가중 기여도를 이용한 퍼지 Q-learning

Fuzzy Q-learning using Weighted Eligibility

정석일, 이연정
경북대학교 대학원 전자공학과, 경북대학교 전자전기공학부
Tel : 053) 940-8662, E-mail : jsi@palgong.knu.ac.kr, yjlee@palgong.knu.ac.kr

Seok-Il Jeong, Yun-Jung Lee
Department of Electronics Graduate School, Kyungpook National University, Taegu
Tel : +82-53-940-8662, E-mail : jsi@palgong.knu.ac.kr, yjlee@palgong.knu.ac.kr

ABSTRACT

The eligibility is used to solve the credit-assignment problem which is one of important problems in reinforcement learning. Conventional eligibilities which are accumulating eligibility and replacing eligibility make ineffective use of rewards acquired in learning process. Because only an executed action in a visited state is learned by these eligibilities. Thus, we propose a new eligibility, called the weighted eligibility with which not only an executed action but also neighboring actions in a visited state are to be learned. The fuzzy Q-learning algorithm using proposed eligibility is applied to a cart-pole balancing problem, which shows improvement of learning speed.

Ⅰ. 서론

우리가 접하고 있는 실세계는 복잡한 모델로 표현되거나 모델로 표현될 수 없는 경우도 있다. 또, 모델이 시간에 따라 변하기도 하고, 제어의 성공 여부가 오랜 시간 뒤에 결정되기도 한다. 실세계의 이러한 특징들로 인해, 제어 대상의 모델을 이용해 제어 입력을 결정하는 제어 방법들이 실세계에 적용되기에는 한계가 있다. 따라서, 모델을 동정화(identification)하여 제어하기나, 모델을 이용하지 않고 제어하는 학습 제어기의 필요성이 대두된다. 학습 제어기란 같은 일을 수행함에 있어서 다음 수행 시에는 환경에 적응하여 제어기의 파라미터를 제어하고자 하는 방향으로 변화시키는 시스템을 말한다.

이러한 학습 제어기를 구성하는 여러 가지 방법 중에서 강화 학습법은 제어 대상의 모델을 필요로 하지 않으며, 제어 행위의 성공 여부만으로 학습이 가능한 방법으로, 에이전트와 환경으로 모델링될 수 있다. 에이전트(agent)는 제어기와 같은 역할을 하고, 환경(environment)은 제어 대상의 역할을 한다. 에이전트는 환경의 현재 상태 s_t를 인식하고, 그 상태에 적절한 행위 a_t를 출력한다. 환경은 s_t와 a_t에 의해 다음 상태 s_{t+1}으로 천이(transition)된다. 상태 천이 후에, 환경은 s_{t+1}이 제어 목적에 부합하는 상태인지 아닌지를 알려주는 보답(reward) r_t을 에이전트에게 건네준다. 에이전트는 r_t를 이용해 a_t가 현재 상태에서 수행하기에 좋은 행동인지 아닌지를 학습하게 된다. 이러한 방식으로 강화 학습은 환경에 대한 모델 없이 간단한 보답 신호만으로 현재 상태에 대한 최적의 행위를 학습한다.

강화 학습의 전통적인 문제점으로 신뢰 할당 문제(credit-assignment problem)를 들 수 있다. 일반적으로 강화 학습에서 보답은 시스템 제어의 성공 여부가 가려진 이후에야 주어지므로, 보답이 주어지기 전까지는 학습이 수행되지 않는 문제가 발생한다. 또한, 주어진 보답을 그동안 수행된 각각의 제어 행위에 어떻게 분배하느냐가 중요한 문제가 된다. 이러한 문제들을 신뢰 할당 문제라고 한다.

강화 학습에서 신뢰 할당 문제에 대한 접근법에는 크게 두 가지가 있다. 첫 번째 방법은, 보답이 주어지기 전까지 학습이 수행되지 않는 문제를 해결하기 위해 에이전트 스스로 현재 상태에서 수행한 행위를 평가하는 내부 보답을 생성하여 학습에 이용하는 방법이다. 이러한 내부 평가 함수를 도입한 대표적인 방법으로는 Barto[1]의 Actor-Critic과 Watkins[2]의 Q-learning을 들 수 있다.

두 번째 방법은, 보답이 주어질 때 그 동안 수행된 제어 행위들에 주어진 보답을 배분하는 문제를 해결하기 위해 기여도(eligibility)를 이용하는 방법이다. 기여도란 어떤 상태에서 수행한 행위가 현재의 보답에 얼만큼의 기여를 했는가를 나타내는 지표이다. 이러한 기여도에는 누적 기여도(accumulating eligibility)[1]와 대체 기여도(replacing eligibility)[3]가 있다.

이들 기존의 기여도들은 현재 상태에서 수행된 행위에 대해서만 기여도를 기록하기 때문에, 직접 방문한 상태 행위 쌍(state action pair)에 대해서만 학습이 수행된다. 따라서, 전체 상태들에 대해 학습을 수행하는데 오랜 시간이 걸린다. 이러한 단점을 보완하기 위해, 본 논문에서는 새로운 기여도를 제안한다.

기존의 기여도와는 달리, 제안하는 기여도는 현재 상태에서 수행된 행위에 대해서만 기여도가 주어지는 것이 아니라, 수행되는 행위와의 거리에 따라서 수행되지 않는 행위에도 가중치(weight)를 둔 기여도가 부여된다. 이러한 가중 기여도(weighted eligibility)를 이용하면, 직접 방문하지 않은 상태 행위 쌍에 대해서도 학습이 수행되고, 그로 인해 전체 상태를 학습하는데 걸리는 시간이 줄어들게 된다.

본 논문의 구성은 다음과 같다. 2장에서 기존의 기여도를 소개하고, 가중 기여도를 제안한다. 3장에서는 퍼지 Q-learning 알고리즘을 제시하고, 4장에서는 제안한 기여도와 기존의 기여도를 이용한 퍼지 Q-learning을 도립 진자 시스템(cart-pole system)에 적용한 결과를 비교하고, 5장에서 결론을 맺는다.

Ⅱ. 가중 기여도

서론에서 언급했듯이, 기존의 기여도에는 누적 기여도와 대체 기여도가 있다. 각각을 그림으로 나타내면 다음과 같다.

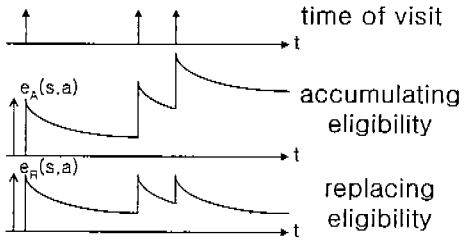

그림 1. 누적 기여도와 대체 기여도

그림 1에서 보듯이, 누적 기여도는 직접 방문한 상태 행위 쌍의 현재의 기여도가 이전의 기여도에 누적되는 방식으로, 수식으로 표현하면 다음과 같다.

$$e_A(s,a)=\begin{cases} e_A(s,a)+\dfrac{a_i(s)}{\sum_{i=1}^{N} a_i(s)}, & \text{if } s=s_t,\ a=a_t^i \\ \lambda\gamma e_A(s,a), & \text{otherwise} \end{cases} \quad (1)$$

반면에, 대체 기여도는 현재의 기여도가 이전의 기여도를 대체하는 방식으로, 다음과 같은 식으로 표현된다.

$$e_R(s,a)=\begin{cases} \dfrac{a_i(s)}{\sum_{i=1}^{N} a_i(s)}, & \text{if } s=s_t,\ a=a_t^i \\ \lambda\gamma e_R(s,a), & \text{otherwise} \end{cases} \quad (2)$$

식 (1), (2)에서 볼 수 있듯이, 기존의 기여도들은 직접 방문한 상태 행위 쌍에 기여도를 기록하고 그 값을 시간에 따라 감쇠시키면서, 학습에 사용한다. 이러한 기여도들을 갱신도(backup diagram)로 표현하면 다음과 같다.

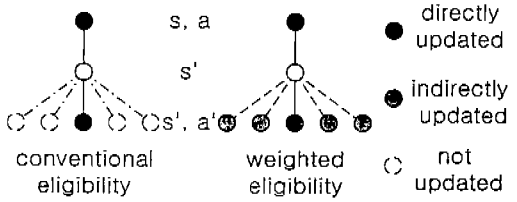

그림 2. 기여도에 따른 갱신도

그림 2에서 보듯이, 기존의 기여도들은 현재 수행된 상태 행위 쌍의 기여도만을 갱신하기 때문에 환경으로부터 얻은 보답을 충분히 이용

하지 못한다고 할 수 있다.

즉, 기존의 기여도들은 현재 방문한 상태 행위 쌍이나 이전에 방문했던 상태 행위 쌍에 대해서만 기여도를 부여하기 때문에, 직접적으로 방문한 상태 행위 쌍에 대해서만 학습이 이루어진다. 이러한 학습 방식으로는 상태의 개수가 많은 경우에 모든 상태에 대해서 학습을 하기 위해서 상태 전부를 직접적으로 방문을 해야 하기 때문에, 전체 상태에 대한 학습이 거의 불가능하거나 가능하다 할 지라도 오랜 시간이 걸리게 된다.

이러한 문제를 개선하기 위해, 본 논문에서는 가중 기여도를 제안한다. 그림 2에서 보듯이, 제안된 방법은 직접 수행된 상태 행위 쌍뿐만이 아니라, 수행되지 않은 상태 행위 쌍들에 대해서도 기여도를 부여하여, 간접적으로도 기여도가 갱신되게 한다. 즉, 기존의 대체 기여도에 현재 상태에서 수행된 행위와의 거리에 따른 가중치를 곱한 값을 수행되지 않은 상태 행위 쌍에 부여함으로써, 주어진 보답에 대해서 더 많은 상태 행위 쌍이 학습에 참여하게 되고, 그로 인해 학습 수행 속도가 향상된다. 제안한 기여도를 식으로 표현하면 다음과 같다.

$$e_W(s,a)=\begin{cases}\dfrac{a_i(s)}{\sum_{i=1}^{N}a_i(s)}\cdot w(d), & \text{if } s=s_t,\ \forall a\\ \lambda\gamma e_W(s,a), & \text{otherwise}\end{cases} \quad (3)$$

여기서,

$$w(d)=1-\frac{d}{\zeta} \quad (4)$$

$$d=|a-a_t^i| \quad (5)$$

$$a_t^i=\pi(Q_i(s_t,A_i),A_i) \quad (6)$$

이다. 식 (4), (5)에서 d는 수행된 행위 a_t^i와 수행되지 않은 행위 a 사이의 거리를 말하고, ζ는 거리에 따른 가중치를 결정하는 파라미터이다. 식 (6)에서 A_i는 현재 상태에서 취할 수 있는 행위들의 집합을 말하고, $Q_i(s_t,A_i)$는 현재 상태 s_t에서 A_i에 속한 행위들을 수행할 때 그 이후에 받을 수 있는 보답들의 합을 나타내는 값이다. 식 (5), (6)에서 a_t^i는 현재 상태에서 ε-greedy 방법에 따라 결정된 행위를 말한다. ε-greedy 방법이란, 현재 수행할 행위를 선택할 때 $1-\varepsilon$의 확률로는 현재 상태에 대한 최적의 행위를 선택하고, 나머지 ε의 확률로는 최적의 행위가 아닌 임의의 행위를 선택하는 방법을 말한다.

Ⅲ. 퍼지 Q-learning

퍼지 Q-learning[4-7]은 퍼지 이론(fuzzy theory)을 Q-learning에 접목하여, 연속적인 상태와 행위 공간에 대해 학습이 가능하도록 한 알고리즘이다. 이는 연속적인 상태 변수값들에 대해 퍼지화(fuzzification)를 수행하여 퍼지화된 상태를 얻고, 그 상태를 이용해서 추론한 결과를 비퍼지화기(defuzzifier)를 통해 연속적인 값으로 출력하는 퍼지 논리 제어기의 장점을 Q-learning에 도입한 것이다.

2장에서 제안한 기여도를 적용하기 위해서 Q-learning을 대신해, 기여도를 이용하는 Watkins의 Q(λ) 방법을 이용하였다. Q(λ) 방법에 퍼지 이론을 접목한 알고리즘에는 여러 가지[5-7]가 있는데, 그들을 종합하여 정리한 알고리즘은 다음과 같다.

단계 1. 모든 상태, 행위에 대해서 $Q(s,a)$와 $e(s,a)$를 0으로 초기화한다.

단계 2. 현재 상태 s_t에 대해서 수행할 행위 a_t를 구한다.

$$a_t=\frac{\sum_{i=1}^{N}a_i(s_t)\cdot a_i}{\sum_{i=1}^{N}a_i(s_t)} \quad (7)$$

여기서,

$$a_i(s_t)=\prod_{j=1}^{P}\mu_{ij}(s_t^j) \quad (8)$$

$$a_t^i=\pi(Q_i(s_t,A_i),A_i) \quad (9)$$

이다.

단계 3. $Q(s_t,a_t)$와 $e(s_t,a_t)$를 구한다.

$$Q(s_t,a_t)=\frac{\sum_{i=1}^{N}a_i(s_t)\cdot Q_i(s_t,a_t^i)}{\sum_{i=1}^{N}a_i(s_t)} \quad (10)$$

단계 4. a_t를 환경에 가하고, 다음 상태 s_{t+1}과 r_t을 얻는다.

단계 5. s_{t+1}에 대한 최적의 행위 a_{t+1}^*와 $Q(s_{t+1},a_{t+1}^*)$를 구한다.

$$a_{t+1}^*=\frac{\sum_{i=1}^{N}a_i(s_{t+1})\cdot a_{t+1}^{i*}}{\sum_{i=1}^{N}a_i(s_{t+1})} \quad (11)$$

$$Q(s_{t+1}, a^*_{t+1}) = \frac{\sum_{i=1}^{N} a_i(s_{t+1}) \cdot Q_i(s_{t+1}, a^{i*}_{t+1})}{\sum_{i=1}^{N} a_i(s_{t+1})} \quad (12)$$

여기서,

$$a_i(s_{t+1}) = \prod_{j=1}^{P} \mu_{ij}(s_{t+1}) \quad (13)$$

$$a^{i*}_{t+1} \leftarrow \arg\max_b Q_i(s_{t+1}, b) \quad (14)$$

이다

단계 6. 모든 상태, 행위에 대해서 $Q(s,a)$를 갱신한다.

$$Q(s,a) \leftarrow Q(s,a) + \beta\, \delta\, e(s,a) \quad (15)$$

여기서,

$$\delta \leftarrow r + \gamma Q(s', a^*) - Q(s,a) \quad (16)$$

이다.

단계 7. 만약 $a_t \neq a^*_t$이면 모든 상태, 행위에 대한 $e(s,a)$를 0으로 초기화한다.

단계 8. 현재 시도가 실패이면, s_t를 s_{t+1}으로 갱신하고 단계 2로 간다.

위 알고리즘의 단계 3에서 사용되는 기여도 $e(s_t, a_t)$에는 기존의 누적 기여도나 대체 기여도, 또는 제안된 가중 기여도가 사용된다. 단계 7은 현재 상태에서 최적의 행위를 취할 때에만 기여도를 부여해 Q-함수값을 갱신하기 위해 사용된다.

Ⅳ. 모의 실험

제안된 가중 기여도와 기존의 기여도를 이용한 퍼지 Q-learning을 도립 진자 시스템에 적용해 각각의 결과로부터 두 기여도의 학습 속도에 미치는 영향을 비교한다.

도립 진자 시스템의 동특성은 4개의 상태 변수 $\theta, \dot{\theta}, x, \dot{x}$을 가지며 다음 식들로 표현된다.

$$\ddot{\theta} = \frac{g\sin\theta + \cos\theta\left[\dfrac{-f - m_p l \dot{\theta}^2 \sin\theta + \mu_c sgn(\dot{x})}{m_c + m_p}\right] - \dfrac{\mu_p \dot{\theta}}{m_p l}}{l\left[\dfrac{4}{3} - \dfrac{m_p \cos^2\theta}{m_c + m_p}\right]}$$

$$\ddot{x} = \frac{f + m_p l[\,\dot{\theta}^2 \sin\theta - \ddot{\theta}\cos\theta] - \mu_c sgn(\dot{x})}{m_c + m_p}$$

여기서 g는 중력가속도, m_c는 수레의 질량, m_p는 진자의 질량, l은 진자의 길이의 반, μ_c는 수레와 지면과의 마찰 계수, μ_p는 수레와 진자와의 마찰 계수를 나타낸다. 모의 실험에 사용된 값은 각각 $9.8m/s^2$, $1.0kg$, $0.1kg$, $0.5m$, 0.0005, 0.000002 이다. 각각의 상태 변수에 대한 제약 조건은 $-12^\circ \le \theta \le 12^\circ$, $-2.4m \le x \le 2.4m$, $-10N \le f \le 10N$ 이다.

제어기가 진자로부터 획득하는 보답은 다음과 같다.

$$r_t = \begin{cases} -1, & \text{if } |\theta| > 12^\circ \text{ or } |x| > 2.4\text{m} \\ 0, & \text{otherwise} \end{cases}$$

퍼지 Q-learning에 사용된 파라미터들은 다음과 같다. 학습율(learning rate) β는 0.3, 감가율(discount rate) γ는 0.99, 기여도 감쇠율(decay rate for eligibility) λ는 0.3, 가중율(weight rate) ζ는 20.0, 탐색율(exploration rate) ε은 0.001을 사용하였다.

도립 진자의 상태 변수의 초기값은 매 시도마다 $\theta = 1.5$, $\dot{\theta} = 0, x = 0, \dot{x} = 0$로 초기화된다. 퍼지 논리 제어기의 각각의 상태 변수에 대한 소속도 함수의 모양은 다음과 같다.

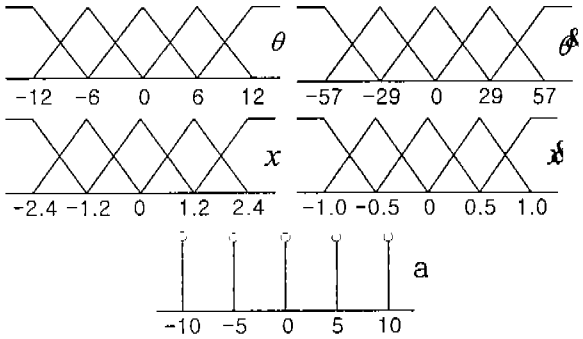

그림 3. 상태 변수들과 행위의 소속도 함수

기존의 기여도와 가중 대체 기여도를 이용한 도립 진자 시스템의 학습 결과는 그림과 같다.

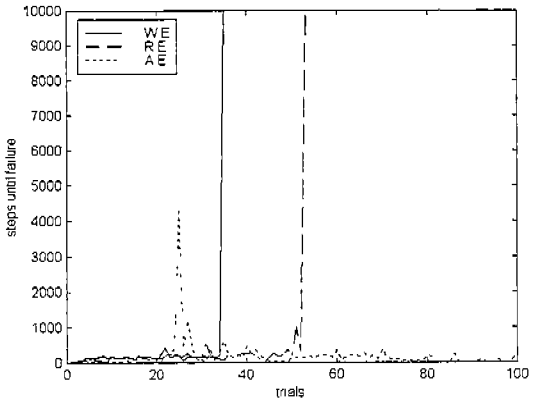

그림 4. 기여도에 따른 학습 속도

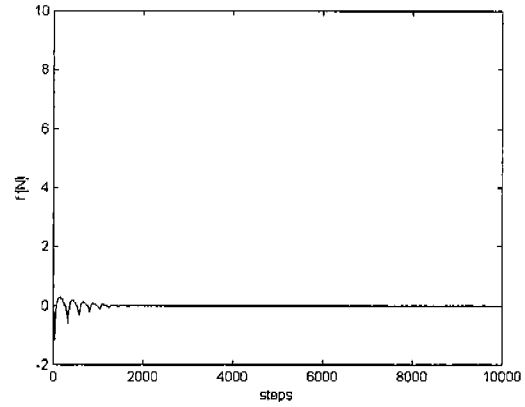

그림 5. 학습된 $f(t)$의 궤적

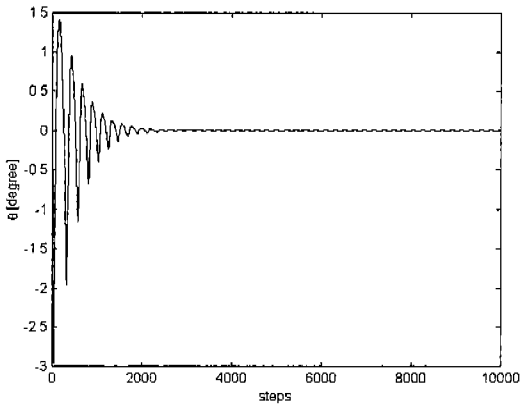

그림 6. 학습된 $f(t)$에 의한 $\theta(t)$의 궤적

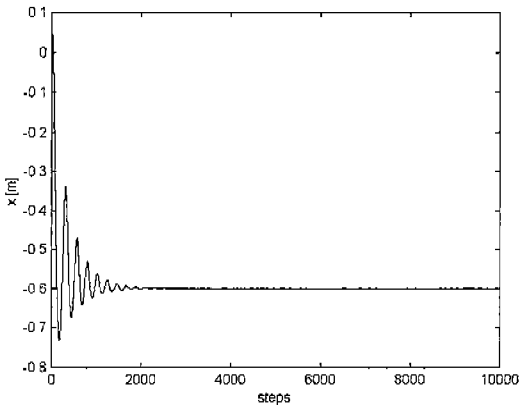

그림 7. 학습된 $f(t)$에 의한 $x(t)$의 궤적

그림 4에서 실선은 가중 기여도, 쇄선은 대체 기여도, 점선은 누적 기여도를 이용한 결과를 나타낸다. 가중 기여도를 이용한 경우가 기존의 기여도를 이용한 경우에 비해 학습 속도가 향상되었음을 알 수 있다. 이는 가중 대체 기여도를 이용한 경우에 더 많은 상태 행위 쌍이 학습에 참여한 결과로 볼 수 있다.

그림 5, 6, 7은 본 논문에서 제안된 가중 기여도를 이용한 퍼지 Q-learning 알고리즘을 통해 학습을 완료한 후, 학습된 Q-함수에 대해서 현재의 상태에서 항상 가장 큰 Q-함수값을 가지는 행위만을 선택하여 도립 진자 시스템을 제어한 결과이다.

V. 결론

본 논문에서는 직접적인 방문뿐만 아니라 간접적인 방문에 의한 학습을 수행하는 가중 기여도를 제안하였다. 또, 제안한 기여도를 이용하기 위해, 기존의 여러 퍼지 Q(λ) 알고리즘을 종합해서 정리하였다. 정리한 알고리즘에 제안한 기여도를 이용하여 도립 진자 시스템에 대해 학습시킨 결과와 기존의 기여도를 이용한 결과를 비교함으로써, 학습 속도의 향상을 보였다.

VI. 참고 문헌

[1] A. G. Barto, R. S. Sutton, and C. W. Anderson, "Neuronlike adaptive elements that can solve difficult learning control problems," *IEEE Trans. on Sys., Man, and Cyber.*, vol. 13, no. 5, 1983.

[2] C. J. C. H. Watkins and P. Dayan, "Technical note: Q-learning," *Machine Learning*, vol. 8, 1992.

[3] S. P. Singh and R. S. Sutton, "Reinforcement learning with replacing eligibility traces," *Machine Learning*, vol. 22, 1996.

[4] P. Y. Glorennec and L. Jouffe, "Fuzzy Q-learning," *IEEE Conf. on Fuzzy Systems*, vol. 2, pp. 659-662, 1997.

[5] T. Horiuch, A. Fujino, O. Katai, and T. Sawaragi, "Fuzzy interpolation-based Q-learning with continuous states and actions," *IEEE Conf. on Fuzzy Systems*, vol. 1, 1996.

[6] L. Jouffe, "Fuzzy inference system learning by reinforcement methods," *IEEE Trans. on Sys., Man, and Cyber.*, vol. 28, no. 3, pp. 338-354, Aug., 1998.

[7] M. Kim, S. Hong, and J. Lee, "Self-organizing fuzzy inference system by Q-learning," *IEEE Conf. on fuzzy Systems*, pp. 372-377, Aug., 1999.