Acrobot Swing Up 제어를 위한 Credit-Assigned-CMAC 기반의 강화학습

Credit-Assigned-CMAC-based Reinforcement Learning with application to the Acrobot Swing Up Control Problem

신 연 용, 장 시 영, 서 승 환, 서 일 홍
(Yeon Yong Shin, Si Young Jang, Seung Hwan Seo, and Il Hong Suh)

*한양대학교 전자전기제어계측공학과 전화:(031)408-5802, 팩스:(031)408-5803, E-mail : ihsuh@hanyang.ac.kr)

Abstract : For real world applications of reinforcement learning techiniques, function approximation or generalization will be required to avoid curse of dimensionality. For this, an improved functio approximation-based reinforcement learning method is proposed to speed up convergence by using CA-CMAC(Credit-Assigned Cerebellar Model Articulation Controller). To show that our proposed CACRL(CA-CMAC-based Reinforcement Learning) performs better than the CRL(CMAC- based Reinforcement Learnning), computer simulation results are illustrated, where a swing-up control problem of an acrobot is considered.

Keywords : CMAC, Credit-Assigned, Function Approximation, Reinforcement Learning, Acrobot

Ⅰ. 서론

Acrobot은 비선형성을 가지면서 말단장치(end-effector)의 위치제어를 위해 필요한 모터의 최소 개수(자유도)보다 모터의 개수가 부족한 underactuated 시스템이다. 그림 1.1에 Acrobot의 구조를 나타내었으며, 모터는 두 번째 조인트에만 연결되어 있다. Acrobot이라는 이름과 연구는 1991년 Murray 와 Hauser의 underactuated mechaniscal system 연구[1]로부터 시작하였으며, 그 후에도 Acrobot 제어를 위한 많은 연구가 진행되었다.

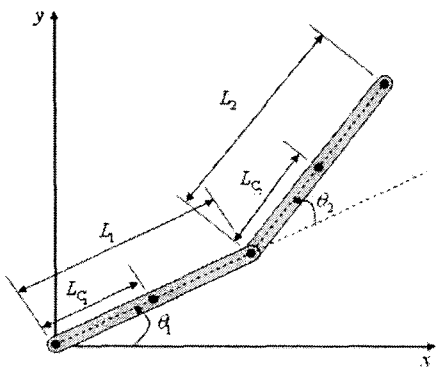

그림 1.1 Acrobot의 구조
Fig. 1.1 Structure of the Acrobot

Acrobot의 제어목적에는 말단장치를 일정 위치 이상으로 올리는 swing-up 제어와 inverted pendulum의 제어와 유사하게 두개의 링크를 수직으로 세워서 균형을 잡는 balancing 제어가 있다. Acrobot 제어를 위한 대표적인 연구로는 M. W. Spong의 partial feedback linearization[3]을 이용한 swing-up 제어 방법이 있으며, 그 후 이를 개선한 A. D. Luca, G. Oriolo의 Iterative State Steering 기법을 이용한 robust feedback control 방법[8]이 있다. swing-up 제어를 위한 또 다른 방법으로는 G. Boone의 Acrobot의 운동방정식과 energy 방정식을 이용한 energy based 제어방법[5], S. C. Brown 과 K. M. Passino의 adaptive fuzzy control 방법을 이용한 balancing 제어 방법[9] 등이 있다.

Acrobot에 대한 모델을 알고, 이 모델에 기반하여 제어하는 Model-based 방법과는 달리 Acrobot에 대한 모델을 알 수 없다는 가정 하에서 이를 학습을 통하여 배워나가며 제어하는 Model-free 방법으로는 R. S. Sutton의 Sparse Coarse Coding 기법을 이용한 강화학습 방법[6]이 있다.

강화학습(Reinforcement Learning)이란 로봇이 미지의 환경에서 행동과 보답을 주고 받으며, 임의의 상태에서 가장 적합한 행위를 학습하는 방법이다. 강화학습을 이용하여 주어진 문제를 해결하기 위해서는 1)상태공간과 행위공간의 정의, 2)임의의 상태에서 적절한 행위를 선택하기 위한 행위책략의 정의, 그리고, 3)환경에서 받은 보답으로부터 행위함수를 학습시키는 방법의 정의가 필요하다.

강화학습을 실제환경에 적용 시 문제점으로는 연속 상태공간을 불연속 상태공간으로 표현하면서 생기는 많은 용량의 메모리(curse of dimensionality)와 이에 따른 긴 학습시간이 필요하다는 것이다. 이를 해결하기 위한 방법으로 함수 근사화(Function Approximation) 또는 일반화(Generalization)가 있으며, 이 방법은 연속 상태공간 상의 임의의 상태를 불연속 상태공간 상의 상태들의 조합으로 근사화하는 방법이다. 함수 근사화에 대한 연구로는 불연속 상태공간 상의 상태들의 행위함수의 선형 벡터 조합으로 실제 상태를 표현하는 Region-based Q-Learning 방법[11], 실제 상태의 행위함수를 불연속 상태들의 행위함수들의 Local Average 기법을 이용하여 계산하는 Kernel-based Reinforcement Learning 방법[12], 그리고, Sutton의 Sparse Coarse Coding을 이용한 강화학습방법 등이 있다. Sutton은 J. A. Boyan과 A. W. Moore가 강화학습에는 함수 근사화를 적용하기 어렵다고 주장한 논문[10]을 반박하면서 CMAC(cerebellar model articulation controllers)과 같은 구조를 이용한다면 강화학습의 상태공간을 함수 근사화하는게 가능하다고 주장하고, CRL(CMAC-based 강화학습)방법을 제안하였다. 함수 근사화를 이용한 강화학습에 관한 많은 연구들이 진행되어 왔음에도 불구하고, 실제 로봇에 강화학습을 적용하기에는 긴 수렴시간이 걸리며, 이를 줄이기 위한 연구는 강화학습이 실제 로봇 속에서 자리매김하는데 필수적이라 할 수 있다.

본 논문에서는 Sutton의 CRL 방법의 수렴속도를 개선하기 위한 연구로 S. Su, T. Tad, 그리고, T. Hung이 제안한 CA-CMAC(Credit-Assigned CMAC) 기법[7]을 이용한 CACRL (Credit-Assigned CMAC-based 강화학습) 방법을 제안하고, 그 수렴속도 개선 성능을 CRL 방법과 비교하였으며, 마지막으로 그 유용성을 Acrobot을 적용한 모의실험을 통하여 보였다.

II장에서 Credit-Assigned CMAC 기법과 CACRL 방법에 대해 설명하였다. III장에서는 CRL과 CACRL, 두 방법의 수렴속도 개선 성능을 Acrobot의 swing-up 제어 모의실험를 통하여 비교하고, 제안하는 CACRL 방법의 유용성을 검증하였다. 마지막으로 IV장에서 논문의 결론을 맺었다.

II. CA-CMAC and CACRL

1. CA-CMAC(Credit-Assigned CMAC)

CA-CMAC 방법은 S. Su, T. Tao, 그리고 T. Hung에 의해 제안되었다. S. Su 등은 CMAC의 수렴속도가 지연되는 이유 중의 하나를 임의의 상태에 연관된 타일들의 데이터 갱신 값을 동등하게 두는데 있다고 생각하였다. 연관된 타일 중 많은 학습을 통하여 신뢰도가 높은 타일도 있을 수 있고, 반대로 신뢰도가 낮은 타일도 있을 경우, 현재의 출력 오차로 부터의 각 타일의 갱신 비율은 신뢰도에 따라 차등을 두어야함에도 불구하고, 신뢰도가 낮은 타일의 영향으로 인한 현재의 출력오차를 동등한 비율로 신뢰도가 높은 타일의 갱신에 적용함으로써 CMAC내의 데이터들의 비효율적인 학습으로 인한 수렴지연현상이 일어 날 수 있다. 이를 해결하기 위하여 S. Su등은 다음과 같은 가정 하에 CA-CMAC 방법을 제안하였다.

> **가정:**
>
> 많은 학습을 한 타일 내의 데이터 일수록 더욱 정확한 데이터를 가지고 있다.

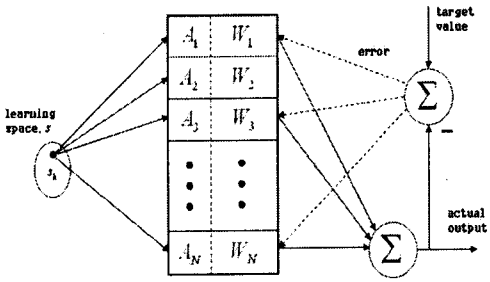

그림 2.1 CMAC의 동작

Fig. 2.1 Operation of CMAC

$$y_{s_k} = \sum_{j=1}^{N} c_{s_k,j} w_j$$

$$w_j^{(i)} = w_j^{(i-1)} + \frac{\alpha}{m} c_{s_k,j} \left(\overline{y}_{s_k} - \sum_{j=1}^{N} c_{s_k,j} w_j^{(i-1)} \right) \quad (2.1)$$

CMAC의 동작과 갱신수식은 그림 2.1과 수식 (2.1)과 같다. CA-CMAC의 타일 내의 저장된 데이터의 갱신 수식은 다음과 같다.

$$w_j^{(i)} = w_j^{(i-1)} + \alpha c_{s_k,j} \left\{ \frac{(f(j)+1)^{-1}}{\sum_{l}^{m} (f(l)+1)^{-1}} \right\} \left(\overline{y}_{s_k} - \sum_{j=1}^{N} c_{s_k,j} w_j^{(i-1)} \right) \quad (2.2)$$

수식 (2.2)에서 $f(j)$는 j-번째 타일의 학습된 횟수이다. CMAC의 갱신수식(2.1)과 CA-CMAC 의 갱신수식(2.2)이 다른 점은 1/m을 $(f(j)+1)^{-1} / \sum_{l}^{m} (f(l)+1)^{-1}$ 으로 대치한 것이다. 임의의 상태와 연관된 m개 타일의 데이터들의 학습된 횟수가 모두 같거나 학습이 한번도 이루어지지 않았다면 CA-CMAC의 갱신값은 CMAC의 갱신값과 같을 것이다. 반면, 학습된 횟수가 다를 경우

에는 학습된 횟수가 많은 타일의 갱신 비율은 1/m 보다 작게 반대로 학습된 횟수가 적은 타일은 갱신 비율이 1/m보다 크게 된다.

2. CACRL(CA-CMAC-based Reinforcement Learning)

CRL은 연속 상태 공간에서의 Acrobot swing-up 제어를 위해서 CMAC 구조를 적용한 강화학습방법이다. Sutton은 CRL의 학습속도 개선을 위하여 Temporal Difference 학습과 Monte Carlo 방법의 장점을 취한 Sarsa(λ)와 Watkins의 Q(λ) 방법을 이용하였고, 이 방법들은 eligibility trace라는 변수를 도입하여 현재 상태로부터 앞으로 몇 step 후의 상태들의 기대치를 현재의 행위 함수의 갱신에 이용한다.

Sutton은 미래의 기대치(Return)를 얻어서 함수를 갱신하는 것과 eligibility trace를 이용하여 과거의 값으로부터 함수를 갱신하는 것이 off-line backup일 경우 등가임을 증명하였다.

그러나, eligibility trace 방법이 미래의 상태들에 대한 누적된 가중치를 이용하여 credit을 assign하기는 하지만, 연속 상태에서 CMAC을 이용할 경우, 행위함수를 계산하는데 참조되는 tile들의 갱신에 cedit을 assign하는 개념과는 구별된다. 기존의 CRL 방법의 tile의 갱신수식은 다음과 같다.

$$\vec{\theta} \leftarrow \vec{\theta} + \frac{\beta}{m}\delta \vec{e} \tag{2.3}$$

$\vec{\theta}$는 CMAC 내의 모든 tile의 벡터이고, β는 step 크기 비율, m은 tile의 개수, δ는 현재 행위함수의 예측되는 오차, 그리고, $\vec{e}$는 모든 eligibility trace를 나타내는 벡터이다. 수식 (2.3)에서 $\vec{e}$는 행위함수간의 credit assignment개념을 갖는 백터로서 현재의 상태와 연관된 tile들의 합으로 행위함수를 계산하고, 이 행위함수로 인하여 발생하는 예측오차를 다시 연관된 tile들의 가중치 학습에 분배하는 비율과는 구분된다. CRL방법에서 이용하는 tile들의 갱신 가중치는 1/m로 일정하게 분포된다.

본 논문에서는 행위함수의 예측오차를 다시 연관된 tile들에 학습시킬 때 eligibility trace 뿐만 아니라 CA-CMAC방법을 이용하여 갱신 가중치를 두어 학습하는 방법으로 다음과 같은 갱신 수식을 제안한다.

$$\theta_i(s,a) \leftarrow \theta_i(s,a) + \beta \left\{ \frac{\left(L_i(s,a)+1\right)^{-1}}{\sum_{k=0}^{m}\left(L_k(s,a)+1\right)^{-1}} \right\} \delta e_i(s,a) \tag{2.4}$$

s는 연속상태공간 내의 현재의 상태이고, $\theta_i(s,a)$는 s상태에서의 행위 a에 대한 행위함수를 계산하는데 참조되는 i번째 tile의 값이며, $e_i(s,a)$와 $L_i(s,a)$는 각각 그 tile에 대한 eibilility trace와 갱신횟수이다. CMAC 내의 모든 tile의 갱신을 위하여 s는 현재의 실제 상태에 대하여 tile의 크기 간격으로 확장한 가상의 불연속 상태 공간의 집합에 속하는 상태로 정의 하였다.

III. 모의 실험

Acrobot의 swing up 제어를 위한 학습 파라미터로써 eligibility trace 감쇠율(λ)은 0.9, step 크기(β)는 0.05, 랜덤비율(ε)은 0.0, 감쇠율(Γ)은 1.0으로 각각 설정하였다. Acrobot의 상태공간은 조인트 1의 각도와 각속도, 조인트 2의 각도와 각속도로 표현되는 4차원의 연속상태공간이며 이를 함수 근사화하기 위하여 10 층의 9×9×9×9 tile을 갖는 구조의 CMAC을 적용하였다.

$$\begin{aligned}\begin{bmatrix}0\\ \tau_2\end{bmatrix} &= \begin{bmatrix} m_2L_2^2 + 2m_2L_1L_2c_2 + (m_1+m_2)L_1^2 & m_2L_2^2 + m_2L_1L_2c_2 \\ m_2L_1L_2c_2 + m_2L_2^2 & m_2L_2^2 \end{bmatrix}\begin{bmatrix}\ddot{\theta}_1\\ \ddot{\theta}_2\end{bmatrix} \\ &+ \begin{bmatrix} -m_2L_1L_2s_2\dot{\theta}_2^2 - 2m_2L_1L_2s_2\dot{\theta}_1\dot{\theta}_2 + m_2L_2gc_{12} + (m_1+m_2)gL_1c_1 \\ m_2L_1L_2s_2\dot{\theta}_1^2 + m_2L_2gc_{12} \end{bmatrix} \\ &= M(\theta)\ddot{\theta} + V(\theta,\dot{\theta}) + G(\theta)\end{aligned} \tag{3.1}$$

모의 실험에서 사용한 acrobot의 운동방정식은 수식 (3.1)과 같으며 조인트 2에 가해지는 토크를 강화학습의 행위로 설정하였다. 현재의 행위에 대하여 운동방정식(3.1)을 풀어서 조인트1, 조인트2의 각가속도에 해당하는 해를 구한 후 이를 Runge Kutta방법으로 적분하여 각각의 조인트의 각속도와 각도를 구하였고, 이렇게 구한 값들을 강화학습의 4차원 상태로 매핑시켰다.

acrobot의 swing-up 제어를 위한 목표 위치의 범위는 다음 그림과 같이 조인트 1의 각도는 5도 이상, 조인트의 각도는 30도 이상의 범위로 설정하였다.

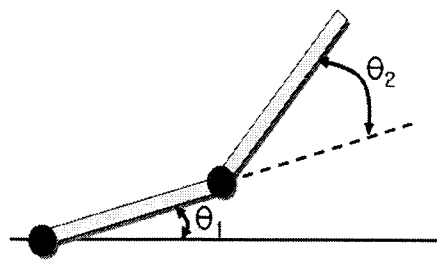

그림 3.1 swing up 제어를 위한 목표 범위
Fig. 3.1 Target boundary for swing up control

CRL과 CACRL의 비교를 위한 모의 실험을 수행하였으며 그림 3.1의 그래프는 episode에 따른 step 수를 나타낸다.

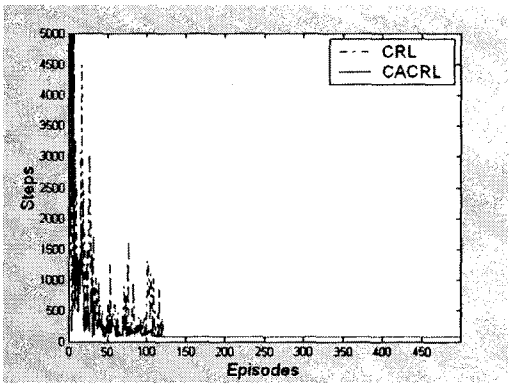

그림 3.2 CRL과 CACRL의 비교
Fig. 3.2 Target boundary for swing up control

Ⅳ. 결론

강화학습이 실제 환경에 적용되기 위해서 해결해야 할 중요한 문제 중의 하나는 연속 상태공간을 불연속 상태공간으로 표현함으로써 생기는 많은 용량의 메모리 문제(curse of dimensionality)이다. 이를 해결하기 위한 방법으로 강화학습 연구 분야에서는 함수 근사화 방법이 연구되어왔다. 본 논문에서는 함수 근사화를 적용한 Sutton의 CRL(CMAC- based 강화학습)방법의 수렴속도를 개선하기 위한 방법으로 S. Su 등이 제안한 CA-CMAC 기법을 적용한 CACRL (Credit-Assigned-CMAC-based 강화학습)방법을 제안하였고, CRL과 제안한 CACRL 방법을 Acrobot의 swing-up 제어에 적용한 모의실험을 통하여 비교 하여 수렴속도의 우수성을 보였다.

추후 연구로는 실제 Acrobot에 적용한 실험을 통하여 제안한 방법의 유용성을 검증하는 것과 Acrobot의 balancing제어를 위한 불연속 상태공간 및 불연속 행위공간에서 연속 행위의 생성 방법에 대한 연구가 필요하다.

참고문헌

[1] R. M. Murray and J. Hauser, "A Case Study in Approximate Linerization: The Acrobot Example," Electronics Research Lab. College of Engineering, University of California, Berkely, Apirl 1991.

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

[3] J. A. Boyan and A. W. Moore, "Generalization in Reinforcement Learning: Safely Approximation the Value Function," NIPS-7. San Mateo, CA: Morgan Kaufmann, 1995.

[4] M. W. Spong, "The swing up control problem for the acrobot," IEEE Control Systems Magazine, vol. 15, pp. 49-55, feb. 1995.

[5] G. Boone, "Minimum-time control of the acrobot," International Conference on Robotics and Automation, pp. 3281-3287, 1997.

[6] R. S. Sutton, "Generalization in reinforcement learning: successful examples using sparse coarse coding," in Neural information Proceeding Systems 8, pp. 1038-1044, MIT Press, 1996.

[7] S. Su, T. Tao, and T. Hung, "Credit Assigned CMAC and Its Application to Online Learning Robust Controllers," IEEE Transaction on Systems, Man, and Cybernetics-Part B:Cybernetics, vol. 33, no. 2, April 2003.

[8] A. D. Luca and G. Oriolo, "Stabilization of the Acrobot via Iterative State Steering," Proceeding of the 1998 IEEE International Conference on Robotics & Automation, Leuven, Belgium, May 1998.

[9] S. C. Brown and K. M. Passino, "Intelligent Control of the Acrobot," Journal of Intelligent and Robotics Systems 18: 209-248, 1997.

[10] J. A. Boyan and A. W. Moore, "Generalization in Reinforcement Learning: Safely Approximating the Value function," NIPS-7 San Mateo, 1995.

[11] 김재현, 서일홍, "지능형 로봇 시스템을 위한 영역기반 Q-Learning", 제어자동화시스템공학회 논문지, 제3권 ,제4호, 8월, 1997.

[12] D. Ormoneit and P. Glynn, "Kernel-Based Reinforcement Learning in Average-Cost Problems," IEEE Transaction on Automatic Control, Vol. 47. No. 10, October, 2002.