partial}{\partial w_1}\mu = (a-\mu)\theta_1$$

$$e_2 = e_\mu\frac{\partial}{\partial w_2}\mu = (a-\mu)\theta_2$$

$$e_3 = e_\mu\frac{\partial}{\partial w_3}\mu = (a-\mu)$$

$$e_4 = e_\sigma\frac{\partial}{\partial w_4}\sigma = ((a-\mu)^2-\sigma^2)(1+0.1-\sigma)$$

3) 여기에서, 적격성은 적용된 제어입력 a_t과 연결강도 w_i사이의 연관성을 나타내는 척도가 됨

4) Reinforcement baseline b 값이 0임을 가정하였음

두 번째 조인트를 위한 제어입력 관련 연결강도의 적격성도 같은 방법으로 구해진다. 위의 식들에 등장하는 θ_1과 θ_2는, 2장에서와 비슷하게 각 조인트의 각도 변위가 [-1, 1] 범위가 되도록, 관찰된 각 조인트 각도를 적절하게 스케일링한 결과로 정의되는 관측변수이다. 그리고, 제어입력으로는 확률분포 π에 의해서 선택된 값을 사용하였고, 이산 제어입력 공간을 고려했던 [1]의 경우와 성능을 비교하는 경우를 위한 형평성 유지차원에서 각 조인트에는 각 시간 스텝 당 [-12도, +12도] 범위까지의 움직임만 허용하는 한계성을 부여했다. 학습에 사용된 그 밖의 관련 파라메터는 다음과 같다: 감쇠율 $\gamma = 0.9$, 학습율 $\alpha = 0.01$.

이상에서 설명한 연속 제어입력을 갖는 Kimura의 로봇에 SGA 학습기법을 적용하여 매트랩 시뮬레이션을 수행한 결과 그림 4와 같이 2장의 경우보다 우수한 성능이 관찰되었다. 또한 향후의 관련 연구를 위하여 매트랩 기반 로봇 애니메이션 프로그램이 개발되었는데, 이는 학습율 등의 각종 선택사항의 효과를 시각적으로 인지하는 데 큰 도움이 되었다.

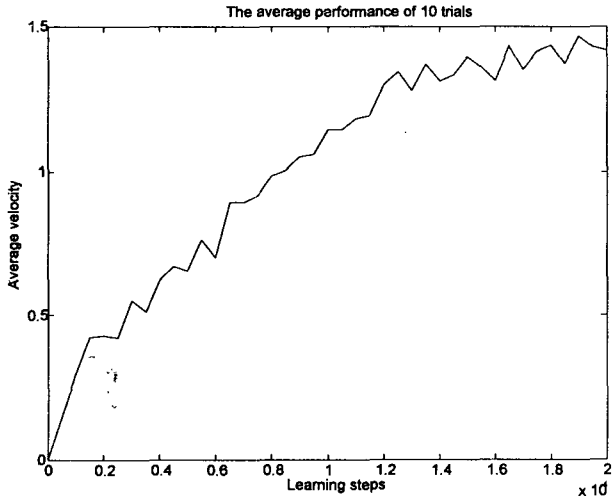

그림 4. 연속제어입력을 갖는 Kimura의 로봇을 SGA 기법으로 학습시킨 결과

4. 결론

본 논문에서는 Kimura의 로봇을 대상으로 하여 SGA 기법을 적용하는 문제를 고려해보았다. 매트랩을 이용하여 실험을 수행해 본 결과, 이 로봇의 제어에는 [1]에서 고려된 이산 제어 입력을 위한 SGA 기법뿐만 아니라, 연속 제어 입력을 위한 SGA 기법도 효과적으로 적용될 수 있음을 관찰하였다. 강화학습 분야에 여러 가지 흥미 있는 새로운 알고리즘이 꾸준히 제안되고 있는 현실을 생각할 때, 본 연구를 통해 확보된 Kimura의 로봇 시뮬레이터는 각종 강화 알고리즘의 효과를 관찰 또는 비교해볼 수 있는 좋은 도구가 될 것으로 기대된다. 향후에 시도해 볼만한 주요한 연구과제로는, 최근에 기계학습 분야에 큰 영향을 미치고 있는 커널 기법을 강화 학습 분야에 접목시키는 학습 알고리즘을 개발한 후 Kimura의 로봇에 적용시켜보는 문제 등을 들 수 있다.

참고문헌

[1] H. Kimura, K. Miyazaki, and S. Kobayashi, "Reinforcement learning in POMDPs with function approximation," In *Proceedings of the 14th International Conference on Machine Learning (ICML'97)*, pp. 152-160, 1997.

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

[3] S. J. Bradtke, B. E. Ydstie, and A. G. Barto, "Adaptive linear quadratic control using policy iteration," In *Proceedings of American Control Conference*, pp. 3475-3479, 1994.

[4] S. H. G. Hagen, *Continuous state space Q-learning for control of nonlinear systems*, PhD Thesis, University of Amsterdam, 2001.

[5] T. Landelius, *Reinforcement learning and distributed local model synthesis*, PhD Thesis, Linkoping University, 1997.

[6] R. J. Williams, "Simple statistical gradient following algorithms for connectionist reinforcement learning," *Machine Learning*, vol. 8, pp. 229-256, 1992.

[7] H. Kimura, M. Yamamura, and S. Kobayashi, "Reinforcement learning by stochastic hill climbing on discounted reward," In *Proceedings of the 12th International Conference on Machine Learning (ICML'95)*, pp. 295-303, 1995.

[8] H. Kimura and S. Kobayashi, "Reinforcement learning for continuous action using stochastic gradient ascent," In *Proceedings of the 5th International Conference on Intelligent Autonomous Systems (IAS-5)*, pp. 288-295, 1998.