한국 퍼지 및 지능시스템학회 논문지 2000, Vol. 10, No. 5 (2000)395~400

강화 학습에 기반한 뉴로-퍼지 제어기

Neuro-Fuzzy Controller Based on Reinforcement Learning

박영철 · 심귀보

Young-Chual Park and Kwee-Bo Sim

중앙대학교 전자전기공학부

요 약

본 논문에서는 강화학습에 기반한 새로운 뉴로-퍼지 제어기를 제안한다. 제안한 시스템은 개체의 행동을 결정하는 뉴로-퍼지 제어기와 그 행동을 평가하는 동적 귀환 신경회로망으로 구성된다. 뉴로-퍼지 제어기의 후건부 소속함수는 강화학습을 한다. 한편, 유전자 알고리즘을 통하여 진화하는 동적 귀환 신경회로망은 환경으로부터 받는 외부 강화신호와 로봇의 상태로부터 내부강화 신호를 만들어 낸다. 이 출력(내부강화신호)은 뉴로-퍼지 제어기의 교사신호로 사용되어 제어기가 학습을 지속하도록 만든다. 제안한 시스템은 미지의 환경에서 제어기의 최적화 및 적응에 사용할 수 있다. 제안한 알고리즘은 컴퓨터 시뮬레이션 상에서 자율 이동로봇의 장애물 회피에 적용하여 그 유효성을 확인한다.

ABSTRACT

In this paper, we propose a new neuro-fuzzy controller based on reinforcement learning. The proposed system is composed of neuro-fuzzy controller which decides the behaviors of an agent, and dynamic recurrent neural networks(DRNNs) which criticise the result of the behaviors. Neuro-fuzzy controller is learned by reinforcement learning. Also, DRNNs are evolved by genetic algorithms and make internal reinforcement signal based on external reinforcement signal from environments and internal states. This output(internal reinforcement signal) is used as a teaching signal of neuro-fuzzy controller and keeps the controller on learning. The proposed system will be applied to controller optimization and adaptation with unknown environment. In order to verify the effectiveness of the proposed system, it is applied to collision avoidance of an autonomous mobile robot on computer simulation.

1. 서 론

일반적으로 기계학습은 교사신호의 유무에 따라 교사 학습과 비교사 학습, 그리고 간접교사에 의한 강화학습으로 크게 분류할 수 있다[1]. 강화학습[1-4]이란 용어는 실험 심리학에서 동물의 학습방법 연구에서 비롯되었으나, 최근에는 공학 특히 인공 지능분야에서 신경회로망의 학습 알고리즘으로 많은 관심을 끌고 있다. 강화 학습은 일반적으로 제어기 또는 에이전트의 행동에 대한 보상을 최대화하는 상태-행동 규칙이나 행동 발생 전략을 찾는 것이다. 본래 강화신호인 보상과 벌칙은 학습자의 외부로부터, 이를테면 교사에 의해 주어지는 것이라고 하는 의미가 있지만, 인공생명의 입장에서는 입력자극에 대한 학습자 측의 해석으로서 좋고 나쁨을 판단하는 것으로 생각하는 편이 자연스럽다. 기계학습의 분야에서 행해지는 강화학습 연구는, 이산 시간에서의 입력자극의 샘플링, 의사결정, 그리고 행동의 반복 등의 구조에 기반 하는 것이 대부분이다. 이것은 수리적으로 마르코프(Markov) 결정 문제에 해당한다. 일반적으로 동물이나 로봇의 학습에서는 보수나 벌칙은 어떤 길이의 행동계열의 실행결과로서 얻어지는 경우가 많고, 학습자의 의사결정에 대해서 그 직후에 평가가 얻어지는 경우는 없다. 교사신호가 있는 패턴분류 학습이나 그 외의 다른 학습법에서는 학습자의 의사결정의 직후에 그것에 대한 평가가, 정답인가 아닌가가 주어지지만, 강화학습에서는 행동에 대한 평가가 어느 정도의 시간적인 지연을 수반한다. 이러한 경우 지속적인 학습을 위해서는 과거 행동에 대한 신뢰 평가가 이루어져야 한다. 이것을 신뢰할당문제(Credit-Assignment Problem)라고 하며 강화 학습에 있어서 매우 중요하다. 이 문제에 대한 가장 일반적인 접근 방법은 강화 신호를 생성하는 외부 평가 함수보다 더 자세한 정보를 얻을 수 있는 내부 평가 함수를 구현하는 것이다.

퍼지시스템과 신경회로망의 융합 모델은 인간의 사고 작용과 같이 학습하고 추론하며 기억하고 인식하는 등의 기능을 하는 보다 지능적인 시스템을 만드는 것을 목표로 하는 모델이다. 퍼지 시스템은 구조적 지

식을 수치적으로 표현하며, 신경회로망 시스템은 비구조적 지식을 수치적으로 표현하는 시스템이다. 그러나 인간의 지식 정보처리 과정은 전문가 시스템이나 퍼지 시스템, 신경회로망도 아닌 여러 지식 정보처리의 통합체이므로 신경회로망과 퍼지 시스템의 융합은 매우 중요하다. 본 논문에서는 퍼지 시스템과 신경회로망의 융합 모델 즉, 뉴로-퍼지 모델[5,6]을 사용한다. 한편 학습방법으로는 강화학습을 이용한다.

본 논문에서는 뉴로-퍼지의 추론 및 학습능력에 의하여 자율이동 로봇의 행동을 결정하고 자율이동로봇의 행동과 주변 상태의 변화를 동적 귀환 신경회로망[6]의 입력으로 하여 뉴로-퍼지 제어기에 의한 자율이동로봇의 행동 결정에 대하여 평가를 한다. 만약 뉴로-퍼지 제어기의 행동결정이 올바른 방향으로 이루어졌다면, 보상이라는 값을 동적 귀환 신경회로망으로부터 입력으로 받아 뉴로-퍼지 제어기의 학습이 이루어지며, 그 반대의 경우에는 벌칙이라는 값을 입력받아 학습을 한다. 한편 동적 귀환 신경회로망의 연결강도는 유전자 알고리즘에 의하여 최적의 연결강도를 찾아 최적의 로봇 행동에 대한 평가가 이루어지도록 한다. 본 논문에서는 동적 귀환 신경회로망으로 구성된 평가 네트웍은 진화를 시키고 뉴로-퍼지 제어기는 강화학습을 하는, 진화와 학습을 이용한 알고리즘을 제안한다. 2장에서 내부강화 신호를 발생시키는 평가 네트웍으로서 동적 귀환 신경회로망을 설명하고, 3장에서 뉴로-퍼지 제어기 및 학습 알고리즘을 소개한다. 4장에서 시뮬레이션을 통하여 유효성을 검증하며, 5장에서 결론을 맺는다.

2. 평가 네트웍의 구성

2.1 동적 귀환 신경회로망에 의한 내부 강화 신호 생성

동적 귀환 신경회로망(DRNN)[6]은 환경으로부터 교사 신호의 일종인 외부 강화 신호와 에이전트의 상태를 입력받아 뉴로-퍼지의 학습 지표가 되는 내부 강화 신호 값을 발생한다. 내부 강화 신호에 의해 뉴로-퍼지 제어기는 학습을 하여 환경에 적합하게 뉴로-퍼지 후건부 소속 함수를 변화시킨다(그림 1). 본 논문에서는 교사신호를 내보내는 동적 귀환 신경회로망의 연결강도를 유전자 알고리즘에 의해서 찾아 최적의 교사 신호를 발생하도록 한다.

이와 같은 구성은 인간의 추론능력과 적응능력을 모방한 것으로서 예상치 못한 환경이나 여러 가지 환경의 잡음에 강건하기 때문에 복잡한 모델링 없이 변화하는 환경에 능동적으로 대처할 수 있다. 기존 전 방

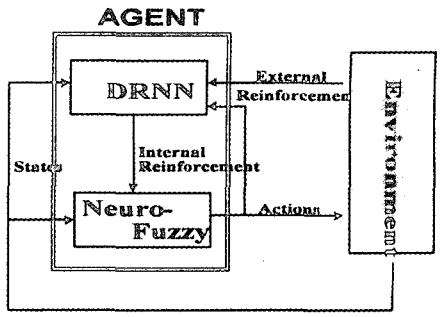


그림 1. 뉴로-퍼지 제어기 및 강화 학습의 개념도
Fig. 1. Conceptual Diagram of Neuro-Fuzzy Controller and Reinforcement Learning

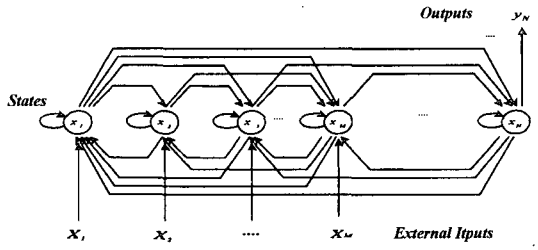


그림 2. 동적 귀환 신경회로망 구조
Fig. 2. Structure of Dynamic Recurrent Neural Networks

향 네트워크는 구조상 정적인 사상만을 학습할 수 있으므로 시변 신호나 시계열 동적 사상만을 학습하기 위해서는 제어기에 시스템의 입력을 Tapped Delay Line으로 구성을 하거나, 시스템의 입력을 신경회로망 입력으로 피드백을 하는 등의 변형이 필요하다. 반면 동적 귀환 신경회로망은 내부적으로 상태 피드백이 있기 때문에 입·출력 사이의 동적 사상을 학습할 수 있다[3]. 그림 2는 동적 귀환 신경회로망의 구조이다. 그림 2에서 보인 동적 귀환 신경회로망의 i번째 뉴런의 동적 방정식은 (1)식과 같다.

$$\tau_i \dot{y}_i = -y_i + f\left(\sum_j \omega_{ij} y_j\right) + X_i \tag{1}$$

단, ω_{ij}는 뉴런 j에서 뉴런 i로의 연결강도, τ_i는 뉴런 i의 시간 감쇠 계수, y_i는 뉴런 i의 출력, y_j는 뉴런 j의 출력, X_i는 뉴런 i의 외부 입력이 된다. $f(\cdot)$는 시그모이드 함수로서, 일반적으로 출력이 -1~1의 값이며, (2)식과 같은 함수를 사용한다.

$$f(x) = \left(\frac{2}{1+e^{-\beta x}} - 1\right) \tag{2}$$

단, x는 뉴런의 Net 입력이고, β는 시그모이드 함수의 기울기이다.

동적 귀환 신경회로망에서 얻어지는 출력 값은 (3)식에 의해 최종적으로 내부 강화 신호를 출력한다.

$$\hat{r}[t+1]=\begin{cases} r(t+1)-y[t,t] & \text{(if penalty)} \\ r[t+1]+(\gamma-1)y[t,t] & \text{(if reward)} \end{cases} \quad (3)$$

단, $\hat{r}$은 내부강화 신호로서 연속적인 값을 가지며, 환경으로 받는 r은 외부강화 신호, $y[t, t]$는 t시간에서의 외부입력과 t시간에서의 연결강도에 의해 계산된 동적 귀환 신경회로망의 출력 값이고, γ는 할인율로서 임의의 양수 값을 갖는다. 뉴로-퍼지 제어기는 내부 강화 신호인 $\hat{r}$에 의해 학습이 이루어진다.

2.2 동적 귀환 신경회로망의 가중치 코드화

동적 귀환 신경회로망의 각 연결강도 값은 유전자 알고리즘에 의해 선택, 교차, 돌연변이의 일련의 과정을 통해 다음 세대의 새로운 연결강도 값을 결정하게 된다. 즉, 최적의 연결강도 값을 진화를 하여 찾게 된다. 본 논문에서 동적 귀환 신경회로망의 각 노드간의 가중치는 유전자 알고리즘의 비트 열로 하여 진화를 시킨다. 즉, 가중치를 [(w_{11}, w_{12}, w_{13}, w_{14}, w_{15}, w_{16}, w_{17}), (w_{21}, w_{22}, w_{23}, w_{24}, w_{25}, w_{26}, w_{27}), ⋯, (w_{n1}, w_{n2}, w_{n3}, w_{n4}, w_{n5}, w_{n6}, w_{n7})], (단, w_{ij} : j번째 노드에서 i번째 노드로의 가중치임)와 같이 코드와 하여 진화한다.

3. 뉴로 - 퍼지 제어기 설계

자율이동 로봇의 뉴로-퍼지 제어기는 동적 귀환 신경회로망으로부터 내부강화 신호와 자신의 상태를 입력받아 다음 행동을 결정한다. 뉴로-퍼지 제어기는 5개의 층으로 이루어지며, 각 층은 퍼지 연산과정에 대응된다. 뉴로-퍼지 제어기의 입력부와 출력부의 소속함수는 삼각형 소속함수를 사용한다. 각 층의 연결은 전 방향 네트워크 구조로 이루어지며 각 층의 연산은 지역적인 연산을 수행한다. 본 논문에서 사용하는 뉴로-퍼지의 추론 과정은 다음과 같다. 첫 번째 층에서 제어 대상물의 상태를 입력받는다. 두 번째 층에서는 입력 값에 대해 입력 소속함수 값을 결정한다. 세 번째 층에서는 퍼지 룰을 결정한다. 네 번째 층에서는 뉴로-퍼지의 출력에 대한 소속함수 값을 결정한다. 맨 마지막 층에서는 뉴로-퍼지의 출력 값을 계산한다. 여기에서 뉴로-퍼지에 의해서 출력된 값을 실제 제어 대상물에 바로 적용을 시키지않고 SAM(Stochastic Action Modifier)를 통과한 값을 제어대상물에 적용시킨다. SAM은 내부 강화신호에 따라 보상신호면 실제 추론값에 가깝게 벌칙신호는 추론값과 다르게 확률적

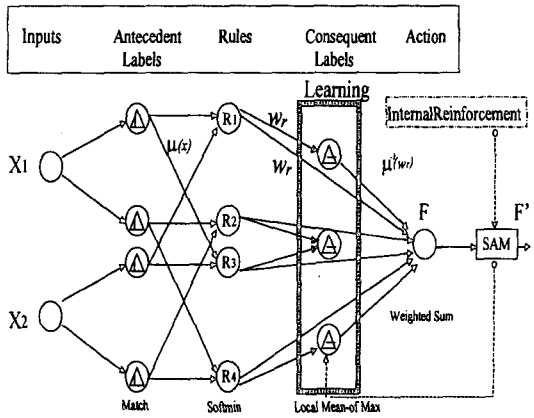


그림 3. 뉴로-퍼지 제어기
Fig. 3. Neuro-Fuzzy Controller

인 출력을 내보낸다.

한편 뉴로-퍼지 제어기는 동적 귀환 신경 회로망으로부터 받은 내부 강화 신호를 이용해 제어기 출력 소속함수의 모양을 학습한다. 그림 3은 본 논문에서 사용한 뉴로-퍼지 제어기[5]의 구조를 나타낸다.

뉴로-퍼지 제어기 각 층의 연산과정을 설명하면 다음과 같다.

layer 1:

입력 층으로서 에이전트의 상태 값을 선형화하여 뉴로-퍼지 입력으로 한다.

layer 2:

$\mu_{c_V,s_{VL},s_{VR}}(x)$: 입력 x에 대한 소속함수 값으로 두 번째 층에서 이 값을 정의한다. 이때, V는 언어항을 구별하기 위하여 사용을 하며, c_V는 소속함수의 중심 값을 나타내며, s_{VL}은 좌변, s_{VR}은 소속함수의 우변에 대한 소속함수의 폭이다. 소속함수의 모양이 삼각형인 경우 소속함수의 멤버쉽 값은 다음과 같은 연산 과정을 통하여 입력에 대한 소속함수 값을 결정한다.

$$\mu_{c,s_L,s_R}(x)=\begin{cases} 1-|x-c|/s_R, x\in[c,c+s_R] \\ 1-|x-c|/s_L, x\in[c-s_L,c] \\ 0 \quad \text{otherwise} \end{cases} \quad (4)$$

layer 3:

각 룰의 전건부에 해당하는 값에 대한 룰의 연결강도를 계산한다. layer 3에서 계산된 값을 O_R로 표시하면 (5)식과 같다.

$$O_{R3}=\omega_r=\frac{\sum_i \mu_i e^{-k\mu_i}}{\sum_i e^{-k\mu_i}} \quad (5)$$

단, i는 룰에 대한 전건부 소속함수의 입력 개수, k는

Softmin Operation의 Hardness, μ는 2번째 층에서 계산한 입력에 대한 소속함수, ω_r는 룰의 연결강도이다.

layer 4:

출력 소속함수 연결강도를 결정한다. 출력 소속함수 값은 (6)식에 의해 계산되어진다.

$$O_{V4}=\left(c_V+\frac{1}{2}(s_{VR}-S_{VL})\right)\left(\sum_r w_r\right)-\frac{1}{2}(s_{VR}-s_{VL})\left(\sum_r w_r^2\right) \tag{6}$$

단, c_V는 출력 소속함수의 중심 값이며, 여기에서 V는 사용된 후건부 소속함수들 중에서 해당되는 소속함수 표시를 나타낸다.

layer 5:

(7)식은 뉴로-퍼지에 의한 추론 값으로서 본 논문에서는 로봇의 회전각이 된다. 디 퍼지화 과정은 무게중심법을 사용하였으며, 또한 계산적으로 연산 과정은 layer 4에서 계산된 값을 layer 3에서 계산된 값으로 나눈 값과 같다.

$$F=\frac{\sum_r w_r\mu^{-1}(e_r)}{\sum_r w_r}=\frac{\sum_V O_{V4}}{\sum_R O_{R3}} \tag{7}$$

뉴로-퍼지의 출력값을 바로 이용하지 않고 다음 (8)식과 같은 연산과정을 통하여 계산된다.

$$F'(t)\approx\Psi(F(t),\sigma(t)) \tag{8}$$

F'는 F값과 표준편차 σ의 랜덤변수에 대한 값이다. 단, $\sigma(\ )$는 단조 감수로서 본 논문에서는 $\exp(-\hat{r})$를 사용한다. 즉, 내부 강화 신호가 크면 퍼지적 뉴로의 실제 출력 값을 감소시켜 자율이동 로봇의 행동을 억제하고, 역이면 자율이동 로봇의 행동을 촉진시킨다.

3.2 뉴로-퍼지 학습

뉴로-퍼지 제어기는 동적 귀환 신경회로망으로부터 내부 강화 신호를 입력받아 출력부의 삼각형 소속함수를 다음 (9), (10)식에 의해 학습을 한다[5].

$$\Delta p_V=\eta s(t)\hat{r}(t)\frac{\partial v}{\partial p_V}=\eta s(t)\hat{r}(t)\frac{\partial v}{\partial F}\frac{\partial F}{\partial p_V} \tag{8}$$

$$s(t)=\frac{F'(t)-F(t)}{\sigma(\hat{r}(t-1))} \tag{9}$$

여기에서 Δp_V는 뉴로-퍼지의 후건부 소속함수의 연결강도이며, η는 학습률로서 작은 양의 정수를 가진다. $s(t)\hat{r}(t)$는 학습 요소로서 만약 큰 외부의 잡음에 의해서 자율 이동 로봇이 좋은 행동을 취하게 되면 주어진 연결강도에 대해 보상을 더 주고, 그 반대이면 연결강도의 영향을 적게 해주기 위해 사용이 된다. V는 학습을 하려고 하는 후건부 소속함수, 중간 값(c_V)과 좌·우변(s_{VR}, s_{VL})의 폭이다. $\partial v/\partial F$는 (10)식에 의해 구하여 부호 함수로서 사용하게 된다.

$$\frac{\partial v}{\partial F}\approx\frac{dv}{dF}\approx sgn\left(\frac{v(t)-v(t-1)}{F(t)-F(t-1)}\right) \tag{10}$$

F는 뉴로-퍼지의 출력이다. F는 식 (11)에 의해 구해진다.

$$F=\frac{\sum_r w_r z_r}{\sum_r w_r} \tag{11}$$

단, $z_V(w_r)=c_V+1/2(s_{VR}-s_{VL})\ (1-w_r)$ (12)

$$\frac{\partial F}{\partial p_V}=\frac{1}{\sum_i w_i}\sum_{V=Con(R_j)}\frac{w_j\partial z_V}{\partial p_V} \tag{13}$$

$$\frac{\partial z_V}{\partial c_V}=1 \tag{14}$$

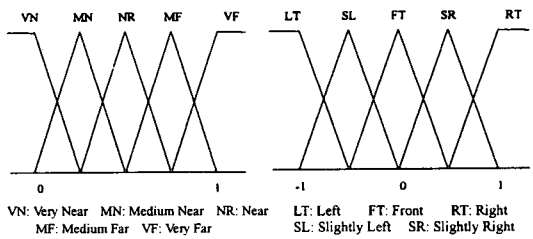
$$\frac{\partial z_V}{\partial s_{VR}}=\frac{1}{2}(1-w_r) \tag{15}$$

$$\frac{\partial z_V}{\partial s_{VL}}=-\frac{1}{2}(1-w_r) \tag{16}$$

(14)~(16)식에 의해 후건부 언어항의 값들이 주어진 환경에 적응하여 학습한다.

4. 시뮬레이션

로봇의 센서로부터 입력되는 장애물의 센싱거리에 대한 언어항과 로봇 출력 값인 로봇의 회전에 대한



(a) 센서 입력 (b)회전각

(a) Sensing input (b) Rotation angle

그림 4. 소속함수

Fig. 4. Membership Function

언어항은 그림 4와 같이 삼각형 소속 함수를 사용한다. 즉, 그림 4의 (a)는 센서입력에 대한 삼각형 소속함수이며, 그림 4의 (b)는 자율이동 로봇의 회전각에 대한 삼각형 소속함수이다. 회전에 대한 삼각형 소속함수 중간 값, 좌·우측의 폭은 학습에 의해 수시로 변화하게 한다.

미로주행의 목적은 적합도 계산식에서 알 수 있듯이 미로와의 충돌을 최소화하면서 출발점을 기점으로 멀리 이동하는 것을 목적으로 한다. 또한 나머지 파라미터의 값은 아래와 같이 설정하였다.

- 개체군의 크기 : 500
- 돌연변이률 : 0.02
- 교차율 : 0.8
- 로봇 최대 이동거리: 1590
- 로봇이 이동할 수 있는 최대 스텝: 80
- Max penalty = 마지막 스텝수의 20%
- 학습률(η) = 3이다.

그림 5는 세대에 대한 최대 적합도 변화 그래프이다. 세대가 증가할수록 적합도는 증가한다. 그러나 최대 적합도인 1에는 도달하지 않는다. 만약 1이 되려면 이상적인 내부강화 신호와 이상적인 로봇 행동에 대한 평가가 이루어져야 한다. 또한 로봇의 이동거리 계산에서 로봇의 이동거리를 미로의 중앙을 직선으로 따라서 갈 경우의 값이다. 그러므로 이상적인 상태라면 로봇은 직선으로 미로의 중앙으로 움직여야한다. 로봇과 장애물과의 거리 값 조정에 의해 영향을 받는다. 그러나 세대가 증가할수록 그림 6과 7에서 보듯이 벌칙 수는 감소하고 이동거리는 증가함을 알 수가 있다. 그림 8은 최적의 동적 귀환 신경 회로망의 연결강도 상태와 뉴로-퍼지의 학습이 이루어진 후 자율이동로봇의 최종적인 미로주행이다. 본 논문에서는 학습뿐만 아니라 학습 결과에 대한 판단여부를 결정하는 부분을 진화를 시켜 최적의 학습이 이루어지면서 최적의 평가가 이루어지도록 한다. 세대가 작으면 시

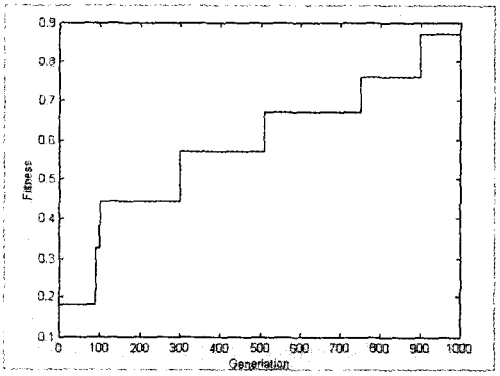


그림 5. 세대에 따른 적합도 변화
Fig. 5. Change of Fitness vs. Generation

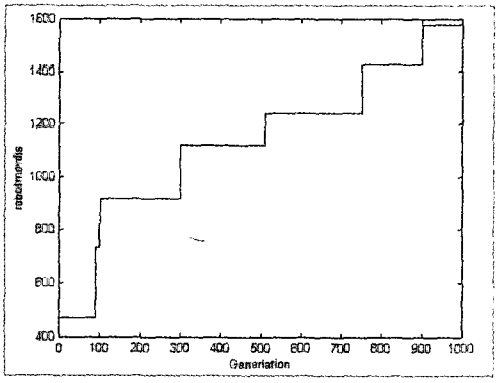


그림 7. 세대에 따른 로봇의 이동거리
Fig. 7. Moving Distance of Robot vs. Generation

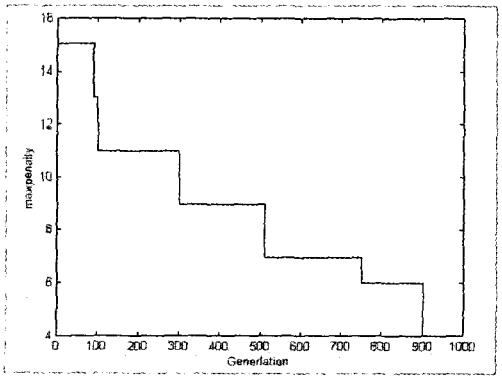


그림 6. 세대에 따른 벌칙 수
Fig. 6. Number of Penalty vs. Generation

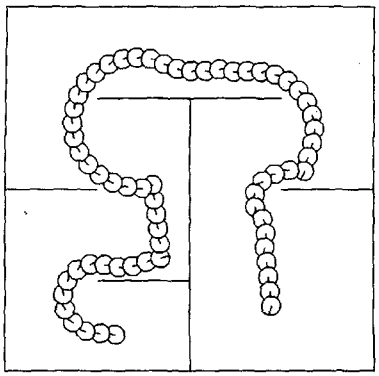


그림 8. 로봇의 궤적
Fig. 8. Trace of Robot

뮬레이션 결과에서 보듯이 학습이 이루어진다고 해도 평가하는 부분이 부적절한 상태여서 그에 따른 판단도 부적절하여 결국 그 시스템은 무용지물이 될 것이다. 세대가 지날수록 학습도 이루어지면서 평가 부분도 진화를 하게 됨을 알 수 있다.

5. 결 론

본 논문에서는 자율이동로봇의 행동 진화를 위해 동적 귀환 신경회로망에 의한 뉴로-퍼지의 강화학습 알고리즘과 유전자 알고리즘을 도입한 방법을 제안하였다. 제안한 방법은 자율 이동로봇의 행동제어기로서 자율 이동 로봇의 미로 탐색 행동에 적용시켜 그 유효성을 검증해 보았다.

참고문헌

[1] A. G. Barto, "Reinforcement Learning," *The Handbook of Brain Theory and Neural Networks*, The MIT Press, Vol. 8, pp. 804-809, 1995.

[2] F. J. Pineda, "Recurrent Back Propagation and the Dynamical Approach to Adaptive Neural Computation," *Neural Computation* 1, pp. 162-172, 1989.

[3] E. Uchibe, M. Asada, K. Hosoda, "Behavior Coordination for a Mobile Robot Using Modular Reinforcement Learning," *Proc. of Int. IROS 96*, pp. 1329-1336, 1996.

[4] Michael A. Aribs, *The Handbook of Brain Theoryand Neural Networks*, Vol. 4, pp. 796-798, 1995.

[5] Hamid R. Berenij, "Learning and Tuning Fuzzy Logic Controllers Through Reinforcements," *IEEE Trans. on Neural Networks*, Vol. 3, No. 5, pp. 724-740, 1992. 9.

[6] Chin-Teng Lin and C. S. George Lee, *Neural Fuzzy Systems : A Neuro-Fuzzy Synergism to Intelligent Systems*, Prentice Hall PTR, pp. 373-381, 1996.

박 영 철 (Young-Chual Park)

제9권 3호 참조

1998년 : 군산대학교 제어계측공학과 공학사

2000년 : 중앙대학교 제어계측과 공학석사

심 귀 보 (Kwee-Bo Sim)

제10권 3호 참조

현재 : 중앙대학교 전자전기공학부 교수