Proceedings of KFIS Autumn Conference 2007 Vol. 17, No. 2.

강화학습 기법과 메타학습을 이용한 기는 로봇의 이동

Locomotion of Crawling Robots Based on Reinforcement Learning and Meta-Learning

문영준, 정규백, 박주영

고려대학교 제어계측공학과

E-mail: {dreamhill, qbaek, parkj} @ korea.ac.kr

요 약

최근 인공지능 분야에서는 강화학습(Reinforcement Learning)에 대한 관심이 크게 증폭되고 있으며, 여러 관련 분야에 적용되고 있다. 본 논문에서는 강화학습 기법 중 액터-크리틱 계열에 속하는 RLS-NAC 알고리즘을 활용하여 Kimura의 기는 로봇의 이동을 다룰 때에 중요 파라미터의 결정을 위하여 meta-learning 기법을 활용하는 방안에 고려한다.

Key Words : Reinforcement learning, Meta-learning, Locomotion

1. 서 론

최근 인공지능 분야에서는 강화학습(Reinforcement learning)에 대한 관심이 크게 증폭되고 있으며[1]-[2], 여러 관련 분야에 적용되고 있다 [3].

본 논문에서는 강화학습 기법 중 하나인 액터-크리틱 방법(Actor-Critic Methods)을 활용한다. 액터-크리틱 방법은 액션(Actions)을 선택하는 Policy structure인 액터 부분과 액터에 의해 선택된 액션을 평가하는 크리틱 부분으로 나누어져 있다. 본 연구팀에서 NAC(Natural Actor-Critic) 알고리즘을 개선하여 제안한 RLS-NAC(Recursive Least-Squares based Natural Actor-Critic) 알고리즘[4]을 가지고 Kimura의 기는 로봇[7]을 대상으로 학습이 성공적으로 이루어지게 하는데 큰 영향을 끼치는 학습율(learning rate, α), 할인율(discount factor, γ)에 주목한다. 지금까지 대부분의 강화학습을 적용한 분야에서는 학습율, 할인율을 고정된 값으로 사용하고 있으며, 이 값을 찾기 위해 수많은 반복 실험을 통해 경험적, 대략적으로 적절한 값을 찾아야 하는 번거로움이 있었다. 그래서 본 논문에서는 강화학습의 성능을 좌지우지하는 이 값들을 찾기 위한 방법으로 학습과정을 통해서 스스로 대략적으로 적절한 값을 찾아가는 방법을 제안한 [5],[6]의 방법론을 이용하여 Kimura의 기는 로봇에 적용하고 범용성을 판단한다.

본 논문의 구성은 다음과 같다: 우선 2장에서는 강화학습과 메타 파라미터(meta-parameter)의 중요성에 대해서 설명하고 3장에서는 이 메타 파라미터를 적응적으로 찾는 메타학습에 대한 소개, RLS-NAC를 적용한 Kimura의 기는 로봇에 대한 설명과 메타학습을 적용한 실험 결과에 대해 살펴 볼 것이다. 마지막으로 4장에서는 결론과 향후 연구 방향 등을 제시한다.

2. 강화학습과 메타 파라미터의 중요성

본 논문에서는 Markov property를 가지며 환경과 에이전트가 상호작용하는 MDP (Markov Decision Problem)에 대해 고려한다 [8].

강화학습은 학습을 통하여 얻게 되는 보상값(reward)의 총합이 최대가 되게 하는 것이 목표이다. 이를 평가하기 위해 향후 얻게 되는 보상값의 합의 기댓값인 상태 가치함수(state value function)를 고려하며 다음과 같이 정의한다.

$$V^{\pi}(s(t)) = E\left\{\sum_{k=0}^{\infty}\gamma^{k} r_{t+k} | s(t) = s, \pi\right\} \quad (1)$$

여기서, r_k는 상태 s_t에서 정책 π를 따를 때

얻게 되는 보상값이며, γ는 $0<\gamma<1$의 범위를 가지는 할인율(discount factor)이다. 할인율은 향후 얻게 될 보상값의 가치를 얼마만큼 둘 것인가를 표현하는 것으로, 만약 작은 값을 가지면 단지 즉시 얻게 되는 보상값만 고려하는 것이며, 1에 가까울수록 먼 미래에 얻게 될 보상값까지 고려한다. 주로 할인율은 거의 1에 가까운 값을 선택하면 가능하지만, 대부분의 시스템은 제한된 수명을 가지고 있으며, 확률적, 동적으로 변하는 시스템에 대해서는 향후 얻게 될 보상값의 신뢰성이 떨어지므로 이 값들을 선정함에 있어서 무조건 큰 값을 가지는 것만이 좋은 것은 아니다. 가치함수는 이전 상태와 변화 후의 상태에 대해서 식 (2)와 같이 회귀적으로 표현할 수 있다.

$$V(s(t))=E\{r(t)+\gamma V(s(t-1))\} \quad (2)$$

식 (2)로부터 TD(Temporal Difference) 에러를 다음과 같이 정의한다.

$$\delta(t)=r(t)+\gamma V(s(t))-V(s(t-1)) \quad (3)$$

식 (3)은 평균적으로 0이 되어야 하고, 식 (4)와 같이 가치함수를 갱신할 수 있다.

$$V(s(t))\leftarrow V(s(t))+\alpha(r(t)+\gamma V(s(t+1))-V(s(t))) \quad (4)$$

α는 학습율(learning rate)이며, $0\leq\alpha\leq1$의 범위를 가지고 학습 속도에 영향을 끼치는 파라미터이다. 만약 너무 작은 값을 가지면, 학습이 천천히 진행되고, 너무 큰 값을 가지면 학습이 제대로 이루어지지 않게 된다.

이 중요 파라미터(학습율, 할인율)를 메타 파라미터(Meta-parameter)라 부르고 학습을 통해 찾는 방법인 메타학습(Meta-learning)은 다음 장에서 설명하겠다.

3. 메타학습을 기는 로봇에 적용한 실험 및 결과

이전 장에서 설명하였듯이, 메타 파라미터는 강화학습의 성능에 지대한 영향을 끼치므로 값 선택에 신중을 요하게 된다. 이런 파라미터는 주로 경험적으로 찾아왔던 것을 본 논문에서는 메타학습을 통해 적절한 값을 찾아보는 것에 주목한다. 메타학습에 대한 기본 개념과 상세한 설명은 [5],[6]을 참조하기 바란다.

메타학습은 시스템의 상태변화에 따라 적절한 액션을 선택함으로써 얻게 되는 보상값과 메타 파라미터가 서로 영향을 주고받으면서 학습이 진행되는 동안 적절한 값을 찾아 가는 방법이다. 알고리즘은 아래의 표 1에 나타난 것과 같다. 아래의 표에서는 할인율(γ)에 대한 메타 학습을 표현했지만, 학습율(α)에 대해서도 가능하다. 또한 강화학습에 사용되는 기본 알고리즘과도 서로 영향을 끼치지 않고 독립적으로 적용된다 [6].

표 1. 메타학습 알고리즘[6].

① n-step 동안의 short-term과 long-term의 보상값 합의 평균을 얻는다.

$$\Delta\bar{r}(t)=\frac{1}{\tau_1}\left(-\bar{r}(t)+r(t)\right)$$

$$\Delta\bar{\bar{r}}(t)=\frac{1}{\tau_2}\left(-\bar{\bar{r}}(t)+\bar{r}(t)\right)$$

τ_1,τ_2 : *time constants*

② n-step마다 short-term과 long-term의 차와 perturbation으로 γ_{b_0}를 갱신한다.

$$\Delta\gamma_{b_0}=\mu\left(\bar{r}(t)-\bar{\bar{r}}(t)\right)\sigma_\gamma(t)$$

③ gamma 뉴런을 갱신한다.

$\gamma_b(t)=\gamma_{b_0}+\sigma_\gamma(t)$

γ_{b_0} : 평균

σ_γ : 평균이 0이고 분산이 ν인 가우시안 분포를 따르는 노이즈 부분

④ γ를 갱신한다.

$$\gamma(t)=1-\frac{1}{e^{\gamma_b(t)}}$$

위의 방법론을 RLS-NAC 알고리즘과 연속 제어입력을 고려하는 Kimura의 기는 로봇에 적용한다. 기는 로봇에 대한 자세한 설명은 참고 문헌 [7]을 참조하기 바라며, 여기에서는 간단히 소개한다. 아래 그림 1과 같이 중력이 작용하는 공간에서 움직이는 두 개의 링크를 가진 평면형 머니퓰레이터로써, 에이전트는 로봇 및 환경에 대한 구체적인 정보 없이 직접 경험을 통해서 이 로봇에 주어진 임무인 가능한한 최대한 빠르게 앞으로 전진하는 것을 이행하도록 하는 것이다. 보상값은 각 스텝당 전진한 거리로 정의하고, 각 조인트의 움직일 수 있는 범위가 제한되며, 이 상태에 따라 각각의 제어입력을 가한다.

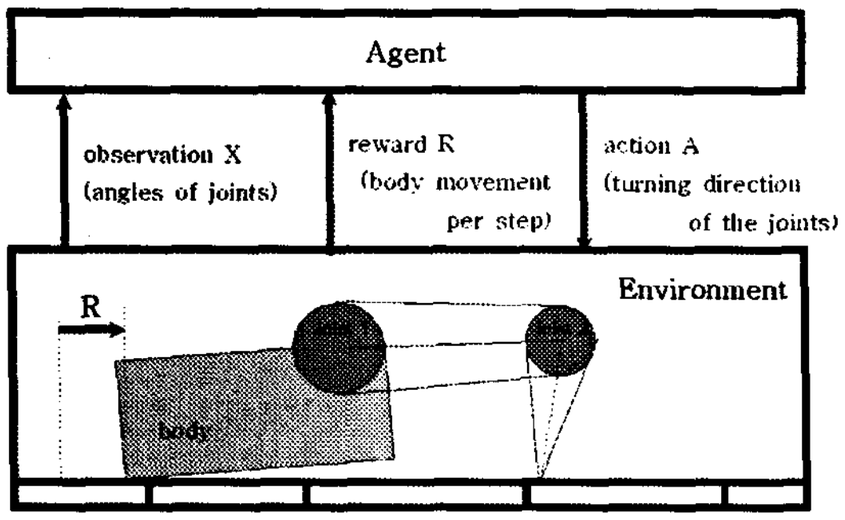

그림 1. Kimura의 기는 로봇[7].

그리고 연속 상태와 연속 제어입력을 고려한다[9]. 할인율은 대부분의 시스템에서 대체로 1 근처의 값으로 선택하면 학습이 충분히 이루어진다. 그러나 학습율은 시스템에 따라 값이 다양하게 바뀌고 선정이 어렵기 때문에 학습율에 대해서만 이 방법론의 성능을 관찰하며 유용성을 판단한다. 실험은 거의 0에 가까운 초기 학습율을 가진 조건하에서 수행한 실험과 고정된 학습율과 메타학습을 사용한 환경에 따라 적응적으로 변하는 학습율을 비교하는 실험을 수행하였다. 그리고 실험에서 사용한 초기 파라미터는 표 2에 나타내었다.

표 2. 메타학습에 필요한 초기 파라미터들.

파라미터	값
Time constant τ_1,τ_2	100
Step n	200
Learning rate μ	2
Std. $\sqrt{\nu}$	0.005

3. 1 초기 학습율=0.0001일 때의 메타학습

그림 2는 초기 학습율을 0.0001에서 시작하여 스텝=10000까지 실험을 수행하여 보상값의 변화에 따라 학습율의 변화를 관찰한 것이다.

상단에서부터 첫 번째 그래프는 short-term과 long-term의 변화, 두 번째는 보상값의 변화에 따른 학습율의 변화를 보여준다. 메타학습의 방법론과 그림 2에서 볼 수 있듯이 short-term과 long-term의 차와 perturbation의 방향에 따라 학습율이 변한다. 만약 perturbation이 방향이 (-)를 취해서 보상값이 줄어들면(-), perturbation의 반대 방향 (+)으로 학습율이 변한다. 이와 같은 방법으로 실험을 수행한 결과 대략 0.005~0.006정도에서 수렴함을 볼 수 있었다.

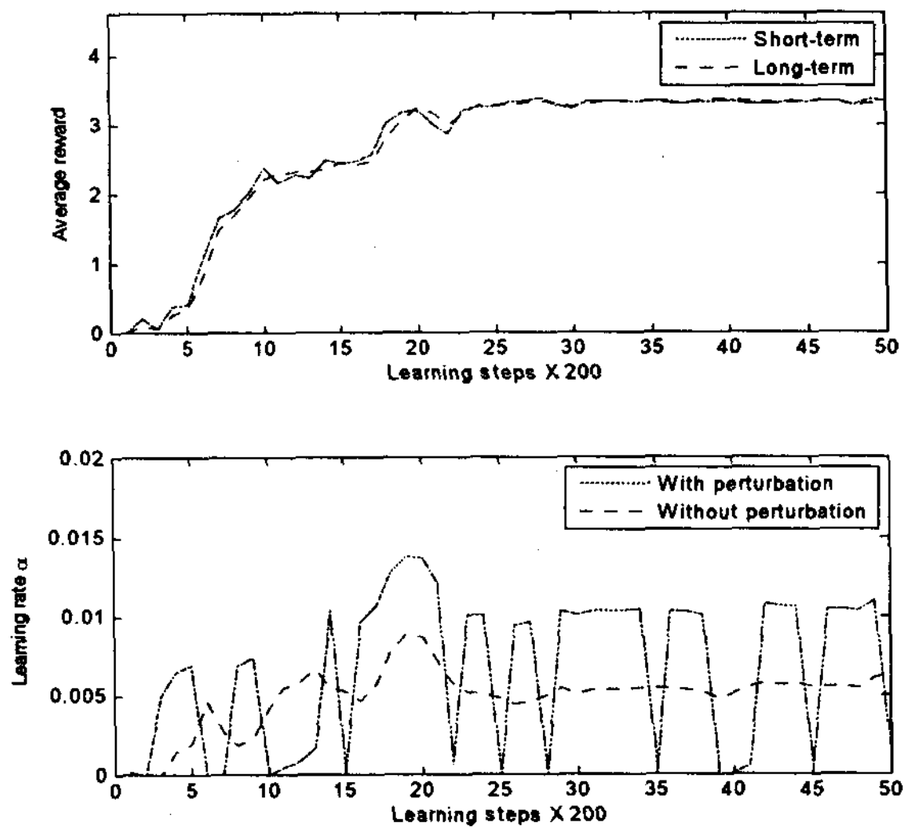

그림 2. Short-term, Long-term과 할인율의 변화.

3. 2 고정된 학습율과의 성능 비교

지금까지 메타학습을 적용시켜 학습율이 환경에 따라 적응적으로 변하면서 학습이 성공적으로 이루어지는 것을 보았다. 이번 실험은 위의 실험을 통해 얻은 결과와 현재까지 해왔듯이 사용자가 임의의 값으로 고정된 학습율을 사용한 실험을 수행하여 성능을 비교한다.

그림 3은 임의의 고정된 학습율과 적응적으로 변하는 학습율을 적용시켜 보상값의 합을 200 스텝마다 얻은 결과이다.

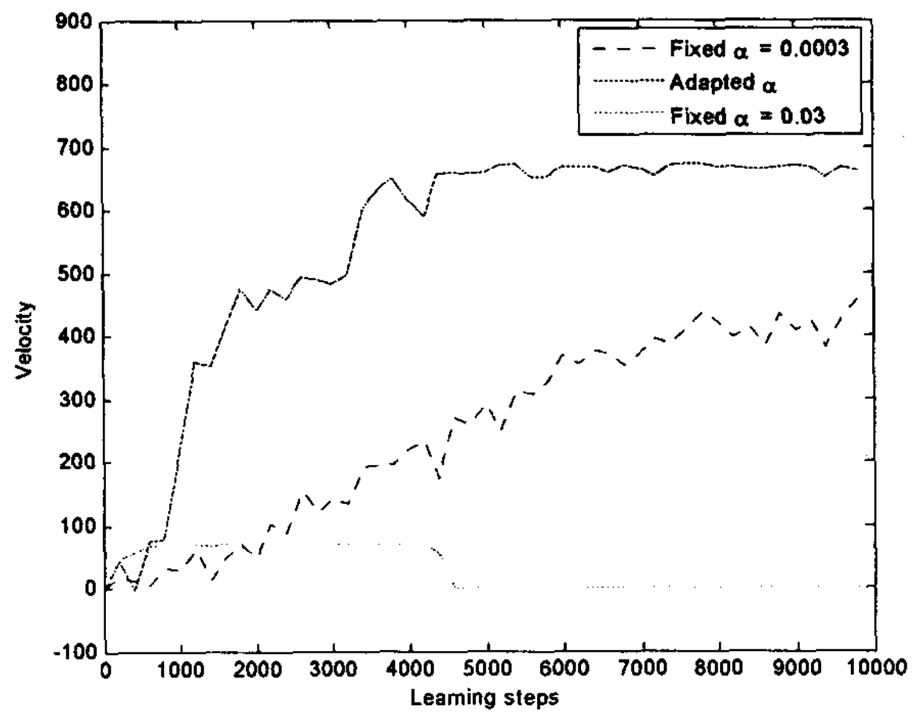

그림 3. 학습율에 따른 속도의 변화.

위 그래프와 같이 시스템마다 어느 정도는 차이가 있을 수 있지만, 학습율에 따라 학습 성능이 현저히 차이가 있음을 볼 수 있다. 너무 작은 값이 선택했을 때는 성능이 떨어지고 너무 큰 값은 학습이 실패하였다. 이 결과를 통해 사용자는 학습율 선정에 있어서 많은 실험을 통해서 선택해야 하는 어려움이 따를 수

있음을 볼 수 있었으며, 메타학습의 타당성을 검증해 볼 수 있었다.

4. 결론 및 향후 연구방향

본 논문에서는 연속 상태와 연속 제어입력을 고려한 Kimura의 기는 로봇을 대상으로 학습의 성능에 크게 영향을 끼치는 중요한 메타파라미터에 관심을 가지며, 학습이 성공적으로 이루어질 수 있는 적절한 값을 찾기 위해 [6]에서 제안한 방법론인 메타학습(Meta-learning)을 적용시켜 학습을 수행하였다. 학습이 성공적으로 이루어지면서 학습율(learning rate) 또한 적절한 값으로 수렴하는 것을 볼 수 있었다.

향후 본 연구실에서는 다양한 모델을 선정하여 강화학습을 적용시켜 봄과 동시에 지금까지 경험적으로 수많은 반복 실험을 통해 찾아야 했던 번거로움을 덜 수 있도록 메타학습 방법을 적용시켜 범용성을 판단하면서 선행연구를 이어가겠다.

참 고 문 헌

[1] 한림 심포지엄 논문집 20, 2006 KAST International Symposium on Learning from the Perspective of Natural Science, Social Science and Engineering, Nov. 30 - Dec.01, 2006, 서울.

[2] Proceedings of 2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning, April 1-5, 2007, Honolulu, Hawaii, USA.

[3] Reinforcement Learning Repository at University of Massachusetts, Amherst, http://www-anw.cs.umass.edu/rlr/.

[4] J. Park, J. Kim, and D. Kang, "An RLS-based natural actor-critic algorithm for locomotion of a two-linked robot arm," Lecture Notes in Artificial Intelligence, vol. 3801, pp. 65-72, December, 2005.

[5] K. Doya, "Metalearning and Neuromodulation", Neural Networks, vol. 15, no. 4, pp. 495-506, June, 2002.

[6] N. Schweighofer, K. Doya, "Meta-learning in reinforcement learning", Neural Networks, vol. 16, no. 1, pp. 5-9, Jan, 2003.

[7] 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.

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

[9] 박주영, 정규백, 문영준. "강화학습에 의해 학습된 기는 로봇의 성능 비교", 한국퍼지 및 지능시스템학회 논문집, 17권, 1호, pp. 33-36, 2007.