論文99-36S-2-4

多入力 시스템의 자율학습제어를 위한 차등책임 적응비평학습

(Differentially Responsible Adaptive Critic Learning(DRACL) for the Self-Learning Control of Multiple-Input System)

金 炯 奭 *

(Hyong-Suk Kim)

요 약

재 강화 학습 방법을 다수의 제어입력을 가진 시스템에 대한 자율적 제어 기법 습득에 활용하기 위해서 차등책임 적응비평 학습구조를 제안하였다. 재 강화 학습은 여러 단계의 제어동작 끝에 얻어지는 최종 비평값을 활용하여 그 전에 행해졌던 제어 동작을 강화 혹은 약화 학습하는 자율적 학습방법이다. 대표적인 재 강화학습 방법은 적응비평학습 구조를 이용하는 방법인데 비평모듈과 동작모듈을 이용하여 외부 비평 값을 최대로 활용함으로써 학습효과를 극대화시키는 방법이다. 이 학습방법에서는 단일한 제어입력을 갖는 시스템으로만 적용이 제한된다는 단점이 있다. 제안한 차등책임 적응비평 학습 구조에서는 비평함수를 제어 입력 인자의 함수로 구축한 다음 제어인자에 대한 차별 화된 비평 값을 부분미분을 통하여 산출함으로써 다수의 제어입력을 가진 시스템의 제어기술 학습이 가능하게 하였다. 제안한 학습제어 구조는 학습속도가 빠른 CMAC 신경회로망을 이용하여 구축하였으며 2 개의 제어입력을 갖는 2-D Cart-Pole 시스템과 3 개의 제어입력을 갖는 인간구조 로봇시스템의 앉는 동작의 학습제어 시뮬레이션을 통하여 효용성을 확인하였다.

Abstract

Differentially Responsible Adaptive Critic Learning technique is proposed for learning the control technique with multiple control inputs as in robot system using reinforcement learning. The reinforcement learning is a self-learning technique which learns the control skill based on the critic information obtained after a long series of control actions. The Adaptive Critic Learning (ACL) is the representative reinforcement learning structure. The ACL maximizes the learning performance using the two learning modules called the action and the critic modules which exploit the external critic value obtained seldomly. Drawback of the ACL is the fact that application of the ACL is limited to the single input system. In the proposed Differentially Responsible Action Dependant Adaptive Critic learning structure, the critic function is constructed as a function of control input elements. The responsibility of the individual control action element is computed based on the partial derivative of the critic function in terms of each control action element. The proposed learning structure has been constructed with the CMAC neural networks and some simulations have been done upon the two dimensional Cart-Pole system and robot squatting problem. The simulation results are included.

* 正會員, 全北大學校 電氣·電子·制御 工學部, 메카트로닉스 硏究所
(School. of Electrical Eng., Mechatronics Research Center, Chonbuk National Univ.)

接受日字:1998年3月9日, 수정완료일:1998年11月2日

Ⅰ. 서 론

제어기술의 학습시 각 상태에 대한 학습정보가 주어지지 않고 제어가 완료된 시점에서 주어지는 성공이나 실패 등으로 표현되는 경우에는 가용한 정보를 효과적으로 활용하여 각 상태에 대한 최적의 제어기술을 익혀야 한다. 이를 위해서는 제어의 시작에서부터 제어가 완결될 때까지 경과한 모든 상태들을 역 추적하여 각 상태에 대한 제어동작을 학습해야 하는데, 제어결과가 성공일 경우에는 기 행한 제어동작을 강화학습하며, 실패일 경우에는 비평 학습한다. 이 학습을 재강화 학습(Reinforcement Learning)이라고 한다.

재 강화 학습 방법은 Bush 와 Mosteller[1]에 의해 제안된 이래 Narendra와 Thathachar등에 의해 Learning Automata에 적용되어 발전되었다[2]. 그 후, 많은 연구가 되어졌는데[1,3,4], 그 핵심은 행한 동작결과가 만족스러울 경우, 차후에 동일 상황에서 동일한 동작을 할 확률을 강화하고 만족스럽지 않다면 그 확률을 줄임으로써 같은 동작을 반복하지 않게 하는 원리를 이용하였다. Fu등은 이 연구들을 재 강화 학습제어 기법으로 발전시켰으며[5], Nikolic등은 확률적 접근 방법(stochastic approach)을 사용하여 비평함수를 모델링하여 이용하였다[6]. 선험적 경험정보(priori information)를 이용하여 학습속도를 개선하는 연구도 있었다[7,8].

다른 그룹의 학자들은 재 강화 학습에 신경회로망을 이용함으로써 좋은 성과를 거두었다. Widrow등은 퍼셉트론 ADALINE에 블랙잭 게임을 재강화 학습방법을 이용하여 학습시켰으며[9], Klopf은 게임학습에 흥분신경과 억제신경 섬유로 구성된 신경회로망을 사용하였다[10]. 이 재 강화학습 방법은 Barto등의 적응비평학습(Adaptive Critic Learning, ACL)[11] 방법이 발표됨으로써 획기적인 발전의 전기가 마련되었다. 적응비평 학습방법은 비평함수의 학습을 위한 비평모듈(Critic Module)과 제어동작의 학습을 위한 제어동작 모듈(Action Module)을 이용하여 각 양자화된 상태에 할당된 신경 소자(Neuron-like element)에 의해 가용한 정보를 효과적으로 활용하여 학습하는 방법이다. 이 학습방법은 상태의 수가 많은 다 차원의 문제에 적용할 경우, 신경회로망의 크기가 커질 뿐아니라 학습시간이 과다하게 필요하다는 문제가 있었다. Anderson 등은 다층신경회로망을 이용하여 압축 학습함으로써 크기 증가문제는 완화시켰으나 장시간의 학습이 요구된다는 문제를 아직 갖고 있었다.[12] Lin 등은 CMAC 신경회로망[13]을 적응비평 학습구조에 이용함으로써 학습속도를 개선하였을 뿐아니라[14] 학습시 결정해야할 파라미터 선택방법도 제시하였다.[15]

이와 같은 많은 연구에도 불구하고 기존의 적응비평 학습구조들은 한 개의 입력을 가진 시스템에만 적용되는 단점을 가지고 있다. 본 연구에서는 여러 개의 제어입력을 갖는 시스템의 경우에도 효과적인 자율학습이 가능한 차등책임 적응비평 학습(Differentially Responsible Adpative Critic Learning,DRACL)구조를 제안하였다. 제안한 학습 구조에서는 비평함수를 제어 입력 인자의 함수로 구축한 다음 각 제어인자의 부분 미분 값에 비례하는 비평 값을 이용하여 인자별 제어기술을 학습할 수 있게 하였다. 제안한 학습구조가 다수의 제어입력을 가진 시스템에 적용할 수 있음을 보이기 위해 2 차원 cart-pole 시스템과 3 개의 제어입력이 필요한 인간구조 로봇시스템의 학습제어에 적용하였다.

본 논문의 제 II 절에서는 '적응비평학습의 원리에 대해 기술하였다. 적응비평학습구조에서는 학습을 용이하게 하기 위해 CMAC 신경회로망을 이용하는 것이 유리하므로 제 III 절에서는 이 CMAC을 이용한 적응비평 학습을 설명하였다. 제안한 차등책임 적응비평 학습 구조는 IV 절에서 소개하였으며 V 절에서는 제안한 학습구조의 학습제어 결과를 제시하였다. VI 절은 이에 대한 결론이다.

Ⅱ. 적응비평 학습

1. 적응비평학습 구조

적응비평 학습구조는 재강화 학습에 적합한 구조로서 ACE(Adaptive Critic Elements) 라는 비평 모듈과 ASE(Adaptive Search Elements)라는 제어동작 모듈로 구성된다. 비평모듈은 누적된 제어 동작결과로 얻어지는 비평 값의 활용을 극대화시키기 위해서 각 제어상태에 대한 비평함수 학습을 담당하며, 제어동작 모듈은 비평모듈로부터 얻은 상태 비평 값을 이용하여 최적의 제어동작 학습을 담당한다. 그림 1은 Barto의 적응비평 학습구조인데[11] 제어기의 상태변수는 양자 화되고 ASE와 ACE 내에는 각 양자 화된 상태를 위한 신경회로망 weight(w_{ai} 및 w_{ci})가 할당

되어 정보를 학습한다. ASE의 출력값은 *noise*가 첨가된 후 ±1로 양자화되어 플렌트의 제어신호로 인가되며 플렌트의 상태 S_K는 양자화되고 디코딩되어 ASE, ACE의 입력신호로 인가된다. 누적된 제어 동작의 결과로 플렌트가 제어에 실패하면 제어에 대한 비평값 R(f)= -1 값을 ACE에 인가하여 비평함수를 학습할 수 있게한다.

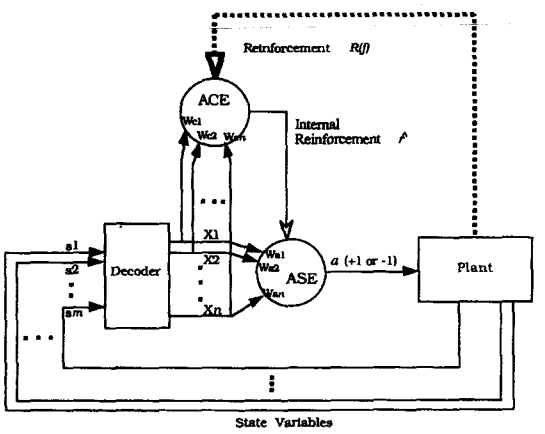


그림 1. Barto 등의 적응비평학습 구조
Fig. 1. ACL structure of Barto et al.

2. ACE 학습

ACE는 적응비평 학습구조에서 시스템의 비평함수를 학습하기 위한 신경회로망이다. 이 신경회로망에서 학습 목표 값과 실제 신경회로망 출력과의 차이를 error라고 하면 신경회로망의 weight는 Widrow-Hoff rule에 의해

$$\Delta \mathbf{w}_c = -\frac{1}{2}\beta \nabla_w error^2$$
$$= -\beta error \nabla_w error \quad (1)$$

로 학습될 수 있다. 여기서 β는 학습율이고 아래 첨자 c는 비평(critic)모듈임을 의미한다.

x를 신경회로망의 입력상태변수라 하고, r을 신경회로망 학습목표가 되는 비평값, k를 시간 변수라 하면 시간 k-1때의 입력 x(k-1)을 weight w(k-1)인 비평모듈에 인가함으로써 r(k-1)인 출력을 얻을 수 있다. 이 신경회로망을 이용한 재 강화 학습에 있어서, 상태 k-1 때 학습되어야 할 목표 값을 모르는 경우의 error는 r(k)와 r(k-1)의 차로부터 λ 만큼 감쇠된 값을 사용한다. 여기서 λ의 범위는 (0,1)이다.

$$error = \lambda (r(k) - r(k-1)) \quad (2)$$

식 (2)를 (1)에 대입하면 시간 k-1 에서의 weight 갱신량 $\Delta \mathbf{w}_c(\mathbf{k}-1)$은

$$\Delta \mathbf{w}_c(\mathbf{k}-1) = \beta (r(k) - r(k-1)) \lambda^2 \nabla_w r(k-1)$$
$$= \alpha (r(k) - r(k-1)) \nabla_w r(k-1) \quad (3)$$

이 된다. 여기서 $\alpha = \beta\lambda^2$이다. 그런데, k회의 누적된 제어동작 결과로서만 외부의 비평 값을 얻는 재강화 학습에 있어서는 r(k)를 얻게되는 책임이 상태 k-1 뿐아니라 그 이전에 선택된 모든 상태들에도 있을 것이므로 이들도 역시 갱신 학습되어야 할 것이다. 따라서 $\Delta \mathbf{w}_c(\mathbf{k}-1)$는

$$\Delta \mathbf{w}_c(\mathbf{k}-1) = \alpha (r(k) - r(k-1)) \sum_{j=0}^{k-1} \lambda^j$$
$$\nabla_w r(k-1-j) \quad (4)$$

로하여 weight를 갱신한다. 식 (4)에서 시간 k에서의 비평값 r(k)는 제어가 완결되어 외부로부터 비평값을 얻기 이전에는 얻을 수 없는 값이므로 제어가 완결되지 않은 시간 k에서는 신경회로망의 출력에 의해 예측된 값을 사용한다. 이 값은 어느 정도 불확실한 값이므로 학습에서는 r(k)에 discount 인자 γ 를 곱한 값을 목표 값으로 하여 사용한다[11]. 또한 제어가 완결되어 외부로부터 실제의 비평값 R을 얻었다면 r(k)는 R을 discount 없이 사용한다. 이 관계를 정리하면 $\Delta \mathbf{w}_c(\mathbf{k}-1)$는

$$\Delta \mathbf{w}_c(\mathbf{k}-1) = \alpha \hat{r}(k-1) \sum_{j=0}^{k-1} \lambda^j \nabla_w r(k-1-j) \quad (5)$$

이며, $\hat{r}(k-1)$는

$$\hat{r}(k-1) = \begin{cases} R-r(k-1); \text{외부로부터 비평값을 얻는 경우} \\ \gamma\, r(k)-r(k-1); \text{외부로부터 비평값을 얻지못하는 경우} \end{cases} \quad (6)$$

로 계산된다.

3. ASE 학습

ASE 모듈은 ACE로부터 제공되는 비평 정보를 활용하여 시스템의 최적 제어기술을 학습키 위한 모듈이다. 이 모듈의 학습 규칙을 유도하기 위해서 목적함수(objective function) J를 다음과 같이 정의한다.

$$J = \frac{1}{2} critic\, a^2 \quad (7)$$

여기서 a 는 제어 동작 값이며, critic은 제어동작 a 에 대한 비평 값 이다. 이 식에서 동작 a에 대한 비평이 긍정적이라면($critic>0$ 인 경우) a 의 절대 값이 클수록 좋은 효과(큰 J 값)를 얻으며, 부정적이라면 a 의 절대 값이 클수록 나쁜 효과(작은 J값)를 얻게됨을 나타내는 함수이다. 신경회로망은 이 비평함수에 의한 비평 값이 증가하는 방향으로 weight를 갱신해야만 최적의 제어 값을 얻을 수 있을 것이다. 따라서, Widrow-Hoff의 학습규칙에 의한 신경회로망의 학습은 다음과 같다.

$$\Delta \mathbf{w}_a = \beta \cdot \nabla_w J$$
$$= \beta \cdot critic \cdot a \cdot \nabla_w a \quad (8)$$

여기서 아래첨자 a는 제어동작모듈 임을 의미하며 β 는 학습률이다. 이 신경회로망에서 출력은 $a = \mathbf{w} \cdot \mathbf{x}$ 로 표현되므로 윗식은

$$\Delta \mathbf{w}_a = \beta \cdot critic \cdot a \cdot \nabla_w (\mathbf{w} \cdot \mathbf{x}) \quad (9)$$

가 된다. 여기서 $critic$은 상태 값에 대한 차이 값에 의해 표현할 수 있으므로 ACE에서의 $\hat{r}$를 사용할 수 있다. 윗 식에 시간 개념을 추가하여 정리하면, 시간 k-1에서의 weight 갱신 식은

$$\Delta \mathbf{w}_a(\mathbf{k-1}) = \beta \cdot critic(k-1) \cdot a(k-1) \cdot \nabla_w (\mathbf{w(k-1)} \cdot \mathbf{x(k-1)}) \quad (10)$$

이 된다. 그런데, ASE의 경우와 마찬가지로 시간 k-1에서 얻은 $critic$ 값은 이전에 거쳐왔던 상태들에도 책임이 있기 때문에 이전의 상태에 대한 weight도 갱신되어야 한다. 이 경우, $critic$(k-1)에 대한 책임은 시간 k로부터 멀어질수록 작을 것이므로 재 강화 학습 강도는 δ의 율로 감쇠된 학습을 사용한다. 따라서 동작모듈의 학습은

$$\Delta \mathbf{w}_a(\mathbf{k-1}) = \beta \cdot critic(k-1) \cdot \sum_{j=1}^{k} \delta^{j-1} \cdot a(k-j) \cdot \nabla_w (\mathbf{w(k-j)} \cdot \mathbf{x(k-j)}) \quad (11)$$

가 된다.

Ⅲ. CMAC 기반 적응 비평학습

제안한 적응비평학습 구조에서 사용한 신경회로망은 학습 속도가 매우 빠르다는 특징이 있는 CMAC 신경회로망을 사용하였다. 이 절에서는 CMAC 기반 적응비평제어의 원리를 설명한다.

1. CMAC 신경회로망

CMAC 신경회로망은 인간의 소뇌를 모델링한 신경회로망인데[13] 입력상태 공간을 양자 화하여 분할하며 각 양자화된 상태공간에는 다수개의 weight를 할당하여 학습을 담당하게 한다는 특징이 있다. 또한 이웃의 양자화된 상태공간과는 일부의 메모리가 공유되게 함으로써 일반화 기능을 갖게 하는 신경회로망이다. 그림 2 는 CMAC 신경회로망의 개념을 보여준다.

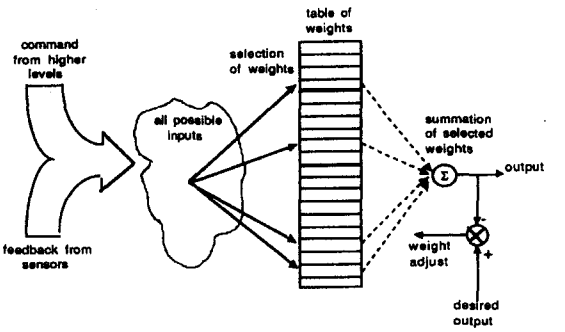


그림 2. CMAC 신경회로망 학습개념
Fig. 2. Learning concept of CMAC neural networks.

CMAC 신경회로망에서는 입력벡터를 양자화 하여 분할하고, 각 양자 화된 상태마다 고유한 숫자 이름을 할당하며 이 숫자를 seed로하여 랜덤하게 실제메모리를 할당한다. CMAC 신경망은 랜덤하게 메모리를 할당하는 과정 때문에 사용 메모리의 크기 면에 있어서 융통성이 있다.

CMAC 신경회로망에 크기가 P_m인 메모리가 주어지고 이 메모리의 값으로 구성된 벡터를 M이라고 하자. 또한 양자화된 상태 변수 s(k)를 디코딩하여 크기가 P_m이며 1의 개수가 N_e인 이진수로 표현된 벡터를 U(s(k))라 하자. 시간 k에서의 상태 x(k)에 대한 CMAC 신경회로망의 출력 I(x(k))는 다음과 같이 표현될 수 있다.

$$I(x(k)) = \frac{1}{N_e} \mathbf{M} \cdot \mathbf{U(x(k))} \quad (12)$$

2. CMAC 기반 적응비평 학습기법

그림 1의 적응비평 신경회로망 구조에서 ACE와 ASE를 CMAC 신경회로망으로 교체하여 구성하고

ACE 메모리를 $\mathbf{M}_c$, ASE 메모리를 $\mathbf{M}_s$, r을 I로 대신하면 식 (5), 식 (11) 및 식 (12)에 의해 비평모듈은

$$\Delta \mathbf{M}_c(\mathbf{x(k-1)}) = \alpha \hat{r}(x(k-1)) \sum_{j=1}^{k} \lambda^{k-j} \nabla_{Mc} I(x(j-1))$$

$$= \frac{\alpha}{N_e} \hat{r}(x(k-1)) \sum_{j=1}^{k} \lambda^{k-j} \mathbf{U}(x(j-1)) \quad (13)$$

와 같이 학습되고 제어동작 모듈은

$$\Delta \mathbf{M}_a(\mathbf{x(k-1)}) =$$

$$\frac{\beta}{N_e} \hat{r}(x(k-1)) \sum_{j=1}^{k} \delta^{k-j} \mathbf{M} \cdot \mathbf{U}(x(j-1)) \quad (14)$$

로 학습된다. 여기서 $\hat{r}(x(k-1))$는

$$\hat{r}(x(k-1)) = \begin{cases} R(f) - \mathbf{M} \cdot \mathbf{U}(x(k-1)) & \text{; 외부의 비평값을 얻을 수 있는 경우} \\ \gamma \mathbf{M} \cdot \mathbf{U}(x(k)) - \mathbf{M} \cdot \mathbf{U}(x(k-1)) & \text{; 외부의 비평값이 없는 경우} \end{cases} \quad (15)$$

로 계산할 수 있다.

Ⅳ. 차등책임 적응비평학습

Barto와 Lin등의 적응 비평학습 제어 방법의 문제점은 다수의 제어입력이 인가되는 시스템에는 적용하기 어렵다는 것이다. 예를 들면, 5 개의 관절을 가진 보행로봇의 경우, 5 관절 중 1 관절에 대한 잘못된 제어로도 로봇은 넘어질 수있다. 기존의 방법에서는 잘못된 1 관절의 제어방법을 수정하기 위해서 제어가 잘된 4 개의 관절들에 대한 제어도 비판하여 수정 학습하게 되므로 학습이 어려워지게 된다. 만약 비평모듈 내의 비평함수를 제어입력인자의 함수로 만든다면, 비평함수의 부분 미분 값을 이용하여 제어가 잘못된 관절만을 비평학습 할 수있게된다. 이를 위하여 제어입력인자를 비평모듈의 입력으로 사용하여 비평 값을 학습케 함으로써 비평함수가 제어입력 인자의 함수로 구성되게 하였다.

그림 3은 차등책임 적응비평을 위한 학습구조로서 비평모듈에 시스템의 상태 벡터 **X(k)**와 제어 입력 벡터 **A(k)** 가 인가된다. 또한 각 제어인자에 대한 비평책임 값은 비평모듈 내에 설치된 부분 미분기를 이용함으로써 각 제어입력 인자에 대한 비평성분 벡터를 계산하며 이 비평 값을 동작 모듈 학습에 활용한다. 상태 벡터를 $\boldsymbol{X(k)}=[x_1(k), x_2(k), x_3(k). \ . \ .]$라 하고 제어 입력 벡터를 $\boldsymbol{A(k)}=[a_1(k), a_2(k), a_3(k), \ , \ ,]$ 라고 하면 식 (12)으로부터 비평 값 r(k)는

$$r(k) = C\left(\begin{bmatrix} \boldsymbol{X(k)} \\ \boldsymbol{A(k)} \end{bmatrix}\right) = \frac{1}{N_e} \mathbf{M} \cdot \mathbf{U}\left(\begin{bmatrix} \mathbf{X(k)} \\ \mathbf{A(k)} \end{bmatrix}\right) \quad (16)$$

로서 $\boldsymbol{X(k)}$와 $\boldsymbol{A(k)}$ 의 함수로 표현할 수 있다. 여기서 C는 비평함수이다.

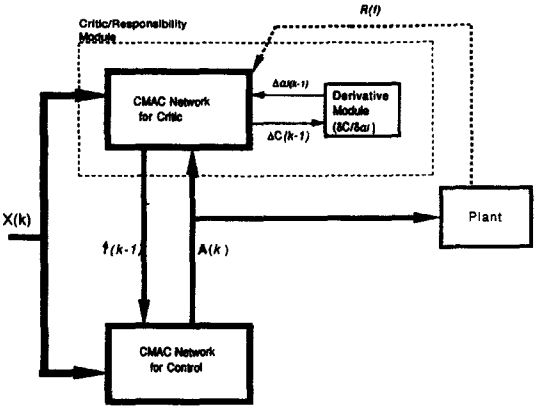


그림 3. 제안한 차등책임 적응비평학습 구조
Fig. 3. Proposed DRACL structure.

이 비평 값을 제어 입력인자 a_i로 부분 미분한 값을 d_i라고 하면, 시간 k-1에서의 부분미분 값은

$$d_i = \frac{\partial C\left(\begin{bmatrix} \boldsymbol{X(k-1)} \\ \boldsymbol{A(k-1)} \end{bmatrix}\right)}{\partial a_i(k-1)} = \frac{\partial\left\{\frac{1}{N-e} \mathbf{M} \cdot \mathbf{U}\left(\begin{bmatrix} \mathbf{X(k-1)} \\ \mathbf{A(k-1)} \end{bmatrix}\right)\right\}}{\partial a_i(k-1)} \quad (17)$$

로 표현할 수 있다. 여기서 d_i 는 비평값에대한 제어입력 인자 a_i의 기여 정도를 의미하므로 제어입력인자 a_i를 위한 메모리의 갱신값 $\hat{r}_i(k-1)$는 $\hat{r}(k-1)$에 d_i 에 의한 가중치가 곱해진 값을 사용할 수 있다. 즉

$$\hat{r}_i(k-1) = \frac{d_i}{\|\boldsymbol{D}\|} |\hat{r}(k-1)| \quad (18)$$

이다. 여기서 **D** 는 d_i 로 구성된 벡터이다. 따라서 동작모듈의 메모리 갱신은 제어 입력 인자 a_i 에 대해 개별적으로 수행해야하며 갱신식 (14)에서의 $\hat{r}(k-1)$ 대신 $\hat{r}_i(k-1)$ 값을 사용하여 다음과 같은 식으로 표현할 수 있다.

$$\Delta \mathbf{M}_{ai}(\mathbf{x}(k-1)) = \frac{\beta}{N_e}\hat{r}_i(x(k-1))\sum_{j=1}^{k}\delta^{k-j}\,\mathbf{I}(x(j-1))\cdot\mathbf{U}(x(j-1)) \quad (19)$$

V. 실험 및 검토

제안한 차등책임 적응비평 학습구조가 여러 개의 제어 입력을 갖는 시스템의 자율학습 제어에 활용될 수 있음을 보이기 위해서 2 차원 Cart-Pole 시스템과 인간 구조로봇의 앉는 동작 중의 균형유지 제어 문제를 대상으로 컴퓨터 시뮬레이션 하였다. 기존의 방법과의 학습 성능비교를 위해서 CMAC 신경회로망을 기반으로 두 학습 구조를 구성하였으며 시뮬레이션 결과를 제시하였다.

1. 2-D Cart-Pole 시스템에 대한 학습제어: 2 개의 제어입력 가진 시스템 제어

[2-D Cart-Pole 시스템]

기존의 Cart-Pole 시스템은 일 차원 축상을 움직일 수 있는 대차 위에 동일한 방향으로 자유롭게 회전할 수 있는 막대가 세워진 시스템이다. 2-D Cart-Pole 시스템은 일차원 Cart-Pole 시스템이 단순 확장된 시스템으로 X와 Y의 두 방향으로 움직일 수 있는 대차 위에 앞 뒤 좌우로 자유롭게 회전할 수 있도록 하는 막대가 설치된 시스템이다. 이 시스템에서의 상태변수는 수직 축과 막대사이의 각도(θ_x , θ_y), 각속도 ($\dot{\theta}_x$, $\dot{\theta}_y$), 및 수평축 상의 위치 (h_x , h_y), 그리고 속도($\dot{h}_x$, $\dot{h}_y$) 등이다. 여기서 아래첨자 x, y는 각 각 x 축 방향 및 y 축 방향을 의미한다. 또 대차와 막대의 질량을 각 각 m_c 및 m_p라고 하고 막대의 길이를 l 이라고 하면 이 시스템은 4개의 조인트를 갖는 로봇 (2 개의 직선운동 조인트와 2 개의 회전 조인트) 으로 간주할 수 있으며, 로봇의 동력학 방정식 유도과정을 거치면[16] 다음 식을 얻을 수 있다.

$$F_x = (2m_c + m_p)\ddot{h}_x + l m_p \cos\theta_x \cos\theta_y \ddot{\theta}_x - l m_p \sin\theta_y \sin\theta_x \ddot{\theta}_y - l m_p(\cos\theta_y \sin\theta_x \dot{\theta}_x^2 + 2\cos\theta_x \sin\theta_y \dot{\theta}_x \dot{\theta}_y + \cos\theta_y \sin\theta_x \dot{\theta}_y^2) \quad (20a)$$

$$F_y = (m_c + m_p)\ddot{h}_y + l m_p \cos\theta_y \ddot{\theta}_y - l m_p \sin\theta_y \dot{\theta}_y^2 \quad (20b)$$

$$\tau_x = l m_p \cos\theta_x \cos\theta_y \ddot{h}_x + \frac{4}{3}(1-\sin^2\theta_y) m_p l^2 \ddot{\theta}_x - \frac{8}{3} m_p l^2 \cos\theta_y \sin\theta_y \dot{\theta}_x \dot{\theta}_y - g l m_p \cos\theta_y \sin\theta_x \quad (20c)$$

$$\tau_y = -l m_p \sin\theta_x \sin\theta_y \ddot{h}_x + l m_p \cos\theta_y \ddot{h}_y + \frac{4}{3} m_p l^2 \ddot{\theta}_y + \frac{4}{3} m_p l^2 \cos\theta_y \sin\theta_y \dot{\theta}_x^2 - g l m_p \cos\theta_x \sin\theta_y \quad (20d)$$

[2-D Cart-Pole 시스템의 제어 시뮬레이션]

2-D Cart-Pole시스템의 제어시뮬레이션을 위해서 $\tau_x = \tau_y = 0$, $g = 9.8 m/s^2$, $m_c = 1.0 Kg$(카트의 중량), $m_p = 0.1 Kg$ (막대의 중량), $l = 1\,m$ (막대의 길이)를 사용하였다. 이 시스템의 제어 학습을 위해서 x 축 방향과 y 축 방향에 CMAC 신경회로망을 따로 사용하였으며 그 들의 출력을 각 각 a_x와 a_y라고 하여 여기에 잡음(noise)이 더해진 값이 0.0 보다 크면 F_x 혹은 F_y에 + 10 N의 힘을 인가하였고, 0.0 보다 작으면 - 10 N의 힘을 인가하였다. 여기서 잡음(noise)으로는 분산 값이 σ^2 인 랜덤 잡음을 사용하였다. 동작모듈에 인가되는 입력은 x, y축 방향으로 기울어지는 막대의 각도와 각 속도 및 x, y 방향에 대한 카트의 이동속도등 6개로 하였으며 비평모듈에 인가되는 입력은 동작모듈의 입력에 시스템의 x 및 y 방향의 제어입력을 추가하여 8 개로 하였다. 또한 각 상태변수의 범위는 $\theta_x, \theta_y \in (-12^\circ, 12^\circ)$로 제한하여 이 범위를 벗어나는 경우 시스템 제어의 실패로 간주하여 벌칙 신호 (-1)로서 비평 CMAC를 학습시켰다. 이 때의 메모리 갱신을 위한 이전의 상태의 수는 N_u개로 제한했다. 실험에 사용된 각종 파라미터 들은 표 1과 같다.

표 1. 시뮬레이션에 사용된 파라미터의 값
Table 1. Parameters used for the simulation.

	α	β	γ	σ	N_b	N_e	N_u
값	0.1	0.5	0.95	0.01	2	12	12

이 시뮬레이션에서 제어 모듈로부터 제어신호가 시스템에 인가되는 것을 1회의 제어동작이라하고 100,000회 동안(2.78 시간에 해당) 실패 없는 제어가 계속되는 경우 성공으로 간주한다. 그리고 제어동작을 시작하여 시스템이 실패에 이르거나 성공할 때까지의 일련의 제어과정을 1 회의 제어시도(control trial)라고 한다. 또한 제어가 성공하거나 500,000회의 제어동작이 끝날 때까지를 1 회의 제어실험이라고 한다. 다

양한 제어 실험이 되게 하기 위해서 매 회의 제어실험이 끝날 때마다 CMAC 신경회로망의 메모리를 다르게 할당하였으며 각 제어 시도마다 잡음을 랜덤하게 발생시켜 사용하였다. 기존의 학습구조와 제안한 학습구조에서의 비평모듈과 동작모듈에는 모듈마다 10K바이트의 메모리를 할당하여 사용하였다.

시뮬레이션은 Cart-Pole 시스템이 ($\theta_x=0^o$, $\theta_y=0^o$, $\dot{\theta}_x=0^o/\sec$, $\dot{\theta}_y=0^o/\sec$ $h_x=0\ m$, $h_y=0\ m$, $\dot{h}_x=0\ m/\sec$, $\dot{h}_y=0\ m/\sec$)인 초기 상태 $\mathbf{X}(0)$로부터 시작하였다. 각 상태 값을 그림 3의 동작모듈에 인가하면 출력 벡터 $\boldsymbol{A}(k)=\begin{bmatrix} a_x(k) \\ a_y(k) \end{bmatrix}$를 얻을 수 있으며 이 값에 랜덤 잡음(noise)을 더한다음 식 (21)과 같이 카트를 x축 방향과 y 축 방향으로 미는 힘 F_x 및 F_y를 결정하였다.

$$\begin{aligned} &\textit{If} \quad a_x[k]+noise > 0.0, \quad \textit{then} \quad F_x=10N, \\ &\textit{otherwise}, \quad F_x=-10N \\ &\textit{If} \quad a_y[k]+noise > 0.0, \quad \textit{then} \quad F_y=10N, \\ &\textit{otherwise}, \quad F_y=-10N \end{aligned} \tag{21}$$

여기서 얻어진 힘 F_x 및 F_y는 시스템 모델식 (20a) - (20d)에 인가하여 새로운 상태 $\mathbf{X}$(k+1)를 결정한다. 또한 힘으로 변환되기 이전의 제어 신호 $\boldsymbol{A}(k)$는 상태벡터 $\mathbf{X}$(k)와 함께 비평모듈에도 인가함으로써 (16)식에서의 상태 비평값 r(k)를 얻는다. 한편 식 (6)으로부터 $\hat{r}(k-1)$는 r(k)와 r(k-1)로부터 얻을 수 있으며 식 (17)에서의 개별 제어값에 대한 비평값의 변화율 d_i는 비평모듈의 입력에 제어신호를 Δa_i 만큼 변화시켜 인가하여 얻는 $r(k)$의 증분값 $\Delta r(k)$과 Δa_i 값과의 비 $\frac{\Delta r(k)}{\Delta a_i(k)}$로 계산하였다. 이 값은 식 (18) 및 (19)에서의 비평모듈 학습에 이용되었다. 또한 동작모듈의 학습은 단일 입력 시스템제어와 같이 식 (13)을 이용하였으며 비평모듈과 동작모듈의 학습은 제어동작에 의해 시스템의 상태가 갱신될 때마다 on-line 학습이 되게 하였다.

그림 4는 제안한 적응비평 제어 방법의 시뮬레이션 결과를 기존의 적응비평 방법의 결과와 비교한 것이다. 그림의 수평축은 제어 시도의 횟수이고 수직 축은 제어의 성공 횟수를 나타낸다. 기존의 적응비평학습의 초기에는 제어 기술이 일부 학습되어 제어가 성공하는 경우가 잦았으나 학습이 더 진행되어도 제어성능이 개선되지 않고 있음을 보여준다. 그 이유는 어느 한 제어 입력인자의 잘못으로인해 제어가 실패할 경우, 잘된 제어입력인자의 제어까지 비평하여 반대 학습을 하기 때문으로 해석된다. 그러나 제안한 제어구조에서는 비평함수로부터 각 각의 제어입력 인자 별로 잘되고 잘못된 제어를 판단 할 수 있으므로 학습이 계속될수록 제어기술을 안정적으로 습득해 가고 있음을 보여준다. 또 학습의 초기에 제안한 제어구조의 학습 성능이 기존의 구조보다 더딘 이유는 제안한 구조의 비평모듈의 입력변수의 수가 기존의 구조에 비해 많기 때문에 비평함수 형성에 시간이 더 걸리기 때문으로 판단된다.

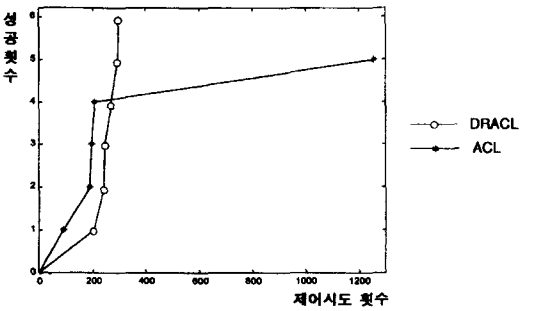


그림 4. 2-D Cart-pole 시스템에 대한 제안한 학습구조와 기존의 학습구조간의 성능비교

Fig. 4. Comparison of the learning performance between the proposed and the previous learning structure for the 2-D cart-pole system

2. 인간구조 로봇의 앉는 동작에 대한 학습제어: 3개의 제어 입력 가진 시스템 제어

[인간구조 로봇의 앉는 동작 제어문제]

3 개의 제어 입력을 가진 시스템의 제어 실험을 위해서 그림 5와 같은 인간 구조로봇의 앉는 동작 학습문제를 대상으로 하였다.

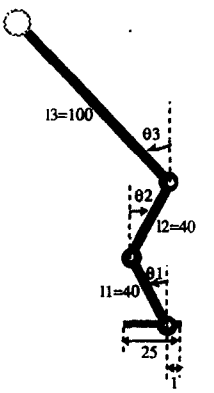


그림 5. 인간구조 로봇

Fig. 5. Human-like-structured robot.

이 로봇 학습제어의 목표는 서있는 자세에서 앉는 동작을 하는 동안에 로봇의 무게 중심을 앞뒤로 적절히 이동하면서 넘어지지 않게 제어하는 방법을 스스로 학습하는 것이다. 이 로봇은 발과 정강이, 무릎 및 몸통에 해당하는 4개의 링크로 구성되며 구조에 대한 제원은 표 2와 같다.

표 2. 인간구조 로봇구조에 대한 제원
Table 2. Specification of the human-like-structured robot model.

링크의 길이(cm)	무게 (Kg)	조인트의 동작 범위
발: 25	발: 0	$-45^\circ \le \theta 1 \le 45^\circ$
정강이: 40	정강이: 3	$-180^\circ \le \theta 2 \le 0^\circ$
무릎: 40	무 릎: 5	$-40^\circ \le \theta 3 \le 140^\circ$
몸통: 100	몸 통: 10	

[인간구조 로봇의 앉는 동작 제어 시뮬레이션]

로봇의 제어과정에서 로봇이 넘어지면 서있는 자세에서부터 다시 출발하게 하여 제어기술을 학습케 하였다. 로봇의 제어 결과 몸통 끝 부분의 높이가 10cm이하가 되면 성공으로 간주하여 보상 값 r=1.0을 인가하고 몸의 전체 무게 중심이 발의 범위(25cm)를 벗어나면 넘어져서 실패한 것으로 간주하여 비평값 p= -1.0을 인가하였다. 매 제어실험 마다 최대의 제어회수를 300 회로 제한하여 그 동안에 성공하지 못하는 경우도 실패로 간주하였다. 또한 학습시간의 단축을 위해서 엉덩이 부분의 높이가 10cm 만큼 낮아지거나 높아지면 보조 보상(subsidiary reward) 값 r'=0.5와 보조 비평(subsidiary penalty) p'= -0.5가 인가되게 하였다.

이 로봇 학습제어를 위해서 제안한 구조의 비평모듈에 사용된 입력은 3 개의 상태와 3 개의 제어 입력 등 6 개이며 동작모듈에 사용한 입력은 3 개 였다. 또한 학습시 사용된 파라미터는 다음 표 3과 같다.

표 3. 그림 5의 로봇 제어학습 시 사용한 파리미터의 값
Table 3. Parameters used for the learning control of the robot in Fig. 5.

	α	β	γ	σ	N_b	N_ϵ
값	0.01	0.5	0.98	0.01	5	7

학습된 제어구조를 이용한 로봇제어시 i 번째 조인트의 각도는 동작 모듈의 출력값 $[\boldsymbol{A}(k)]_i$에 따라 다음과 같이 매 제어시 마다 인가되게 하였다.

$$\Delta\theta_i = \begin{cases} 1^\circ & , \quad \text{if}([\boldsymbol{A}(k)]_i + noise) \ge 0.0 \\ -1^\circ & , \quad otherwise \end{cases}$$

제안한 차등책임 적응비평 제어 효과를 검증하기 위해서 CMAC에 의해 구성한 기존의 적응비평 구조와 제어 시뮬레이션을 통해 비교하였다. 여기서 매 제어 시뮬레이션마다 랜덤하게 발생시킨 잡음을 인가함으로써 다른 종류의 시뮬레이션이 되게 하였다. 그림 6의 수평축은 제어시도 횟수를 나타내며 수직 축은 최근 20회 시도 동안의 성공한 횟수를 나타낸다. 제안한 제어구조를 사용할 경우, 180회의 시도(trial) 이상에서는 평균 19회 정도의 성공 횟수를 보이는 반면, 기존의 방법에서는 평균 10회 내외의 성공 횟수를 보이고 있다.

학습 초기에는 제안한 제어구조의 학습속도가 기존의 구조에 비해 약간 더디지만 학습이 진행됨에 따라 점차 우수한 학습 제어효과를 보이게 되는 현상은 2-D Cart-Pole 제어 시뮬레이션의 경우와 같다.

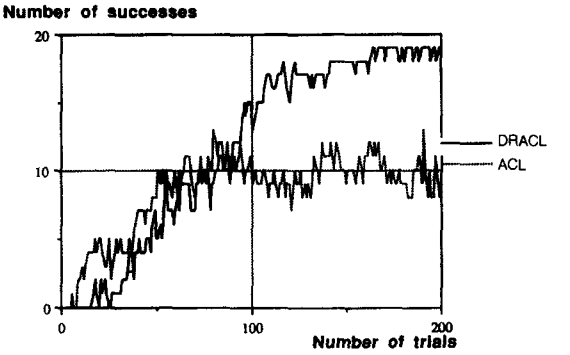


그림 6. 인간 구조 로봇의 앉는 동작에 대한 제어성능 비교
Fig. 6. Comparison of the control performance about the human-like-structured robot's squatting motion.

VI. 결 론

누적된 제어동작 끝에 얻어지는 성공과 실패 등의 애매한 정보를 이용하여 제어기술을 학습해야하는 경우에 적응비평 학습이 효과적이다. 기존의 적응비평 학습들은 한 개의 제어 입력이 필요한 시스템의 제어에 국한되어 적용되었다. 본 논문에서는 제어입력 인자별로 비평 값을 산출하여 재 강화 학습에 이용함으

로써 다수의 제어입력이 필요한 시스템에 대해서도 효과적인 학습이 가능하게 하였다.

제안한 방법은 비평함수가 제어입력 신호의 함수가 되게하여 제어 입력 인자 별로 책임져야할 비평 값을 구하여 학습하는 방법이다. 이를 위하여 적응비평 학습구조의 비평모듈에 제어 입력 벡터를 인가함으로써 비평 값이 제어 입력 인자들의 함수가 되게 하였다. 이 경우, 비평함수는 개별 제어 입력 인자에 대한 부분 미분이 가능하게 되며 각 제어 입력 인자에 대한 제어기술 학습에는 부분 미분 값에 비례하는 비평 값을 이용할 수 있게된다.

제안한 학습구조를 여러 개의 제어 입력을 갖는 시스템에 적용시 효과적임을 보이기 위해서 2 개의 제어입력을 갖는 2-D Cart-Pole 시스템과 3 개의 제어입력을 갖는 인간형 구조 로봇의 앉는 동작에 대한 균형제어에 이용하였다. 두 실험에서 기존의 학습구조는 학습이 계속되어도 제어성능이 개선되지 않는 반면 제안한 학습구조는 안정적으로 제어기술을 습득하여 제어성능이 점차 나아져 가고 있음을 보여주었다. 다만 비평모듈에 시스템의 상태변수 뿐아니라 제어 입력 변수 까지 인가해야 하므로 입력변수의 수가 늘어나게 되어 초기의 학습이 더디다는 문제가 있었다.

본 연구에서는 제안한 학습구조를 CMAC 신경회로망을 이용하여 구축하였지만 다층신경회로망등 다른 종류의 신경회로망을 이용해서도 구성할 수 있는 일반적인 구조이다. 향후의 연구과제는 제안한 학습구조를 2 족 보행 로봇 제어와 같은 실제적인 문제에 적용하는 일이다.

감사의 글

※ 이 논문은 1996년도 전북대학교의 학술 연구 조성비 지원에 의해 연구 되었습니다. 이에 감사드립니다.

참 고 문 헌

[1] R. R. Bush and F. Mosteller, *Stochastic Models for Learning* , New York: Wiley, 1958.

[2] K. S. Narendra and M. A. L. Thathachar, "Learning automata - A survey," IEEE Trans. on Systems, Man, and Cybernetics, vol. SMC-4, no. 4, pp. 323-334, July 1974.

[3] V. I. Varshavski and I. P. Vorontsova, "On the behavior of stochastic automata with variable structure," Automat. Telemekh., vol. 24, pp. 353-360, Mar. 1963.

[4] I J. Shapiro and K. S. Narendra, "Use of stochastic automata for parameter self-optimization with multimodal performance criteria," IEEE Trans. on Systems, man and Cybernetics, vol. SSC-5, pp. 352-360, Oct. 1969.

[5] M. D. Waltz and K. S. Fu, "A heuristic approach to reinforcement learning control systems," IEEE Trans. Automatic Control, vol. AC-10, no. 4, pp. 390-398, Oct. 1965.

[6] Z. J. Nikolic and K. S. Fu, "An algorithm for learning without external supervision and its application to learning control systems," IEEE Trans. on Automatic Control, vol. AC-11, no. 3, pp. 414-422, July 1966.

[7] L. E. Jones, "On the choice of subgoals for learning control systems," IEEE Trans. on Automatic Control, vol. AC-13, no. 6, December 1968.

[8] L. E. Jones and K. S. Fu, "On the selection of a subgoal and the use of *a priori* information in learning control systems," Automatica, vol. 5, pp. 705-720, Pergamon Press, 1969.

[9] B. Widrow, N. K. Gupta, and S. Maitra, "Punish/Reward: Learning with a critic in adaptive threshold systems," IEEE Trans. on Systems, Man, Cybernetics, vol. 5, pp. 455-465, September 1973.

[10] A. H. Klopf, *The Hedonistic Neuron: A Theory of memory, Learning, and Intelligence*, Washington, DC: Hemisphere, 1982.

[11] A. G. Barto, R. S. Sutton and C. W. Anderson, "Neuronlike adaptive elements that can solve difficult learning control

problems, "IEEE Trans. on Systems, Man, and Cybernetics, vol. SMC-13, no. 5, pp. 834-846, Sept./Oct. 1983.

[12] C. W. Anderson, "Strategy learning with multilayer connectionist representations," Proceedings of the Fourth International Workshop on Machine Learning, Irvine, CA, pp. 103-114, 1987.

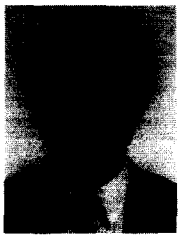
[13] J. S. Albus, *Theoretical and experimental aspects of a cerebellar model*, Ph. D. Thesis, U. of Maryland, 1972.

[14] C. -S. Lin and H. Kim, "CMAC-based Adaptive Critic Self-learning Control," IEEE Transactions on Neural Networks, vol. 2, no. 5, pp. 530-533, September 1991.

[15] C. -S. Lin and H. Kim, "CMAC-based Adaptive Critic Learning and Selection of Its Learning and Structure Parameters," IEEE Transactions on Neural Networks, vol. 6, no. 3, pp. 642-647, May, 1995.

[16] W. A. Wolovich, robotics: Basic Analysis and Design, CBS College Publishing, 1987.

저 자 소 개



金 炯 爽(正會員)

1956년 1월 21일생. 1980년 한양대 전자공학과 졸업. 1982년 전북대 대학원 전기공학과 졸업(석사). 1992년 Dept. of Electrical and Computer Eng., University of Missouri, Columbia (Ph.D.). 1982년 ~ 1993년 국방과학연구소 선임연구원. 1993년 ~ 현재 전북대 전기·전자·제어공학부 조교수. 주관심분야는 신경회로망 모델 개발, VLSI에 의한 신경회로망 하드웨어 구현, 로봇비젼 및 무인헬기 원격제어