퍼지 추론에 의한 리커런트 뉴럴 네트워크 강화학습

전 효 병°, 이 동 욱, 김 대 준, 심 귀 보

로보틱스 및 지능제어 시스템 연구실
중앙대학교 공과대학 전기,전자,제어공학부
Tel : 820-5319, Fax : 817-0553, E-mail : hbjun@jupiter.cie.cau.ac.kr,
kbsim@juno.cie.cau.ac.kr

Fuzzy Inference-based Reinforcement Learning for Recurrent Neural Network

Hyo-Byung Jun°, Dong-Wook Lee, Dae-Jun Kim, Kwee-Bo Sim

Robotics and Intelligent Control System Lab.
Faculty of Electrical, Electronic and Control Engineering Chung-Ang University
Tel : +82-2-820-5319, Fax : +82-2-817-0553
E-mail : hbjun@jupiter.cie.cau.ac.kr, kbsim@juno.cie.cau.ac.kr

Abstract

In this paper, we propose the Fuzzy Inference-based Reinforcement Learning Algorithm. We offer more similar learning scheme to the psychological learning of the higher animal's including human, by using Fuzzy Inference in Reinforcement Learning. The proposed method follows the way linguistic and conceptional expression have an effect on human's behavior by reasoning reinforcement based on fuzzy rule. The intervals of fuzzy membership functions are found optimally by genetic algorithms. And using Recurrent Neural Network composed of dynamic neurons as action-generation network, not only current state but also past state is considered to make an action in dynamical environment.

We show the validity of the proposed learning algorithm by applying to the inverted pendulum control problem.

Keyword : Fuzzy Inference, Reinforcement Learning, Associative Search Unit, Genetic Algorithm, Recurrent Neural Network

1. 서 론

강화학습은 실험 심리학에서 동물의 학습방법 연구에서 비롯되었으나, 최근에는 공학 특히 인공 지능분야에서 뉴럴 네트워크의 학습 알고리즘으로 많은 관심을 끌게되었다.[1]

다른 기계학습론에서는 환경에 대한 정확한 모델링을 통해 교사 신호에 의한 학습이 주가 되었으나, 강화 학습은 일반적으로 비교사 학습법으로서 동적으로 변화하는 환경하에서 제어기 또는 에이전트의 행동에 대한 보상을 최대화하는 상태-행동 규칙이나 행동 발생 전략을 찾는것이다.

그러나 많은 실세계의 경우에 있어서 목표상태에 도달할 때까지는 중간 단계의 행동에 대한 즉각적인 보상이 주어지지 않는다. 이러한 경우 외부로부터의 강화 신호가 없기 때문에 학습이 일어나지 않게 된다. 그러한 경우에도 목표상태에 도달하기 위해서는 계속적인 학습이 이루어져야 하므로 일시적인 credit 또는 blame이 주어져야한다.

이러한 문제는 credit-assignment problem이라고 하여 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. 대표적인 방법으로는 Sutton의 TD-method에 의한 Actor-critic architecture와 Watkin의 Q-learning이 있다.[1][4][5]

Actor-critic 구조에서는, 상태와 외부 강화 신호를 사용하여 critic network에서 생성한 내부 강화 신호를 action network에서 행동 학습에 사용하였다. 여기에서 critic network는 TD-method에 의해 학습을 행하는데, 출력을 다음과 같이 감쇄 기대값 평균으로 두면,

$$p(t) = E\left\{\sum_{k=0}^{\infty}\gamma^k \cdot r(t+k)\right\} = \sum_{i=1}^{n}\omega_i(t)\cdot x_i(t) \quad (1)$$

γ 는 감쇄 계수, r(t)는 시간 t에서의 보상

이때 결합 가중치 변경식은 다음과 같다.

$$\Delta\omega(t) = \eta[r(t)+\gamma\cdot p(t+1)-p(t)]\cdot x(t) \quad (2)$$

η : 학습계수

그러나 $p(t+1)$이 $\omega(t+1)$이 아니라 $\omega(t)$에 의해 계산되어야 하기 때문에 근사화에 의한 방법에 의존할 수밖에 없게된다.[1]

즉 TD-method에서는 동적으로 변화하는 환경이 Markovian 환경이라는 가정 하에 근사에 의해 연속적인 상태에만 의존하여 강화 신호를 예측하였다.

그리고 Q-learning에서는 이산화된 상태공간과 행동공간을 필요로 하기 때문에 대용량의 메모리를 필요로 하게 되고 연속적인 출력을 요구하는 문제에 적용할 수 없게된다.

따라서 본 논문에서는 그림 1과 같이 퍼지 추론과 동적 궤환 네트워크를 사용하여 인간을 포함한 고등 동물의 심리학적 학습 방법에 보다 가까운 강화 학습 알고리즘을 제안한다.

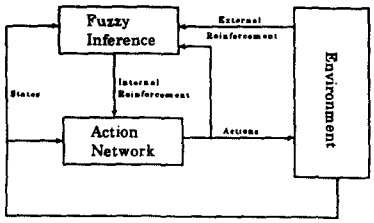

그림 1. 퍼지추론에 의한 강화학습

2장에서는 상태와 행동에 대해 평가를 하는 퍼지 추론, 3장에서는 Associative Search Unit, 시계열적인 데이터에 대해 동적인 특성을 가지는 Recurrent Neural Network의 강화 학습 알고리즘, 4장에서는 도립진자 시스템에 적용한 시뮬레이션 결과를 보이고 5장에서 결론을 맺는다.

2. 퍼지 추론에 의한 강화 신호 생성

퍼지추론은 크게 퍼지 관계의 합성 법칙에 의한 추론법, 다치 논리에 퍼지니스를 도입한 퍼지 논리에 의한 추론법, 후건부가 전건부의 선형결합으로 나타나는 추론법으로 나눌 수 있다. [10]

본 논문에서는 퍼지 관계의 합성 법칙에 의한 추론법으로서 정성적, 언어적 표현에 의해 규칙을 생성하고, max-min 합성 중심법을 사용하여 추론한다. 모델의 상태뿐만 아니라 제어기, 즉 뉴럴 네트워크의 출력을 전건부로 하고 뉴럴 네트워크의 파라메타를 조절할 수 있는 강화신호를 후건부로하여 규칙을 구성한다.

퍼지 멤버십 함수는 그림 2 에서와 같이 5개의 라벨로 구성하였다. 멤버십 함수의 최적 구간 결정에는 Genetic Algorithms을 사용하였다.

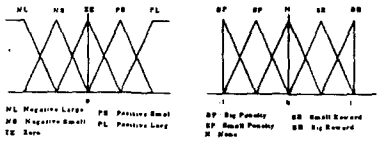

그림 2. 퍼지 멤버십 함수

퍼지 추론부의 입력이 m개, 퍼지 제어 규칙이 n개라 하면, 일반적인 퍼지 프로덕션 규칙은,

$$R^i : \text{IF } x_1 \text{ is } A_{i1}, x_2 \text{ is } A_{i2}, \cdots, x_m \text{ is } A_{im} \text{ THEN } y \text{ is } B_i \quad (3)$$

로 나타낼 수 있고, 퍼지 관계 R로 나타내면,

$$R = R_1 \cup R_2 \cup \cdots \cup R_n = \bigcup_{i=1}^{n} R_i \quad (4)$$

여기서 $R_i = (A_{i1}\times A_{i2}\times\cdots\times A_{im})\times B_i$

로된다. 지금 입력이 $A_1^0, A_2^0, \cdots, A_m^0$ 이라 하면, 전건부 변수 x_m 은 확정한 수치로 관측되므로 출력 B^0 는,

$$B^0(y) = R(x_1^0, x_2^0, \cdots, x_m^0, y) \quad (5)$$

와 같이 간단히 된다. 그러므로 추론값 y는

$$B^0(y) = \bigvee_{i=1}^{n}[\omega_i \wedge B_i(y)] \quad (6)$$

$$\omega_i = A_{i1}(x_1^0)\wedge A_{i2}(x_2^0)\wedge\cdots\wedge A_{im}(x_m^0)$$

ω_i : i 번째 규칙의 적합도

하중 평균 무게 중심법을 사용한 비퍼지화를 방법을 그림 3에 나타낸다.

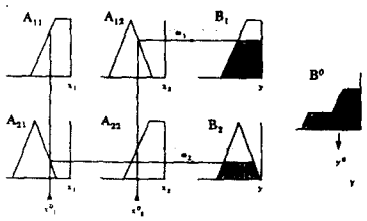

그림 3. 무게 중심 비퍼지화

3. 동적 특성을 갖는 확률 뉴런

3.1 Associative Search Unit

Associative search unit는 연상 강화 학습중의 하나인 Associative Search Rule 에 사용되는 뉴런으로서 Klopf(1982)에 의해 제안되었다.[1] 이것은 본질적으로 Hebbian Learning Rule과 같은 학습법이며, 행동의 다양성을 확보하기 위해 뉴런의 출력을 활성화 정도에 따른 확률에 의존하는 random variable로 구성한다.

$$s(t)=\sum_{i=1}^{n} w_i(t)x_i(t) \tag{7}$$

$w_i(t)$: i 번째 weight vectors

$x_i(t)$: i 번째 입력 벡터

일 때 뉴런의 출력은 다음과 같다.

$$y(t)=\begin{cases}1 & p(t)\text{ 의 확률}\\ 0 & 1-p(t)\text{ 의 확률}\end{cases} \tag{8}$$

$p(t)$는 $s(t)$ 에 대해 0 과 1 사이의 값을 가지는 증가 함수 형태의 확률이다.

이때 weight 갱신은 다음과 같다.

$$\Delta w(t)=\eta\cdot r(t)\cdot y(t-\tau)\cdot x(t-\tau) \tag{9}$$

η : 학습계수, $r(t)$: 강화 신호 , τ : 지연시간

본 논문에서는 동적인 뉴런의 출력에 가우시안 확률분포를 가지는 random variable을 더해 줌으로서 행동의 다양성을 유지하면서 뉴런의 출력 값이 연속적이 되도록 하였다.

3.2 Dynamic Recurrent Neural Network(DRNN)

동적인 리커런트 뉴럴 네트워크는 내부적으로 상태 피드백과 self-feedback이 존재하고, 입력 신호를 비선형 처리하므로 네트워크가 동적인 특성을 보이며 시계열 데이터 처리에 유용하다.

완전 연결된 리커런트 네트워크의 구조는 그림 4와 같이 뉴런이 서로 비대칭 결합하고 있는 상호 결합형 뉴럴 네트워크이다. 이 때 i 번째 뉴런의 출력은 다음과 같다.

$$y_i(t)=f(h_i(t-1))+\delta_i(\sigma)$$
$$h_i(t)=\left(\sum_j w_{ij}y_j(t)+x_i(t)\right) \tag{10}$$

여기서 $h_i(t-1)$ 는 시간 t-1에서 i 노드의 net 입력, $x_i(t)$ 시간 t에서의 i 번째 노드의 입력, f(·) 활성화 함수로서 다음과 같은 비선형함수이다.

$$f(x)=\frac{2}{1+\exp(-\frac{2x}{u_0})}-1=\tanh(\frac{x}{u_0}) \tag{11}$$

한편, $\delta(\sigma)$ 는 평균이 0이고 표준편차가 σ 인 임의의 수가 되며, 이 때 표준편차는 연속된 강화신호의 합에 따라 결정된다.

$$\sigma=\begin{cases}\dfrac{\alpha}{\sum r} & r>0\\ 1 & r=0\\ \alpha\sum|r| & r<0\end{cases} \tag{12}$$

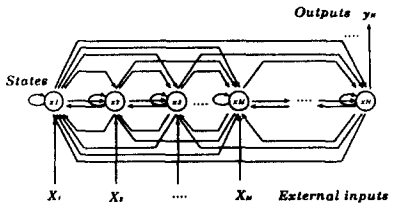

그림 4. 리커런트 뉴럴 네트워크

이 때의 출력층에서의 오차를,

$$E(t)=\frac{1}{2}\sum_k(E_k(t))^2 \tag{13}$$

$$E_k(t)=\begin{cases}(1-r(t))\cdot y_k(t), & r(t)\geq 0\\ (r(t)-1)\cdot y_k(t), & r(t)<0\end{cases} \tag{14}$$

로 두면, 최급 강하법에 의해 다음과 같이 연결강도 변화량을 구할 수 있다.

$$\Delta w_{pq}(t)=-\eta\frac{\partial E(t)}{\partial w_{pq}}=\eta\sum_k E_k(t)\frac{\partial y_k(t)}{\partial w_{pq}} \tag{15}$$

여기서, $\frac{\partial y_i(t)}{\partial w_{pq}}\equiv z^i_{pq}$라 두면 z^i_{pq} 는 다음과 같다.

$$z^i_{pq}(t)=f'(h_i(t-1))\left[\delta_{ip}y_q(t-1)+\sum_j w_{ij}z^j_{pq}(t-1)\right] \tag{16}$$

결과적으로 식 (10)과 식 (14), (15),(16)에 의해,

$$\Delta w_{pq}(t)=\eta\cdot r(t)\cdot\sum_k E_k(t)\cdot z^k_{pq} \tag{17}$$

과 같이 동적 리커런트 뉴럴 네트워크의 연결 강도를 갱신할 수 있다.

4. 시뮬레이션에 의한 검토

제안한 학습 알고리즘을 도립진자 제어 문제에 적용하여 유효성을 확인한다.

먼저 리커런트 뉴럴 네트워크의 학습에 필요한 내

부 강화 신호를 만들어 주기 위해 그림 5 와 같은 퍼지 추론 규칙을 만들었다.

멤버십 함수를 5개(NL,NS,ZE,PS,PL)로 하고 전건부의 수가 n개 라면, 만들 수 있는 규칙의 수는 5^n 개가된다. 멤버십 함수는 유전 알고리즘을 사용하여 최적의 구간을 찾아내었다.[11]

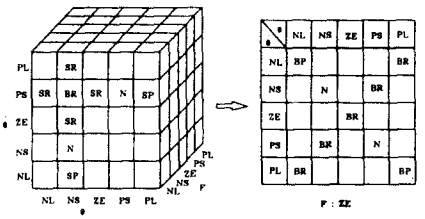

그림 5. 3차원 퍼지 추론 규칙

본 시뮬레이션에서는 도립진자의 상태 2개, θ 와 $\dot{\theta}$ 와 뉴럴 네트워크의 출력, F 를 전건부로 하여 63개의 규칙에 의해 내부 강화 신호 r를 추론하였다.

외부 강화 신호는 도립진자가 쓰러졌을 때 Failure 가 주어져 시뮬레이션이 끝나게 된다.

그림 6에서 보는 바와 같이 도립진자가 쓰러지지 않는 횟수가 학습 횟수에 따라 비례하여 증가함을 알 수 있다.

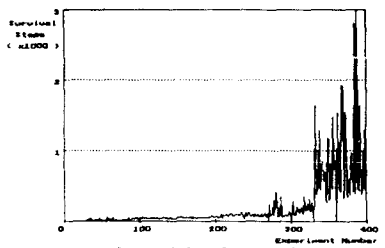

그림 6. 학습 정도 곡선

그림 7에서는 위에서 구성한 퍼지 추론 규칙만으로 도립진자의 안정화와 위치 제어를 동시에 고려한 경우의 위치 변화와 안정화 제어만 고려한 경우의 위치 변화를 비교해 보았다. 안정화와 위치 제어를 동시에 고려한 경우의 강화 신호는,

$$r(t) = \beta \cdot r_\theta(t) + (1-\beta) \cdot r_x(t) \qquad (17)$$

여기서, $r_\theta(t)$ 는 안정화에 대한 강화 신호, $r_x(t)$ 는 위치 제어에 대한 강화 신호이고, 각각은 위에서 구성한 63개의 규칙에 의해 추론된 값이다.

그림에서 ⓐ는 안정화 제어만 고려한 경우, ⓑ ,ⓒ, ⓓ 는 위치 제어까지 고려한 경우의 위치 변화의 정도를 나타낸다. 즉 학습이 진행될수록 위치 편차가 줄어듦을 알 수 있다.

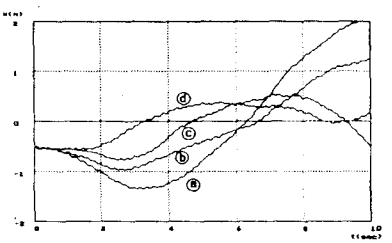

그림 7. 위치 변화 곡선

5. 결론

본 논문에서는 퍼지 추론에 의한 동적 뉴런의 강화 학습 알고리즘을 제안하였다. 제안한 알고리즘은 비선형 시스템인 도립진자 제어 문제에 적용하여 그 유효성을 확인하였다.

참고 문헌

[1] A. G. Barto, *The handbook of Brain Theory and Neural Network*, The MIT Press, pp. 804-809, 1995.

[2] E. Uchibe, M. Asada, K. Hosoda, "Behavior Coordination for a Mobile Robot Using Modular Reinforcement Learning," *Proc. IROS 96*, 1996.

[3] T. Zheng, M. Komori, O. Ishizuka, K. Tanno, H. Matsumoto, "An Adaptive ULR Fuzzy Controller Through Reinforcement Learning," *Fuzy-IEEE/IFES '95*, 1995.

[4] P. Martin, J.R. Millan, "Reinforcement Learning of Sensor-based Reaching Strategies for a Two-Link Manipulator," *Proc. IROS 96*, 1996.

[5] T. Sawaragi, H. Sawada, O. Katai, "Reinforcement Learning for Autonomous Mobile Robots by Forming Approximate Classificatory Concepts," *Proc. IROS 96*, 1996.

[6] C. J. Lin, C. T. Lin, "Reinforcement Learning for ART-Based Fuzzy Adaptive Learning Control Networks," *Fuzz-IEEE/IFES '95*, Vol. 3. pp. 1299-1306, 1995

[7] Long Zhao, Zimin Liu, "A Genetic Algorithm for Reinforcement Learning," *IEEE*, 1996

[8] A. Teller , M. Veloso, "Neural Programmin and an Internal Reinforcement Policy," *1st Asia-Pacific Con. on Simulated Evolution and Learning Proc.* 1996

[9] H.R. Berenji, A. Malkani, C. Copeland, "Tether Control Using Fuzzy Reinforcement Learning," *Fuzz-IEEE/IFES '95*, vol 3, pp.1315-1322, 1995.

[10] C. T. Lin, C. S. George Lee, *Neural Fuzzy Systems*, Prentice Hall PTR , 1996.

[11] D. E. Goldberg, *Genetic Algorithms*, Addison -Wesley, 1989.