한국지능시스템학회 논문지 2009, Vol. 19, No. 4, pp. 519-524

액터-크리틱 퍼지 강화학습을 이용한 기는 로봇의 제어

Control of Crawling Robot using Actor-Critic Fuzzy Reinforcement Learning

문영준[1] · 이재훈[2] · 박주영[2]
Youngjoon Moon[1], Jaehoon Lee[2] and Jooyoung Park[2]

[1]대우조선해양 미래연구소
[2]고려대학교 제어계측공학과

요 약

최근에 강화학습 기법은 기계학습 분야에서 많은 관심을 끌어왔다. 강화학습 관련 연구에서 가장 유력하게 사용되어 온 방법들로는 가치함수를 활용하는 기법, 제어규칙(policy) 탐색 기법 및 액터-크리틱 기법 등이 있는데, 본 논문에서는 이들 중 연속 상태 및 연속 입력을 갖는 문제를 위하여 액터-크리틱 기법의 틀에서 제안된 알고리즘들과 관련된 내용을 다룬다. 특히 본 논문은 퍼지 이론에 기반을 둔 액터-크리틱 계열 강화학습 기법인 ACFRL 알고리즘과, RLS 필터와 NAC(natural actor-critic) 기법에 기반을 둔 RLS-NAC 기법을 접목하는 방안을 집중적으로 고찰한다. 고찰된 방법론은 기는 로봇의 제어문제에 적용되고, 학습 성능의 비교로부터 얻어진 몇 가지 결과가 보고된다.

키워드 : 퍼지 모델, 액터-크리틱 방법, RLS-NAC, 기는 로봇

Abstract

Recently, reinforcement learning methods have drawn much interests in the area of machine learning. Dominant approaches in researches for the reinforcement learning include the value-function approach, the policy search approach, and the actor-critic approach, among which pertinent to this paper are algorithms studied for problems with continuous states and continuous actions along the line of the actor-critic strategy. In particular, this paper focuses on presenting a method combining the so-called ACFRL(actor-critic fuzzy reinforcement learning), which is an actor-critic type reinforcement learning based on fuzzy theory, together with the RLS-NAC which is based on the RLS filters and natural actor-critic methods. The presented method is applied to a control problem for crawling robots, and some results are reported from comparison of learning performance.

Key Words : Fuzzy Model, Actor-Critic Method, RLS-NAC, Crawling Robot

1. 서 론

최근에 이론적 가치 및 실용적 가능성 측면에서 크게 주목받고 있는 기계학습(machine learning)의 한 분야인 강화학습(reinforcement learning)은 인간 또는 동물이 경험을 통해 지식을 습득해 나아가는 과정을 모사하여 사용자가 원하는 목표를 달성하는 방법론이다.

강화학습은 시행착오를 통해 문제의 해결책을 찾아가기 때문에 수식적인 모델링으로 해(solution)를 얻기 힘든 복잡한 비선형 시스템(nonlinear system)을 제어 하는데 장점이 있다. 그리고 제어 분야[1-2] 뿐만 아니라 게임[3], 금융[4] 등 여러 분야에 성공적으로 적용되었으며, 점차적으로 범용성을 넓혀가고 있다.

강화학습은 이산(discrete)인 저차원 공간을 가지는 문제에서는 쉽게 적용되지만, 하드웨어가 결합된 연속적(continuous)인 고차원 시스템에서는 계산속도 문제와 시스템이 가지는 공간을 정확하게 표현할 수 없는 문제 등으로 인해 학습이 성공적으로 이루어지지 않는 경우가 종종 발생한다[5].

그래서 이와 같은 시스템이 가지는 공간을 표현하는 공간의 일반화(generalization)에 관한 문제에 대해서는 이전부터 광범위하게 연구가 진행되어 오고 있다. 이에 대한 예로 타일 코딩(tile-coding)이나 신경망(neural network) 등의 방법과 이를 서로 융합하는 다양한 방법 등을 들 수 있다[5].

본 논문에서는 연속적인 고차원 공간에 강화학습을 적용하는 것을 목표로 하여 액터-크리틱 방법(actor-critic method), 퍼지 이론, RLS(recursive least-squares) 필터 등을 종합적으로 사용하는 방법론을 고려한다. 강화학습 관련 연구에서 가장 유력하게 사용되어 온 방법들로는 가치함수(value function)를 활용하는 기법, 정책(제어규칙, policy) 탐색 기법 및 액터-크리틱 기법 등이 있는데, 본 논문

접수일자 : 2009년 4월 6일
완료일자 : 2009년 6월 26일
"본 논문은 본 학회 2009년도 춘계학술대회에서 선정된 우수논문입니다."
이 논문은 지식경제부의 융복합형로봇전문인력양성사업의 지원을 받아 수행되었습니다.

에서는 이들 중 연속 상태 및 연속 입력을 갖는 문제를 위하여 액터-크리틱 기법의 틀에서 제안된 알고리즘들과 관련된 내용을 다룬다. 특히 본 논문은 퍼지 이론에 기반을 둔 액터-크리틱 계열 강화학습 기법인 ACFRL(actor-critic fuzzy reinforcement learning) 알고리즘[6]과, RLS 필터[7]와 액터-크리틱 기법[8-10]에 기반을 둔 RLS-NAC(recursive least-squares natural actor-critic) 기법[11]을 접목하는 방안을 집중적으로 고찰한다. 고찰된 방법론은 기는 로봇의 제어문제에 적용되고, 학습 성능의 비교로부터 얻어진 몇 가지 결과가 보고된다.

본 논문의 2장에서는 RLS-NAC 알고리즘[11]을 상세히 설명하고, 3장에서는 ACFRL 알고리즘[6]에 대해 설명한 후 RLS-NAC와의 접목을 고려한다. 그리고 4장에서는 실험대상인 Kimura의 기는 로봇(crawling robot)[12]에 대해 설명하고, 시뮬레이션 결과를 고찰한다. 마지막으로 5장에서는 결론과 향후 연구방향을 제시한다.

2. 기초 이론

2.1 강화학습의 기초

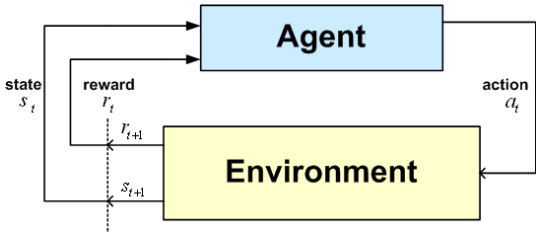


그림 1. 강화학습의 구조[5].
Fig. 1. Structure of reinforcement learning[5].

강화학습(reinforcement learning)은 시간이 진행됨에 따라 최적의 의사결정을 내려야 하는 순차적 의사결정 문제(sequential decision making problems)를 해결하는 방안의 하나로써, 그림 1에서와 같이 에이전트(제어기, agent)와 환경(environment)의 상호작용에 따라 관찰되는 상태(state), 입력(action) 및 보상 값(reward)을 효과적으로 활용하여 최적의 정책을 찾아가는 방향으로 학습을 진행하는 방법론이다[5]. 다음 절에서는 본 논문의 주요 소재가 되는 RLS-NAC 알고리즘[11]에 관한 주요 사항을, 액터-크리틱 관련 기초 이론[5,8]을 바탕으로 하여 기술한다.

2.2 RLS-NAC 알고리즘

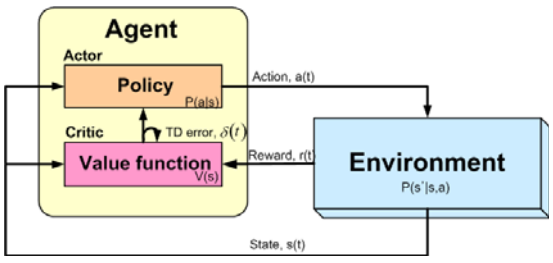


그림 2. 액터-크리틱 구조[5].
Fig. 2. The actor-critic architecture[5].

강화학습에서 널리 사용되는 액터-크리틱 전략은 그림 2과 같이 가치함수와 정책을 분리하여 사용하는 구조를 가지고 있다. 액터 부분은 현재 시점의 상태와 TD 에러를 이용해 정책에 따라 액션을 선택하는 부분이고, 크리틱 부분은 현재의 상태와 보상 값을 이용해 정책에 의해 선택된 액션을 평가하는 부분이다[5].

그리고 강화학습의 목표에 따라서 식 (1)로 정의된 목적함수(objective function)를 최대화한다.

$$J(\pi)=J(\theta)=V^{\pi_\theta}(s_0) = \sum_s d^{\pi_\theta}(s)\sum_a \pi_\theta(a|s)r(s,a) \tag{1}$$

여기에서 θ는 정책의 확률 분포를 나타내는 파라미터이다.

액터(actor) 부분에서는 최적의 정책을 찾기 위해 식 (1)을 θ에 대하여 미분을 취하는 정책 기울기 방법(policy gradient method)[8-9]을 이용하여 식 (2)을 얻을 수 있다.

$$\begin{aligned}\nabla_\theta J(\theta) &= \sum_s d^{\pi_\theta}\sum_a \nabla_\theta \pi_\theta(a|s)\, Q^{\pi_\theta}(s,a) \\ &= \sum_s d^{\pi_\theta}(s)\sum_a \pi_\theta(a|s)\left(Q^{\pi_\theta}(s,a)-V^{\pi_\theta}(s)\right) \\ &= \sum_s d^{\pi_\theta}(s)\sum_a \pi_\theta(a|s)\log\pi_\theta(a|s)A^{\pi_\theta}(s,a)\end{aligned} \tag{2}$$

여기에서 이득 가치 함수(advantage value function)는 $A^{\pi_\theta}(s,a)=Q^{\pi_\theta}(s,a)-V^{\pi_\theta}(s)$로 정의되는 함수로써, 상태 s에서 입력 a를 취했을 때 평균적인 경우보다 얼마나 더 많은 보상합(rewards sum)을 얻을 수 있는지를 나타내며, 다음의 식 (3)으로 근사하는 방안이 널리 사용된다[8,10].

$$A^{\pi_\theta}(s,a)\approx \widetilde{A}_w(s,a)=\nabla_\theta \log\pi_\theta(a|s)^T w \tag{3}$$

액터에서 사용하는 정책의 분포를 표현하는 파라미터θ의 업데이트 규칙을 얻기 위해 자연 기울기 방법(natural gradient method)[10]를 사용할 수 있으며, 참고문헌 [10]의 관찰에 따라 이 업데이트 규칙은 다음의 식 (4)와 같이 단순화될 수 있다.

$$\theta_{t+1}=\theta_t+\alpha\widetilde{\nabla}_\theta J(\theta)\approx\theta+\alpha w \tag{4}$$

크리틱(critic) 부분에서는 식 (3)과 가치 함수를 선형 근사화한 식 (5)을 이용하여 벨만 방정식(Bellman equation)으로 부터 식 (6)과 같은 근사식을 구해낼 수 있다.

$$V^{\pi_\theta}(s)\approx \widetilde{V}_v(s)=\phi^T(s)v \tag{5}$$

$$\begin{aligned}Q^{\pi_\theta}(s_t,a_t) &= A^{\pi_\theta}(s_t,a_t)+V^{\pi_\theta}(s_t) \\ &\approx r_t+\gamma V^{\pi_\theta}(s_{t+1})\end{aligned} \tag{6}$$

식 (6)에서 각 항에 해당하는 근사화된 식을 대입하여 최소 자승법(least-squares method)을 적용시키면 식 (7)과 같이 표현될 수 있다. 이를 통해 근사화된 가치함수의 에러를 최소화하는 파라미터를 찾게 된다.

$$\begin{aligned}&\Psi_t(v,w)\\&=\left\|\sum_{k=0}^{t}z_k\left[\left(\widetilde{V}_v(s_k)+\widetilde{A}_w(s_k,a_k)\right)-\left(r_k+\gamma\widetilde{V}_v(s_{k+1})\right)\right]\right\|^2\\&=\left\|\begin{array}{l}\sum_{k=0}^{t}z_k\left[\phi^T(s_k)-\gamma\phi^T(s_{k+1}),\nabla_\theta\log\pi(a_k|s_k)^T\right]\begin{bmatrix}v\\w\end{bmatrix}\\-\sum_{k=0}^{t}z_k r_k\end{array}\right\|^2\end{aligned}\tag{7}$$

여기에서 z_k는 학습의 속도를 향상시키는 적격성 벡터(eligibility trace vector)로써, 식 (8)과 같이 정의한다[11].

$$\begin{aligned}z_k&=\gamma\lambda z_{k-1}+\left[\phi^T(s_k),\ \nabla_\theta\log\pi(a_k|s_k)^T\right]^T\\z_0&=\left[\phi^T(s_0),\ \nabla_\theta\log\pi(a_0|s_0)^T\right]^T\end{aligned}\tag{8}$$

그리고, v와 w를 알기 위해 식 (9)와 같이 RLS 기법[7]을 적용시키면 [11-12]에서 볼 수 있듯이, 식 (10)에 보여진 파라미터 업데이트 식을 얻을 수 있다.

$$\begin{bmatrix}v_t\\w_t\end{bmatrix}=A_t^{-1}b_t\tag{9}$$

$$\begin{aligned}A_0&=\delta I+\left[\phi^T(s_0),\ \nabla_\theta\log\pi(a_0|s_0)^T\right]^T\\&\quad\left[\phi^T(s_0)-\gamma\phi^T(s_1),\ \nabla_\theta\log\pi(a_0|s_0)^T\right],\\A_t&=\beta A_{t-1}\\&\quad+z_t\left[\phi^T(s_t)-\gamma\phi^T(s_{t+1}),\ \nabla_\theta\log\pi(a_t|s_t)^T\right],\\P_t&=A_t^{-1},\\K_t&=P_t z_t,\\z_t&=\gamma\lambda z_{t-1}+\left[\phi^T(s_t),\ \nabla_\theta\log\pi(a_t|s_t)^T\right]^T,\end{aligned}\tag{10}$$

여기에서 망각상수(forgetting factor) β를 위해서는 1 미만의 양수가 사용되는데, 최근에 얻어지는 데이터를 더 중시하는 역할을 수행한다.

식 (4)과 식 (10)의 업데이트 규칙을 이용하여 RLS-NAC 알고리즘은 보상 값의 합을 최대로 하는 정책 파라미터와 근사화된 가치함수 파라미터를 찾게 된다 [11,13].

3. ACFRL 알고리즘과의 접목

앞에서 설명하였듯이, 강화학습은 상태나 액션 공간이 크거나 연속적인 시스템을 다룰 때 그 공간을 적절하게 표현할 수 있는 방법을 찾아야 한다. 이에 따라 본 논문에서는 강화학습과 퍼지 이론을 혼합한 방법에 대해 고려해 본다.

퍼지이론은 인간의 언어나 사고 등과 관련된 애매모호함을 수리적으로 다루는 분야이며, 근사적이고 정확하게 표현할 수 없는 것을 표현하는데 효과적이다[14-15]

기본적인 퍼지 논리 시스템(fuzzy logic system)은 그림 3과 같이 퍼지화(fuzzification), 퍼지 규칙(fuzzy rule) 선정, 퍼지 추론(fuzzy inference), 비퍼지화(defuzzification) 과정으로 구성되어 있다[14-15].

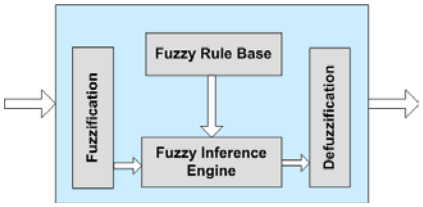


그림 3. 퍼지 논리 구조[14].
Fig. 3. Fuzzy logic system[14].

이와 같은 퍼지 이론에 기초를 두어 본 절에서는 ACFRL[6]과 RLS-NAC[11]의 결합에 대해 고려한다.

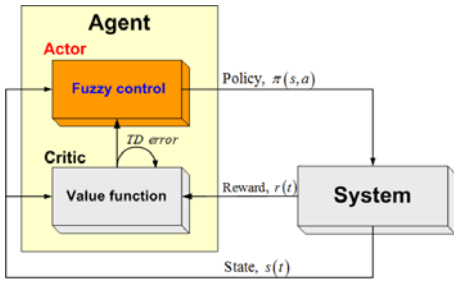


그림 4. ACFRL의 다이어그램[6].
Fig. 4. Diagram of ACFRL[6].

ACFRL은 고차원 공간을 가지는 시스템에 대한 문제를 이에 관련된 파라미터들을 조절하고, 불확실성에 관한 문제에 부딪치는 경우엔 함수 근사화를 사용하는 알고리즘이다.

액터-크리틱 퍼지 강화학습 알고리즘의 주요 특징은 그림 2에서 살펴보았던 액터-크리틱 구조의 액터 부분이 그림 4와 같은 퍼지 룰-베이스 액터(fuzzy rule-base actor)로 이루어지는 것이다.

액터 부분의 퍼지 룰-베이스(fuzzy rule-base)는 여러 개의 퍼지 규칙(fuzzy rule)을 모아 놓은 것으로써, 입력 벡터 $s\in R^K$를 출력 벡터 $a\in R^m$로 맵핑(mapping)하는 함수이다. 그러나 본 논문에서는 입력 벡터 $s\in R^K$를 출력 $a\in R$로 맵핑하는 MISO(multiple input single output) 퍼지 시스템을 이용한다.

그리고 룰-베이스는 Takagi-Sugeno-Kang(TSK) 형태[16]를 고려하게 되는데, 이 TSK 모델은 입력 공간을 퍼지 부분 공간으로 나누고, 각 규칙의 입력과 출력의 관계를 선형적인 모델로 표현한다. 이에 대해 가중치를 고려하며, 이들의 합을 통해 비선형 시스템을 선형화하는 과정이 이루어진다. 그리고 IF-THEN 규칙에선 IF 부분은 퍼지형태가 되고, THEN 부분은 $\overline{a^i}=f_i$와 같이 국소적 선형 함수(local linear function)의 형태가 된다. 그래서 i번째 MISO의 룰(rule)을 식 (11)과 같이 수식적으로 표현할 수 있다.

$$\begin{aligned}&Rule\ i:\ IF\ s_1\ is\ S_1^i\ \text{and}\ s_2\ is\ S_2^i\ \text{and}\\&\qquad\cdots\ \text{and}\ s_K\ is\ S_K^i\\&THEN\ a\ is\ \overline{a^i}=a_0^i+\sum_{j=1}^{K}a_j^i s_j\end{aligned}\tag{11}$$

여기에서 s_k는 입력 공간의 변수, K는 입력 공간 차원, S_k^i는 i번째 룰의 퍼지집합 소속 함수, 그리고 a_j^i는 조절 가

능한 계수이다.

소속 함수(membership function)는 입력이 퍼지 집합에 속하는 정도를 맵핑하는 함수로 식 (12)와 같이 표현되며 가우시안(gaussian) 형태의 소속 함수를 이용하여 전개된다.

$$\mu_{S_j^i}(s_j)=b_j^i\exp\left(-\frac{\left(s_j-\overline{s_j^i}\right)^2}{2\sigma_j^{i^2}}\right) \tag{12}$$

만약 M개의 규칙을 가지고 있을 때, 식 (11)과 식 (12)는 각각의 규칙을 통해 나오는 값들을 가중치 w^i를 고려하여 TSK의 최종 출력을 식 (13)을 이용해 얻는다[14].

$$a=f(s)=\frac{\sum_{i=1}^{M}\overline{a^i}w^i(s)}{\sum_{i=1}^{M}w^i(s)} \tag{13}$$

여기에서 $\overline{a^i}$는 i번째 룰의 출력이고, $w^i(s)$는 i번째 규칙의 가중치이다.

그리고 각 규칙의 적합도를 의미하는 가중치를 계산하기 위해 곱의 추론(product inference)을 사용하며, i번째 룰에 대한 곱의 추론은 식 (14)를 통해서 얻을 수 있다[14].

$$w^i(s)=\prod_{j=1}^{K}\mu_{S_j^i}(s_j) \tag{14}$$

식 (14)을 식 (13)에 대입하여, 식 (15)를 얻는다.

$$\begin{aligned}a=f(s)&=\frac{\sum_{i=1}^{M}\exp\left(-\frac{\left(\overline{a^i}-a\right)^2}{2\sigma^{i^2}}\right)w^i}{\sum_{i=1}^{M}w^i}\\&=\frac{\sum_{i=1}^{M}\overline{a^i}\left(\prod_{j=1}^{K}b_j^i\right)\exp\left(-\sum_{j=1}^{K}\frac{\left(s_j-\overline{s_j^i}\right)^2}{2\sigma_j^{i^2}}\right)}{\sum_{i=1}^{M}\left(\prod_{j=1}^{K}b_j^i\right)\exp\left(-\sum_{j=1}^{K}\frac{\left(s_j-\overline{s_j^i}\right)^2}{2\sigma_j^{i^2}}\right)}\end{aligned} \tag{15}$$

강화학습 적용에선 식 (13)과 같은 결정적인(deterministic) 액션을 사용하는 대신에 액션 공간에 대한 확률 분포(probability distribution)를 고려한다. 따라서 ACFRL 알고리즘에선 액션에 대한 확률 분포를 고려하여 식 (13) 대신 평균이 $\overline{a^i}$이고 분산이 σ^i인 가우시안 분포를 따르는 정책(policy)을 사용하게 된다. 즉, 입력 변수에 대해 RBF(Radial Basis Function)를 이용한 가중치의 평균의 형태가 된다. 만약 현재 시스템의 상태가 s일 때, 액션 a를 취하는 정책은 식 (16)과 같다.

$$\pi_\theta(s,a)=\frac{\sum_{i=1}^{M}\exp\left(-\frac{\left(\overline{a^i}-a\right)^2}{2\sigma^{i^2}}\right)\left(\prod_{j=1}^{K}b_j^i\right)\exp\left(-\sum_{j=1}^{K}\frac{\left(s_j-\overline{s_j^i}\right)^2}{2\sigma_j^{i^2}}\right)}{\sum_{i=1}^{M}\left(\prod_{j=1}^{K}b_j^i\right)\exp\left(-\sum_{j=1}^{K}\frac{\left(s_j-\overline{s_j^i}\right)^2}{2\sigma_j^{i^2}}\right)} \tag{16}$$

여기서 θ는 퍼지 룰-베이스의 액터 부분을 조절하는 파라미터이다.

지난 절에서 상세히 설명된 RLS-NAC의 기본 틀을 사용하되 액터의 구조를 위하여 ACFRL의 퍼지 이론에 기반을 둔 확률분포를 사용하면 두 방법의 접목이 가능하며, 이 접목 과정의 핵심적인 부분인 파라미터 θ를 업데이트를 위하여 식 (16)에 로그를 취하여 θ에 대한 미분을 적용하면 다음의 식 (17)을 얻을 수 있다.

$$\begin{aligned}\nabla_\theta\ln\pi(s,a)&=\frac{\nabla_\theta\pi(s,a)}{\pi(s,a)}\\&=\frac{\frac{\sum_{i=1}^{M}\exp\left(-\frac{\left(\overline{a^i}-a\right)^2}{2\sigma^{i^2}}\right)w^i}{\sum_{i=1}^{M}w^i}\frac{\left(a-\overline{a^i}\right)}{\sigma^{i^2}}s}{\frac{\sum_{i=1}^{M}\exp\left(-\frac{\left(\overline{a^i}-a\right)^2}{2\sigma^{i^2}}\right)w^i}{\sum_{i=1}^{M}w^i}}\end{aligned} \tag{17}$$

여기에서 θ는 각 퍼지 룰의 국소적 선형 모델(local linear model)을 조절할 수 있는 계수이고, s는 입력 벡터이다.

4. 기는 로봇에의 적용

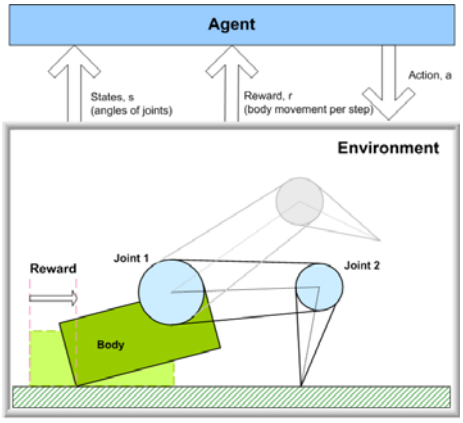


그림 5. 기는 로봇[12].
Fig. 5. Crawling robot[12].

본 논문에서 고려된 알고리즘을 참고 문헌 [12]에서 소개된 Kimura의 기는 로봇에 적용한다. 기는 로봇은 그림 5과 같이 중력의 영향 아래에서 움직이는 두 개의 링크를 가진 평면형 매니퓰레이터(two-linked planar manipulator)이며, 이에 대한 구체적인 설명은 다음과 같다[12,17]. 에이전트는 로봇 및 환경에 대한 구체적인 정보 없이 로봇의 직접적인 경험을 통해(즉, 확률적인 정책에 따라 관찰된 보상 값의 좋고 나쁨을 이용해) 동작한다. 기는 로봇이 움직일 때, 링크 끝부분과 지면 사이의 미끄러짐은 없고, 몸체와 지면 사이의 마찰력은 없다고 가정한다.

보상 값은 각 시간 스텝 당 이동한 거리로 정의 된다. 만약 로봇이 앞이나 뒤로 이동하면, 이에 해당되는 양수 또는 음수의 보상 값을 가지게 되며 이를 통해서 상점과 벌점이 주어진다.

기는 로봇의 상태 벡터는 $s(t)=[x_1(t),\ x_2(t),\ 1]^T$이며, 각 조인트의 범위는 $[-1,\ 1]$ 이다. 여기서 정책은 식 (18)과 같다[17].

$$\pi(a,\mu,\sigma)=\frac{1}{\sqrt{2\pi}\,\sigma}exp\left(-\frac{(a-\mu)^2}{2\sigma^2}\right), \quad (18)$$

where $\mu=\theta^T s$: average

σ^2: variance (18)

로봇의 상세 사양은 표 1과 같다[12].

표 1. 고려된 기는 로봇의 사양[12].
Table 1. Specifications of the considered crawling robot[12].

링크의 길이	link 1: 34㎝, link 2: 20㎝
조인트의 움직일 수 있는 범위	$-4^\circ \le joint1 \le 35^\circ$ $-120^\circ \le joint2 \le 10^\circ$
조인트1의 위치	수평방향으로 32㎝ 수직방향으로 18㎝

강화학습과 퍼지 제어 방법을 이용한 본 논문의 알고리즘은 고차원의 상태를 가지거나 공간의 근사화가 어려운 시스템에 적용할 수 있다. 이 알고리즘을 기는 로봇에 적용하는 문제에서, 액터 부분의 소속 함수는 식 (19)의 삼각형 모양의 함수를 사용하였다. 그림 6은 식 (19)를 이용하여 각 조인트의 소속 함수를 나타낸 것이다.

$$\mu_{S_j^i}=s\in[-1,\ 1]\rightarrow\begin{cases}\dfrac{x-a}{b-a} & when\ a\le s\le b\\ \dfrac{c-x}{c-b} & when\ b\le s\le c\\ 0 & otherwise\end{cases} \quad (19)$$

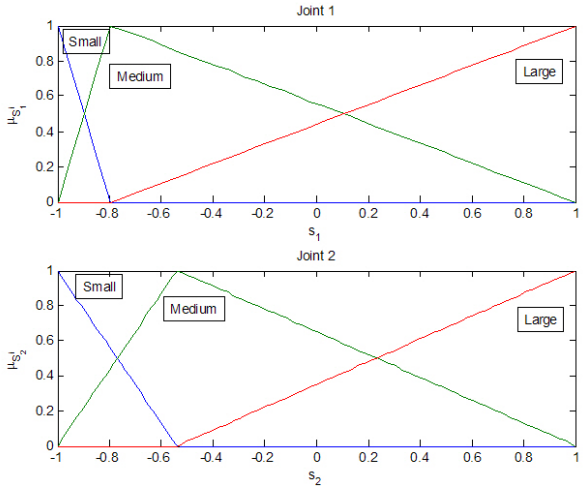


그림 6. 각 조인트의 소속 함수.
Fig. 6. Membership function of each joint

강화학습의 RLS-NAC 알고리즘에 퍼지 제어를 적용한 ACFRL 알고리즘은 총 20,000 시간 스텝동안 학습을 진행시키면서 정책 파라미터 $\theta=[a_j^i,\ b_j^i]^T\in R^{54}$를 찾는다.

ACFRL 알고리즘과의 성능 비교를 위해 그림 7과 같은 상태 공간 일반화(generalization) 방법을 고려하였다.

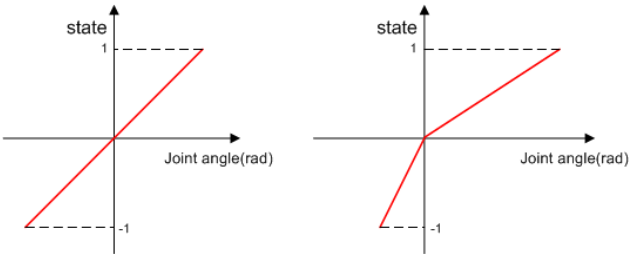


그림 7. 선형 근사화.
Fig. 7. Linear scaling and piecewise linear scaling.

그림 8은 3가지의 경우에 대해 총 10번의 시도 후 시간 스텝 당 평균 속도를 관찰하였다. 단지 $[-1,\ 1]$사이의 범위로 선형 근사화한 것과 $[-1,\ 0]$, $[0,\ 1]$로 구간을 나누어 구간 선형 근사화한 것과 비교할 때, 본 논문의 액터-크리틱 퍼지 강화학습을 적용한 것이 더 우수한 학습 성능을 보이는 것을 확인할 수 있었다.

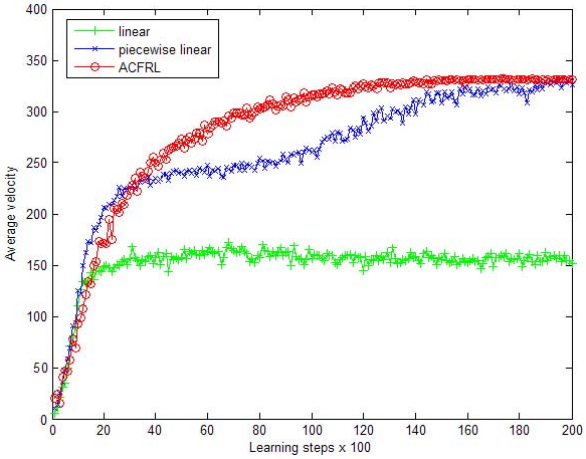


그림 8. 기는 로봇의 성능 비교.
Fig. 8. Velocity of the crawling robot

5. 결론 및 향후 연구방향

본 논문에서는 연속적이며 고차원의 상태 공간을 효과적으로 다룰 수 있는 액터-크리틱 기반 퍼지 강화학습 알고리즘을 구하기 위하여, RLS-NAC 알고리즘[11]을 ACFRL 기법[6]에 접목하는 문제를 고려한 후, 기는 로봇 문제[12]에 적용해보았다.

향후 연구 과제로는, 본 논문에서 고려한 알고리즘을 다양한 예에 적용해보는 문제와 보다 효과적인 전처리 과정의 도입 등을 들 수 있다.

참 고 문 헌

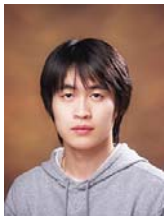
[1] Q. Yang, J. B. Vance, and S. Jagannathan, "Control of nonaffine nonlinear discrete-time systems using reinforcement-learning-based linearly parameterized neural networks," *IEEE Transactions on Systems, Man, and Cybernetics,*

Part B: Cybernetics, vol. 38, no. 4, pp. 994-1001, 2008.
[2] J. Valasek, J. Doebbler, M. D. Tandale, and A. J. Meade, "Improved adaptive-reinforcement learning control for morphing unmanned air vehicles," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 38, no. 4, pp. 1014-1020, 2008.
[3] K.-H. Park, Y.-J. Kim, and J.-H. Kim, "Modular Q-learning based multi-agent cooperation for robot soccer," *Robotics and Autonomous Systems*, vol. 35, no. 2, pp. 109-122, 2001.
[4] J. Moody and M. Saffell, "Learning to trade via direct reinforcement," *IEEE Transactions on Neural Networks*, vol. 12, no. 4, pp. 875-889, 2001.
[5] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, 1998.
[6] H. R. Berenji and D. Vengerov, "A convergent actor-critic-based RFL algorithm with application to power management of wireless transmitters", *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 4, August, 2003.
[7] X. Xu, H. He, and D. Hu, "Efficient reinforcement learning using recursive least-squares methods", *Journal of Artificial Intelligent Research*, vol. 16, pp. 259-292, 2002.
[8] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation", *Advances in Neural Information Processing Systems*, vol. 12, pp. 1057-1063, 2000.
[9] V. Konda and J. N. Tsitsiklis, "Actor-Critic Algorithms", *SIAM Journal on Control and Optimization*, vol. 42. no. 4, pp. 1143-1166, 2003.
[10] J. Peters, S. Vijayakumar, and S. Schaal, "Reinforcement learning for humanoid robotics", In *Proceedings of the Third IEEE-RAS International Conference on Humanoid Robots*, 2003.
[11] 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.
[12] H. Kimura, K. Mivazaki, and S. Kobayashi, "Reinforcement learning in POMDPs with function approximation", In *Proceedings of the 14th International Conference on Machine Learning(ICML 1997)*, pp. 152-160, 1997.
[13] 김종호, *강화학습 알고리즘을 이용한 시스템 제어에 대한 연구*, 고려대학교 제어계측공학과 석사학위논문, 2005.
[14] L. X. Wang, *Adaptive Fuzzy Systems and Control: Design and Stability Analysis*, Prentice-Hall, 1994.
[15] 박종진, 최규석, *퍼지 제어 시스템*, 교우사, 2001.
[16] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 15, pp. 116-132, 1985.
[17] 박주영, 정규백, 문영준, "강화학습에 의해 학습된 기는 로봇의 성능 비교", *한국 퍼지 및 지능시스템 학회 논문집*, 17권, 1호, pp. 33-36, 2007.

저 자 소 개

문영준(Youngjoon Moon)
2007년 : 고려대학교 제어계측공학과 졸업
2009년 : 동 대학원 제어계측공학과 석사
2009년～현재 : 대우조선해양 미래연구소

관심분야 : 지능시스템, 로봇제어.
Phone : 010-8573-8959
E-mail : yjmoon09@dsme.co.kr



이재훈(Jaehoon Lee)
2008년 : 고려대학교 제어계측공학과 졸업
2008년～현재 : 동 대학원 제어계측공학과 석사과정

관심분야 : 지능시스템, 관성항법시스템(INS), 신호융합.
Phone : 010-9141-9779
E-mail : white8704@korea.ac.kr

박주영(Jooyoung Park)
1983년 : 서울대 전기공학과 졸업(학사)
1985년 : KAIST 졸업(석사)
1985년～1988년 : 한국전력 월성원자력발전소 근무
1992년 : University of Texas at Austin 전기및컴퓨터공학과 졸업(박사)
1993년～현재 : 고려대학교 세종캠퍼스 제어계측공학과 교수

관심분야 : 계산지능, 서포트벡터학습, 강화학습
Phone : 041-860-1444
E-mail : parkj@korea.ac.kr