Proceedings of KFIS Spring Conference 2007 Vol. 17, No. 1.

강화학습에 의해 학습된 기는 로봇의 성능 비교

Performance Comparison of Crawling Robots Trained by Reinforcement Learning Methods

박주영, 정규백, 문영준

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

요 약

최근에 인공지능 분야에서는, 국내외적으로 강화학습(reinforcement learning)에 관한 관심이 크게 증폭되고 있다. 강화학습의 최근 경향을 살펴보면, 크게 가치함수를 직접 활용하는 방법(value function-based methods), 제어 전략에 대한 탐색을 활용하는 방법(policy search methods), 그리고 액터-크리틱 방법(actor-critic methods)의 세가지 방향으로 발전하고 있음을 알 수 있다. 본 논문에서는 이중 세 번째 부류인 액터-크리틱 방법 중 NAC(natural actor-critic) 기법의 한 종류인 RLS-NAC(recursive least-squares based natural actor-critic) 알고리즘을 다양한 트레이스 감쇠계수를 사용하여 연속제어입력(real-valued control inputs)으로 제어되는 Kimura의 기는 로봇에 대해 적용해보고, 그 성능을 기존의 SGA(stochastic gradient ascent) 알고리즘을 이용하여 학습한 경우와 비교해보도록 한다.

Key Words : 강화학습, RLS, NAC, SGA, Kimura의 기는 로봇

1. 서 론

최근에 인공지능 분야에서는, 국내외적으로 강화학습(reinforcement learning)에 관한 관심이 크게 증폭되고 있다[1]-[2]. 널리 알려진 바와 같이, 강화학습은 시스템의 상태(state) 및 제어입력(action 또는 control input) 공간이 이산 집합(discrete set)일 때에는 확고한 이론적 기초가 확립되어 있다[3]-[4]. 반면에 주어진 시스템이 연속 상태 및 연속 제어입력를 고려하는 경우에는 극히 제한된 부류의 문제[5]-[7]에 대해서만 엄밀한 증명이 제공되고 있는 형편이다. 하지만 이러한 이론적 엄밀성의 결여에도 불구하고 간단한 게임으로부터 복잡한 로봇 시스템의 제어에 이르기까지, 강화학습을 성공적으로 응용한 예는 최근 들어 날로 증가하는 추세이다[8].

강화학습의 최근 경향을 살펴보면, 크게 가치함수를 직접 활용하는 방법(value function-based methods), 제어 전략에 대한 탐색을 활용하는 방법(policy search methods), 그리고 가치함수와 제어전략 탐색을 위하여 분리된 모듈을 사용하되 학습과정에서 이들을 종합적으로 활용하는 액터-크리틱 방법(actor-critic methods)의 세가지 방향으로 발전하고 있음을 알 수 있다. 이중 세 번째 부류인 액터-크리틱 방법 중 한 기법인 NAC(natural actor-critic) 알고리즘이 본 논문에서 주목하고 있는 대상인데, 여기에서는 NAC 알고리즘을 개선하는 취지로 최근에 본 연구팀에 의해서 제안된 바 있는 RLS-NAC 알고리즘[9]의 범용성을 확인하는 차원에서, [9], [13]-[15] 등의 선행연구에서 미처 수행되지 못하고 미루어 두었던 연속제어입력이 사용되는 경우에 대한 트레이스 감쇠계수 변화 대비 성능비교 고찰을 수행해보고자 한다. 즉, 본 논문에서는 연속제어입력(real-valued control inputs)으로 제어되는 Kimura의 기는 로봇[10]을 대상으로 다양한 트레이스 감쇠계수를 사용하여 RLS-NAC(recursive least-squares based natural actor-critic) 알고리즘을 적용해보고, 그 성능을 기존의 SGA(stochastic gradient ascent) 알고리즘[10]-[12]을 이용하여 학습한 경우와 비교해보도록 한다.

본 논문의 구성은 다음과 같다: 우선 2장에서는 본 논문의 주요 소재가 되는 Kimura의 기는 로봇[10]에 대하여 간단히 설명한다. 그리고, 3장에서는 제어입력이 실수 범위에서 연속적인 값을 취하는 경우를 위한 로봇 제어 문제에 RLS-NAC[9] 알고리즘을 적용하는 단계를 간단히 소개한 후, 시뮬레이션을 통해 얻어진 성능 비교를 SGA[10-12] 알고리즘을 적용한 경우와 비교하여 보고한다. 마지막으

로 4장에서는, 결론과 향후 연구 방향 등을 제시한다.

2. Kimura의 기는 로봇

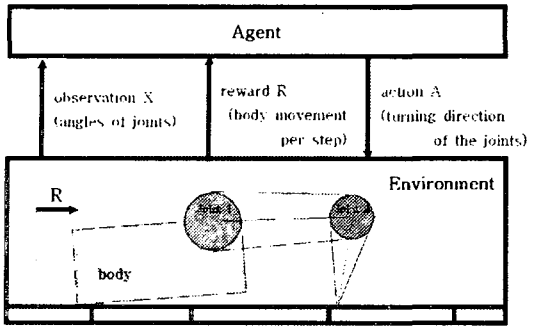

그림 1. Kimura의 기는 로봇[10]

참고문헌 [10]에서 Kimura 등은 강화학습의 효용성을 보이기 위해 간단한 기는 로봇을 응용 문제로 고려한 바 있다. 이 로봇은 적용되고 있는 알고리즘의 효용을 시각적으로 즉각적으로 확인해볼 수 있는 효과적인 실험 대상이므로, 강화학습 관련 연구에서 자주 활용되곤 한다[9], [10], [13]-[15]. Kimura의 기는 로봇의 각 요소에 대한 설명은 다음과 같다[10], [13]: 이 로봇은 두 개의 링크를 가지고 기는 동작을 수행하는 평면형 머니퓰레이터(planar manipulator)로써 그림 1의 구조를 갖는다. 이 로봇에 부과된 임무는 최대한 빠른 속도로 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 방향과 크기를 결정한다. 그리고, 학습 과정에서 이용되는 보상값 r 을 위해서는 해당 시간 스텝 동안 전진한 거리가 사용된다.

본 논문에서 고려하는 로봇 관련 데이터는 [10], [13]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. Kimura의 기는 로봇에 대한 매트랩 기반 시뮬레이터는 [13]-[15] 등에서 소개된 바 있다.

3. 다양한 트레이스 감쇠계수를 사용한 RLS-NAC 알고리즘의 적용 및 성능 비교

본 논문에서 고려하는 RLS-NAC 알고리즘은, 본 논문에 앞서 작성된 바 있는 [9]에서 설명된 대로 다음과 같이 요약될 수 있다(각 용어에 대한 정의 및 상세한 설명을 위해서는 [9], [15] 등을 참조하기 바람):

알고리즘의 적용을 위해 미리 준비할 내용:

- 초기 상태 s_0
- 제어전략 $\pi_\theta(a|s)$과 초기 파라메터 $\theta = \theta_0$, 그리고 관련 미분 벡터 $\nabla_\theta \log \pi_\theta(a|s)$
- 상태가치함수(state value function) 근사기 $\tilde{V}_v(s) = \phi(s)^T v$에 사용하는 기저함수 $\phi(s) = [\phi_1(s), \cdots, \phi_K(s)]^T$
- 액터 파라메터 θ 갱신 때의 학습율 $\alpha > 0$
- 망각계수(forgetting factor) $\beta \in (0,1)$
- 할인율(discount rate) $\gamma \in (0,1)$
- 트레이스 감쇠계수(trace-decay parameter) $\lambda \in [0,1]$
- 행렬 P_0를 가역으로 만들기 위한 상수 $\delta > 0$
- 각 액터 파라메터 크기를 한정시키기 위한 $M > 0$

알고리즘 적용을 통해 달성하고자 하는 목표:

- 제어전략 $\pi_\theta(a|s)$의 파라메터 벡터 θ를 위한 최적해 발견
- 상태가치함수 근사기 $\tilde{V}_v$와 우월가치함수 근사기(advantage value function approximator) $\tilde{A}_w(s,a) = \nabla_\theta \log \pi_\theta(a|s)^T w$의 파라메터 벡터 v와 w를 위한 최적해 발견

알고리즘:

for $t := 0, 1, 2, \cdots$ do

- 제어전략을 위한 확률분포 $\pi_{\theta_t}(\cdot \mid s_t)$로부터 제어입력 a_t를 추출함

- a_t를 적용한 후 보상값(reward) r_t와 다음 상태 s_{t+1}를 관찰함

- w_t와 v_t를 갱신하기 위해 [9]의 (5)와 (6)식에 나와 있는 RLS 규칙을 적용함.

- 제어전략을 위한 확률분포의 파라메터 벡터를 $\theta_{t+1} = \theta_t + \alpha w_t$를 이용하여 갱신함.

- θ_{t+1}의 원소값의 절대값이 M을 초과하는 경우에는 M으로 한정시켜 줌.

end

본 논문에서는 이산제어입력 확률분포를 고려한 이전 논문 [9]와 달리, 정규분포로 표현되는 연속제어입력 확률분포를 고려하였다. 즉, 기는 로봇의 각 조인트를 위한 제어입력 선택 확률분포 π로 다음과 같은 정규분포를 고려하였다:

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

그리고, 첫 번째 조인트를 위한 π의 평균 μ와 σ를 각각

$$\mu = \theta_1 s_1 + \theta_2 s_2 + \theta_3 ,$$

$$\sigma = 0.1 + \frac{1}{1+\exp(-\theta_4)}$$

로, 그리고 두 번째 조인트를 위한 π의 평균 μ와 σ를 각각

$$\mu = \theta_5 s_1 + \theta_6 s_2 + \theta_7 ,$$

$$\sigma = 0.1 + \frac{1}{1+\exp(-\theta_8)}$$

로 정하였다. 따라서, 액터를 위한 확률분포의 파라메터 개수는 총 8개가 된다. 그리고, 위의 정의에서 등장하는 s_1과 s_2는 각 조인트의 각도 변위가 [-1, 1] 범위가 되도록, 관찰된 각 조인트 각도를 적절하게 구간 선형함수를 이용하여 스케일링한 결과로 정의되는 상태변수이다. 제어입력으로는 확률분포 π에 의해서 선택된 값을 사용하였고, 이산 제어입력 공간을 고려했던 [9]의 경우와 유사하게 각 조인트에는 각 시간 스텝 당 [-20도, +20도] 범위까지의 움직임만 허용하는 한계성을 부여했다. 기저함수 ϕ를 위해서는 항등함수가 사용되었고, 초기 상태 s_0로는 로봇이 수평하게 서있는 자세를 사용하였다. 트레이스 감쇠계수 λ를 0.0에서 1.0까지 0.1 간격으로 변화시키면서, $t = 10000$ 시점 때까지 학습을 실행하여 기는 로봇의 속도를 구하는 실험을 5회 실행한 후 이 결과를 그림 2에 정리하였다. 이 그림에서 위쪽과 아래쪽 경계부분은 각각 5회중 최대 속도값과 최소 속도값을 나타내고, 네모로 표시된 부분은 5회 실험 전체의 평균 속도 값을 보여준다. 그림의 내용으로부터 알 수 있듯이, λ 값이 0.1 이하일 때에는 평균적으로 속도가 느릴 뿐만 아니라 속도 값들의 분산의 폭도 커지는 좋지 않은 결과가 관찰되고 있다. 반면에, λ 값이 0.2부터 1.0 구간에서 움직일 때에는 로봇의 이동속도가 적은 폭의 분산을 가지고 평균적으로 높은 값을 가짐을 보여준다. 특히, $\lambda \approx 1.0$일 때에는 매 실험 때마다 거의 같은 결과가 얻어지는 일관성이 관찰되었다.

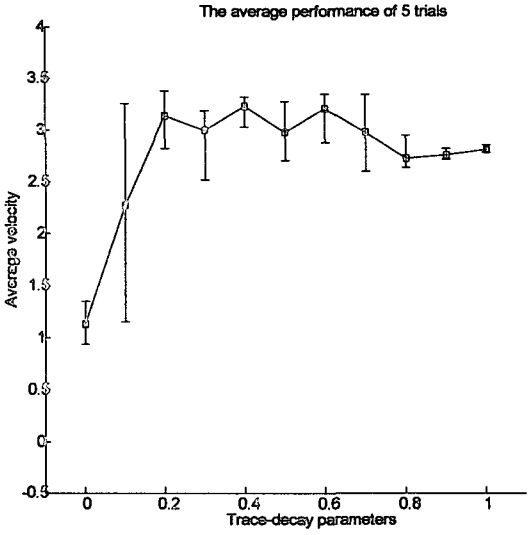

그림 2. 다양한 트레이스 감쇠계수를 사용하여 RLS-NAC 알고리즘을 적용한 결과

성능비교의 목적으로, 트레이스 감쇠계수를 0.0~1.0 범위에서 0.1 간격으로 변화시키면서 SGA 알고리즘 [10], [13]을 적용한 경우에 대해서도 같은 종류의 실험을 수행한 후, 그 결과를 그림 3에 요약하였다. 그림으로부터 관찰할 수 있듯이, SGA 알고리즘을 이용하여 학습한 결과는 λ 값에 거의 무관하게 일정한 성능을 나타내는 특징을 가지며, 평균 속도 측면에서는 RLS-NAC 알고리즘을 사용한 경우보다 상당히 뒤떨어지는 결과가 얻어짐을 알 수 있다. 따라서, 트레이스 감쇠계수 값을 위하여 지나치게 작은 값을 사용하는 상황을 제외하고는, RLS-NAC 알고리즘이 SGA 알고리즘보다 Kimura의 기는 로봇이라는 응용 도메인에 대해서는 더 우수한 성능을 제공함을 결론지을 수 있다.

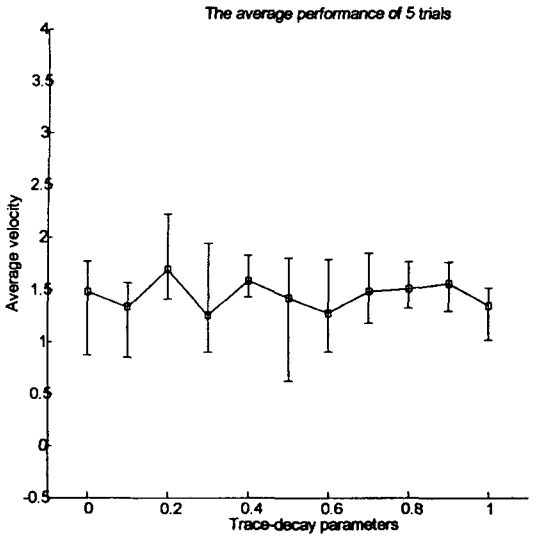

그림 3. 다양한 트레이스 감쇠계수를 사용하여 SGA 알고리즘을 적용한 결과

4. 결론 및 향후 연구방향

본 논문에서는 연속제어입력으로 제어되는 Kimura의 기는 로봇을 대상으로, 다양한 트레이스 감쇠계수를 사용하여 RLS-NAC 기법을 적용해보고 그 성능을 SGA 기법을 적용하는 경우와 비교해보았다. 매트랩을 이용하여 실험을 수행해 본 결과, 이 로봇의 제어에는 RLS-NAC 기법과 SGA 기법이 모두 효과적으로 적용될 수 있으며 RLS-NAC 기법이 보다 더 우수한 학습 성과를 제공함을 확인할 수 있었다. 다음 단계의 연구과제로 생각해볼 수 있는 향후 연구주제로는, 커널 기반 함수근사 기법(kernel-based function approximation methods)을 NAC 알고리즘에 접목시켜서 Kimura의 기는 로봇에 적용시켜보는 문제 등을 들 수 있다.

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] http://liu.ece.uic.edu/ADPRL07, 2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning, April 1-5, 2007, Honolulu, Hawaii, USA.

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

[4] A. Gosavi, Simulation-Based Optimization: Parametric Optimization Techniques and Reinforcement Learning, Kluwer Academic Publishers, 2003.

[5] 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.

[6] S. H. G. Hagen, Continuous State Space Q-learning for Control of Nonlinear Systems, PhD Thesis, University of Amsterdam, 2001.

[7] T. Landelius, Reinforcement Learning and Distributed Local Model Synthesis, PhD Thesis, Linkoping University, 1997.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 박주영, 김종호, 신호근, "SGA 기반 강화학습 알고리즘을 이용한 로봇제어", 한국퍼지및지능시스템학회 2004년도 추계학술대회 논문집, 14권 2호, pp.63-66, 2004년 10월.

[14] 김종호, 강대성, 박주영, "RPO 기반 강화학습 알고리즘을 이용한 로봇제어", 한국퍼지및지능시스템학회 논문지, 15권 4호, pp. 505-510, 2005년 8월.

[15] 김종호, 강대성, 박주영, "RLS 기반 Actor-Critic 학습을 이용한 로봇 이동", 한국퍼지및지능시스템학회 논문지, 15권 6호, pp. 88-93, 2005년 12월.