강화 학습에 기반한 뉴럴-퍼지 제어기

Neural-Fuzzy Controller Based on Reinforcement Learning

*박영철[1], 김대수[2], 심귀보[1]
*Young-Chual Park[1], Dae-Su Kim[2], Kwee-Bo Sim[1]

[1]중앙대학교 공과대학 전자전기공학부
[2]한신대학교 자연과학대학 컴퓨터학과
[1]School of Electrical and Electronic Engineering, Chung-Ang University
Tel: +82-2-820-5319, Fax:+82-2-817-0553 E-mail: kbsim@cau.ac.kr
[2]Computer Science,College of Natural Sciences, HanShin University
Tel: +82-339-370-6784, Fax:+82-339-372-3343 E-mail: daekim@hucc.hanshin.ac.kr

요약문

본 논문에서는 강화 학습 개념을 도입하여 자율이동 로봇의 성능을 개선하고자 한다. 본 논문에서 사용되는 시스템은 크게 두 부분으로 나눌 수가 있다. 즉, 뉴럴 퍼지 부분과 동적귀환 신경회로망이다. 뉴럴 퍼지 부분은 로봇의 다음 행동을 결정하는 부분이다. 또한 동적귀환 신경회로망으로부터 내부 강화 신호를 받아 학습을 하여 최적의 행동을 결정하게 된다. 동적 귀환 신경회로망은 환경으로부터 외부 강화신호를 입력으로 받아 뉴럴 퍼지의 행동결정에 대해 평가를 한다. 또한 내부강화 신호 값을 결정하는 동적 귀환 신경회로망의 웨이트는 유전자 알고리즘에 의해 진화를 한다. 제안한 알고리즘 구조를 컴퓨터 시뮬레이션 상에서 자율 이동 로봇의 제어에 적용을 함으로서 그 유효성을 증명하고자 한다.

Abstract

In this paper we improve the performance of autonomous mobile robot by induction of reinforcement learning concept. Generally, the system used in this paper is divided into two part. Namely, one is neural-fuzzy and the other is dynamic recurrent neural networks. Neural-fuzzy determines the next action of robot. Also, the neural-fuzzy is determined to optimal action internal reinforcement from dynamic recurrent neural network. Dynamic recurrent neural network evaluated to determine action of neural-fuzzy by external reinforcement signal from environment. Besides, dynamic recurrent neural network weight determined to internal reinforcement signal value is evolved by genetic algorithms. The architecture of propose system is applied to the computer simulations on controlling autonomous mobile robot.

1. 서 론

강화 학습은 일반적으로 제어기 또는 에이전트의 행동에 대한 보상을 최대화하는 상태-행동 규칙이나 행동 발생 전략을 찾는 것이다. 그러나 많은 실세계에 경우에 있어서 목표 상태에 도달 할 때까지는 중간 단계의 행동에 대한 즉각적인 보상이 주어지지 않는다. 이러한 경우 외부로부터의 강화 신호가 없기 때문에 학습이 일시적인 신뢰 할당이 이루어져야 한다. 이것을 Credit-Assignment Problem이라고 하며 강화 학습에 있어서 가장 중요한 문제라고 할 수 있으며, 이 문제에 대한 가장 일반적인 접근 방법은 강화 신호를 생성하는 외부 평가 함수보다 더 자세한 정보를 얻을 수 있는 내부 평가 함수를 구현하는 것이다. 본 논문에서는 여러 가지 학습법 중에서 강화학습을 이용한다. 즉, 뉴럴 퍼지의 추론 및 학습능력에 의하여 자율이동 로봇의 행동을 결정하고 자율이동 로봇의 행동과 주변 상태의 변화를 동적귀환 신경회로망의 입력으로 하여 뉴럴 퍼지의 자율이동 로봇의 행동 결정에 대해 평가를 한다. 만약 뉴럴 퍼지의 행동결정이 올바른 방향으로 이루어졌다면, 보상이라는 값을 동적귀환 신경회로망으로부터 입력으로 받아 뉴럴 퍼지의 학습이 이루어지며, 그 반대의 경우에는 벌칙이

라는 값을 입력으로 받게 되어 학습을 한다. 또한 동적 귀환 신경회로망의 웨이트는 유전자 알고리즘에 의하여 최적의 웨이트를 찾아 최적의 로봇 행동에 대한 평가가 이루어 지도록 한다. 즉 본 논문에서는 학습과 진화를 이용하여 자율이동 로봇의 성능을 향상시키고자 한다. 2장에서 동적 귀환 신경회로망에 의한 강화 학습을 소개하고, 3장에서 뉴럴 퍼지에 의한 행동 시스템, 4장에서 시뮬레이션 결과, 5장에서 결론을 맺는다.

2. 동적 귀환 신경회로망에 의한 강화 학습

본 논문에서는 동적 귀환 신경회로망과 뉴럴 퍼지를 자율이동 로봇의 행동학습에 사용한다. 외부의 환경으로부터 받은 연속적인 입력과 출력을 행동-평가 구조를 가진 자율이동 로봇의 행동 계획 알고리즘으로 사용한다. 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[1]. 비선형 동적 시스템을 제어하는데 정적 귀환 신경회로망 보다 내부적으로 상태 피드백이 있어 동 특성을 가지는 동적 귀환 신경회로망이 적합하다. 본 논문에서는 그림 1과 같이 퍼지 추론과 동적 귀환 신경회로망을 자율이동 로봇의 행동 학습에 사용한다.

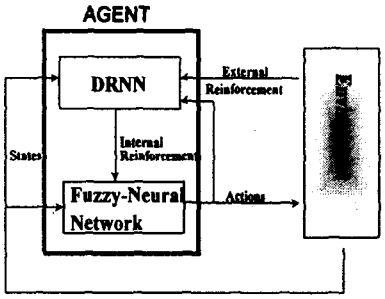

그림 1. 강화 학습의 블록선도

그림 2에서 보인 동적 귀환 신경회로망의 i 번째 뉴런의 동적 방정식은 다음 식과 같다.

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

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

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

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

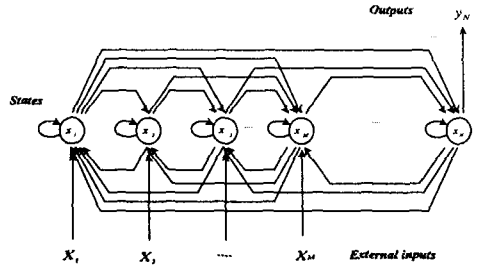

그림 2. 동적 귀환 신경회로망 구조

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

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

단, $\hat{r}$은 내부강화 신호로서 연속적인 값을 가지며, 환경으로 받는 r은 외부강화 신호, $y[t,t]$는 t시간에서의 외부입력과 t 시간에서의 연결강도에 의해 계산된 동적 귀환 신경회로망의 출력 값이고, γ는 할인률로서 임의의 양수 값을 갖는다. 뉴럴 퍼지는 $\hat{r}$의 값에 의해 학습이 이루어진다. 동적 귀환 신경회로망의 웨이트를 이진수로 랜덤하게 발생한 수를 이와 같은 연산 과정을 통하여, 자율 이동 로봇의 행동을 결정하는 신경회로망의 웨이트를 유전 알고리즘의 교배, 돌연변이, 선택 등의 일련의 과정을 통하여 최적화 한다. 본 논문에서는 유전연산자 중에서 2개의 개체간에 염색체를 바꿈으로써 새로운 개체를 생성하는 교배 중에서 일점 교배를, 개체 선택법으로는 엘리트 선택을 한다. 엘리트 보존 선택을 채용하면 그 시점에서의 가장 좋은 개체는 교차나 돌연변이에 의해서 파괴되지 않는 이점이 있다. 본 논문에서는 진화의 방향을 결정하는 적합도 함수를 다음 식(4)와 같이 결정한다.

$$\text{Fitness} = \alpha(P_{\max} - P)/P_{\max} + \beta(D_{mov}/D_{total}) \qquad (4)$$

단, $P_{\max}$: 최대 설정 벌칙수, P : 벌칙수, D_{mov} : 실제 로봇 이동거리, D_{total}: 목표지점 까지 총 이동거리, $0 \le \alpha, \beta, \gamma \le 1$이다.

3. 뉴럴 퍼지에 의한 행동시스템

본 논문에서는 퍼지와 뉴로의 융합 형태인 뉴럴 퍼지 모델을 사용하였다. 즉, 퍼지 뉴로 융합모델 중에서 뉴럴 퍼지를 사용한다. 뉴럴 퍼지는 동적 귀환 신경회로망으로 부터 내부강화 신호와 자율이동 로봇의 상태를 입력받아 자율이동 로봇의 행동을 결정하게 된다. 뉴럴 퍼지는 5개의 층으로 이루어지며, 각 층은 일련의 퍼지 연산과정에 대응된다. 뉴럴 퍼지의 입력부와 출력부의 소속함수는 삼각형 소속함

수를 사용한다. 각 층의 연결은 전 방향 네트워크 구조로 이루어지며. 지역적인 연산을 수행한다. 그림 3은 본 논문에서 사용한 뉴럴 퍼지 구조이다.

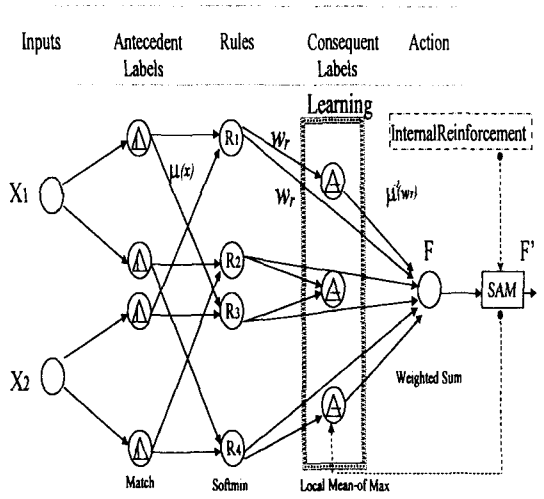

그림 3. 뉴럴 퍼지 구조

그림 4는 자율 이동 로봇(Khepera 로봇)의 장애물 센싱 거리, 센싱 반경, 이동거리를 나타낸다. 자율이동 로봇은 8개 센서를 지니게 된다.

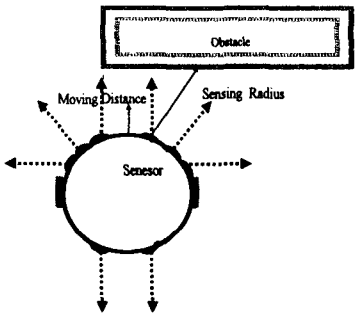

그림 4. 로봇 센서위치

뉴럴 퍼지에 대한 각 층의 연산과정은 다음과 같다[2].

layer1 :입력 층으로서 7개의 입력센서로부터 장애물과 물체까지에 대한 3개의 거리 값을 선형화하여 뉴럴 퍼지 입력으로 한다.

layer2 : $\mu_{CV,SV_L,SV_R}(x)$: 입력 x에 대한 소속함수 값으로 두 번째 층에서 이 값을 정의한다. 이때, V는 언어항을 구별하기 위하여 사용을 하고, C는 소속함수의 중심 값이 되며, SV_L은 좌변, SV_R은 소속함수의 우변에 대한 소속함수의 폭이다. 예를 들어 소속함수의 모양이 삼각형인 경우 멤버의 소속함수 값은 다음과 같은 연산 과정을 통하여 입력에 대한 소속함수 값을 결정한다.

$$\mu_{C,S_L,S_R}(x)=\begin{cases}1-|x-C|/S_R, x\in[C, C+S_R]\\ 1-|x-C|/S_L, x\in[C-S_L, C]\\ 0, \quad otherwise\end{cases} \quad (5)$$

layer 3: 각 룰의 전건부에 해당하는 값에 대한 룰의 연결강도를 계산한다. layer 3에서 계산된 값을 O_R로 표시하면 식 (6)과 같다.

$$O_{R3}=\omega_r=\frac{\sum_i \mu_i \exp(-ku_i)}{\sum_i \exp(-ku_i)} \quad (6)$$

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 식(7)에 의해 계산되어진다.

$$O_{V4}=(C_V+\frac{1}{2}(SV_R-SV_L))(\sum_r \omega_r) - \frac{1}{2}(SV_R-SV_L)(\sum_r \omega^2{}_r) \quad (7)$$

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

layer 5: 식 (8)은 뉴럴 퍼지에 의한 추론 값으로서 본 논문에서는 로봇의 회전각이 된다. 디퍼지화 과정은 무게중심법을 사용하였으며, 또한 계산적으로 연산 과정은 layer 4에서 계산된 값을 layer3에서 계산된 값으로 나눈 값과 같다.

$$F=\frac{\sum_r \omega_r \mu^{-1}(\omega_r)}{\sum_r \omega_r}=\frac{\sum_V O_{V4}}{\sum_R O_{R3}} \quad (8)$$

뉴럴 퍼지의 출력값을 바로 이용하지 않고 다음 식 (9)와 같은 연산과정을 통하여 계산되어진다.

$$F'(t)\approx(F(t), \sigma(t)) \quad (9)$$

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 뉴럴 퍼지의 실제 출력 값을 감소시켜 자율이동 로봇의 행동을 억제하고, 역이면 자율이동 로봇의 행동을 촉진 시킨다. 뉴럴 퍼지는 동적 귀환 신경회로망으로부터 내부 강화 신호를 입력받아 출력부의 삼각형 소속함수를 다음 식 (10), (11)에 의해 학습을 한다[3].

$$\Delta p = \eta S(t)\hat{r}(t)\frac{\partial v}{\partial p} = \eta\frac{\partial v}{\partial F}\times\frac{\partial F}{\partial p} \quad (10)$$

$$S(t)=\frac{F'(t)-F(t)}{\sigma(\hat{r}(t-1))} \quad (11)$$

여기에서 Δp는 퍼지 신경회로망의 후건부 소속함수이며, 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 웨이트에 대해 보상을 더 주고, 그 반대이면 웨이트의 영향을 적

게 해주기 위해 사용이 된다. v는 학습을 하려고 하는 후건부 소속함수의 중간 값과 좌우변의 폭이다. 학습식은 다음 식들과 같다.

$$\frac{\partial v}{\partial F} \approx \frac{dv}{dF} \approx \frac{v(t)-v(t-1)}{F(t)-F(t-1)} \quad (12)$$

단, $z_v(w_r) = C_v + \frac{1}{2}(S_{VR} - S_{VL})(1-w_r)$이며

$$\frac{\partial F}{\partial pv} = \frac{1}{\sum_i w_i} \sum_{V=con(R_j)} \frac{w_j \partial z_v}{\partial p_v} \quad (13)$$

$$\cdot \frac{\partial z_v}{\partial C_v} = 1 \quad (14)$$

$$\cdot \frac{\partial Z_v}{\partial S_{vR}} = \frac{1}{2}(1-w_r) \quad (15)$$

$$\cdot \frac{\partial Z_v}{\partial S_{vR}} = -\frac{1}{2}(1-w_r) \quad (16)$$

이다.

4. 실험 결과

본 논문에서 사용한 파라미터 값은 다음과 같다. • 개체군의 크기, 세대수 : 1000. 500 • 돌연변이률 : 0.02 • 교배율 : 0.8 • 적합도 계산식변수: α=0.5, β=0.5, Totaldistance=2025(미로 I), 2010(미로 II) • Max penalty=마지막 스텝수의 20% • 돌연변이 확률: 0.02 • 교차 확률: 0.8 • 학습률(η)=3이다. 퍼지 룰은 전문가에 의해 설계를 하였다. 그림 5는 미로 I 에서 세대에 대한 최대 적합도 변화, 그림 6은 미로 II에서 세대에 대한 최대 적합도 변화이다. 세대수가 지날수록 적합도가 증가함을 알 수 있다. 그림 7은 미로 I 에서의 최대 적합도를 가졌을 경우의 미로 주행, 그림 8은 미로 II에서의 미로 주행 결과를 나타낸 그림이다.

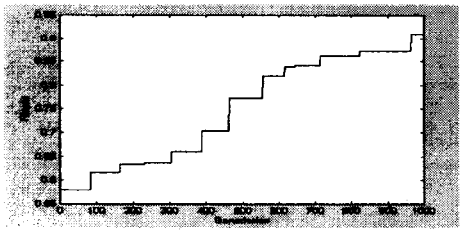

그림 5. 미로 I 에 대한 최대 적합도 변화

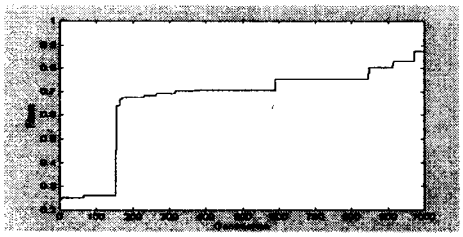

그림 6. 미로 II에 대한 최대 적합도 변화

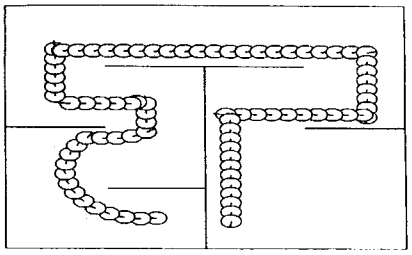

그림 7. 미로 I 에서의 로봇 주행

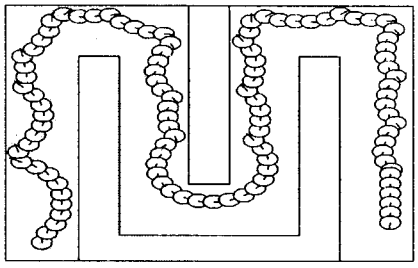

그림 8. 미로 II에서의 로봇 주행

5. 결론 및 추후 과제

자율이동로봇의 행동 진화를 위해 동적 귀환 신경회로망에 의한 뉴럴 퍼지의 강화학습알고리즘과 유전자 알고리즘을 도입한 방법을 제안하였다. 제안한 방법은 자율이동로봇의 행동제어기로서 자율이동 로봇의 미로 탐색 행동에 적용시켜 그 유효성을 검증해 보았다. 그러나 내부강화 신호를 생성하는 동적 귀환 신경회로망에서 입력센서의 개수가 커지면 유전화하는 과정에서 웨이트 변수의 개수가 기하 급수적으로 증가하기 때문에 진화의 효율이 매우 떨어진다. 문제점을 해결하기 위하여 클러스터링 기법이나 러프셋 이론을 도입해 입출력의 수를 줄이는 방법을 함께 연구할 예정이다.

감사의 글

본 연구는 과학기술부의 뇌 과학 프로젝트(Braintech 21)의 지원으로 이루어진 결과임(과제번호: 98-J04-01-01-A-07)

참고 문헌

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

[2] 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.

[3] Chin-Kuan Chiang, "A Self-Learning Fuzzy Logic Controllers Using Genetic Algorithms with Reinforcements,"*IEEE Trans. on Fuzzy Systems*, Vol. 5, No. 3, pp. 460-467, 1997. 8