퍼지 및 지능시스템학회 논문지 2001, Vol. 11, No. 5 pp. 388-394

분포 기여도를 이용한 퍼지 Q-learning

Fuzzy Q-learning using Distributed Eligibility

정석일, 이연정*
Seok-Il Jeong, Yun-Jung Lee

Jatco Korea Engineering
*경북대학교 전자전기공학부

요 약

강화학습은 에이전트가 환경과의 상호작용을 통해 획득한 경험으로부터 제어 규칙을 학습하는 방법이다. 강화학습의 중요한 문제 중의 하나인 신뢰 할당 문제를 해결하기 위해 기여도가 사용되는데, 누적 기여도나 대체 기여도와 같은 기존의 기여도를 이용한 방법은 방문한 상태에서 수행된 행위만을 학습시키기 때문에 학습 과정에서 획득된 보답 신호를 효과적으로 사용하지 못한다. 본 논문에서는 방문한 상태에서 수행된 행위뿐만 아니라 인접 행위들도 학습될 수 있도록 하는 새로운 기여도로써 분포 기여도를 제안한다. 제안된 기여도를 이용한 퍼지 Q-learning 알고리즘을 역진자 시스템에 적용하여 학습 속도면에서 기존의 방법에 비해 우수함을 보인다.

Abstract

Reinforcement learning is a kind of unsupervised learning methods that an agent learns control rules from experiences acquired by interactions with environment. The eligibility is used to resolve the credit-assignment problem which is one of important problems in reinforcement learning. Conventional eligibilities such as the accumulating eligibility and the replacing eligibility are ineffective in use of rewards acquired in learning process, since only one executed action for a visited state is learned. In this paper, we propose a new eligibility, called the distributed 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 the proposed eligibility is applied to a cart-pole balancing problem, which shows the superiority of the proposed method to conventional methods in terms of learning speed.

Key Words : Fuzzy Q-learning, reinforcement learning, distributed eligibility

1. 서 론

실제 제어대상 시스템을 완벽히 모델링하기란 매우 어려운 일이다. 실세계(real world)의 시스템들은 연속적인 값을 가지는 상태들의 복잡한 모델로 표현되거나, 시간에 따라 모델이 변화하거나, 모델링하기 어려운 외란 등의 불확실성을 포함하고 있다. 이와 같은 이유에서, 대상 시스템의 모델을 기반으로 제어기를 설계하고 원하는 제어 입력을 결정하는 방법들이 실세계에 적용되기에는 한계가 있다. 이러한 문제를 해결하는 한 가지 방법으로서, 모델을 이용하지 않고 제어기 또는 제어 입력을 직접 학습을 통하여 결정하는 이른바 학습 제어기가 활발히 연구되고 있다.

학습 제어기를 구성하는 여러 가지 방법 중에서, 강화학습(reinforcement learning)은 제어기가 현재 상태(state)에 알맞은 행위(action)를 수행하고 그 제어결과의 좋고 나쁨에 따라 주어지는 보답(reward)을 이용하여 보다 나은 결과를 얻을 수 있도록 행위 즉 제어 입력을 학습하는 방법이다. 이러한 강화학습은 학습 시에 제어 대상의 모델을 이용하지 않기 때문에, 모델을 필요로 하는 방법에 비하여 모델이 복잡하거나 모델을 얻기 힘든 문제에 보다 효과적으로 활용될 수 있다.

기본적인 강화학습의 개념에서는 일련의 제어 과정이 완료된 후 그 결과의 성공 여부에 따라 보답이 주어지게 된다. 그런데, 일련의 제어 과정 동안 여러 상태를 거치면서 여러 행위가 작용하였기 때문에 주어진 보답을 그 동안 수행된 여러 제어 행위에 직절히 분배하는 문제가 발생한다. 이를 신뢰 할당 문제(credit-assignment problem)[1][9]라고 한다.

Barto[1]는 내부평가함수(internal evaluating function)를 도입한 Actor-Critic 방법을 이용하여 제어 대상으로부터 보답이 주어지지 않더라도 내부적으로 보답을 생성하여 항상 학습을 수행할 수 있는 방법을 제안하였다. 그러나 내부평가함수를 도입하더라도, 제어 대상으로부터 얻은 보답은 그 때 수행된 행위에 대해서만 할당되고 그 이전에 수행된 행위에 분배되지 않는다. 이러한 문제를 해결하기 위해, Watkins[2]는 Q-learning 알고리즘에 기여도(eligibility) 개념을 도입한 Q(λ) learning 방법을 제안하였다. 이 방법에서는 현재 수행된 행위뿐만 아니라 보답이 주어질 때까지 수행된 모든 행위에 대해 일정한 값을 부여하는 기여도(eligibility)를 도입

접수일자 : 2001년 3월 29일
완료일자 : 2001년 8월 29일

하였다. 이러한 기여도를 이용함으로써, 제어 대상으로부터 보답을 얻을 때까지 수행된 모든 행위에 대해서 각 행위가 수행된 시점에 따라 보답이 적절하게 분배되도록 한 것이다.

한편, Actor-Critic 방법과 Q-learning 알고리즘에서는 상태 변수가 가질 수 있는 값을 유한개의 구간으로 나누고, 입력된 값이 속하는 구간에 해당하는 상태를 현재 상태로 인식한다. 따라서, 주어진 문제에 따라 적당한 구간의 분해능을 정하기가 어렵고, 상태 변수의 개수와 분해능이 증가함에 따라 전체 상태의 개수가 지수적으로 증가하는 단점을 가지게 된다. 또한 증가된 상태로 인해, 많은 기억 공간이 필요하게 되고 학습 시간도 길어지게 된다. 그리고 제어기에 의해 수행되는 행위들도 일정한 분해능으로 나뉘어져 학습 시에 선택되기 때문에, 제어기에서 출력되는 제어 입력도 불연속적이 되는 단점을 가진다.

이러한 단점들을 개선하기 위해 퍼지 이론(fuzzy theory)[3]을 Q-learning에 접목한 퍼지 Q-learning 방법이 연구되어 왔다. 퍼지 Q-learning 알고리즘에서는 상태 공간을 퍼지분할하여 연속적인 상태 할당이 이루어지도록 한다. 이와 관련한 기존의 주요 연구들을 살펴보면 Lin[4]은 퍼지 추론(fuzzy inference)을 수행하기 위해 신경회로망(neural network)을 이용하였으며, Horiuch[5]는 T-S(Takagi-Sugeno) 모델[6], Glorennec[7]은 간략화된 T-S 퍼지 추론을 이용한 방법들을 제안하였다.

본 논문에서는 기본적으로 Glorennec의 퍼지 Q-learning 방법과 Watkins의 Q(λ) learning 방법을 종합한 알고리즘의 토대 위에 분포 기여도(distributed eligibility)라는 새로운 기여도를 이용한 강화학습 방법을 제안한다. 분포 기여도의 필요성에 대하여 살펴보면 다음과 같다.

전술한 기존의 강화학습법에서는 어떤 상태에서 직접적으로 수행된 행위에 대해서만 기여도 값을 부여하는 방법을 사용된다. 따라서, 방문한 상태에서 수행된 행위에 대해서만 학습이 이루어진다. 그러나 사람이 학습하는 것을 관찰해보면, 어떠한 행위를 수행하여 얻은 결과와 그와 비슷한 행위를 수행하였을 때 얻는 결과는 비슷할 것이라는 생각을 가지고 직접 취해 보지 않은 행위에 대해서도 이미 행해본 행위와의 관계에 따라 그 결과를 추측하여 학습함을 볼 수 있다. 이러한 관찰을 기반으로 본 논문에서는 직접 수행된 행위에 대해서만 학습을 하는 기존의 강화학습법들과는 달리 직접 수행된 행위뿐만 아니라 수행되지 않은 인접 행위들의 기여도에도 직접 수행된 행위와의 거리에 따른 가중치(weight)를 곱한 기여도 값을 부여하는 분포 기여도라는 새로운 기여도를 이용한 방법을 제안한다. 이를 사용함으로써, 현재 방문한 상태에서 수행된 행위뿐만이 아니라 직접적으로 수행되지 않은 행위들에도 기여도가 주어져, 더 많은 상태 행위 쌍(state action pair)이 학습에 참여하게 되어 결과적으로 학습 수행 속도가 향상될 것으로 예상할 수 있다.

실제로 본 논문에서는 기존의 기여도와 제안한 기여도를 사용한 퍼지 Q-learning 방법을 역진자 시스템(inverted pendulum system)에 적용해 얻은 각각의 결과를 비교함으로써, 제안된 기여도를 사용한 방법이 보답을 더 효율적으로 사용해 학습 속도의 향상을 가져옴을 보일 것이다.

본 논문은 다음과 같이 구성된다. 2장에서는 기존의 기여도인 누적 기여도와 대체 기여도를 소개하고, 기존의 기여도의 단점을 개선한 분포 기여도를 제안한다. 3장에서는 제안한 분포 기여도를 이용한 퍼지 Q-learning 알고리즘을 제시하고, 4장에서 기존의 기여도와 분포 기여도를 이용한 퍼지 Q-learning 방법을 역진자 시스템에 적용한 모의 실험 결과들을 비교한다. 끝으로 5장에서 결론을 맺는다.

2. 분포 기여도

2.1 누적 기여도와 대체 기여도

강화학습의 가장 중요한 문제 중의 하나인 신뢰 할당 문제의 접근법에는 내부평가함수의 구현과 기여도의 도입이 있다. 환경으로부터 보답이 주어지지 않을 때에도 내부평가함수 값의 변화량으로 학습을 수행할 수 있으며, 기여도를 도입하여 보답이 주어질 때 그 동안 수행된 모든 행위들에 대하여 보답을 적절히 분배할 수 있다. 기존에 사용되어 온 기여도에는 누적 기여도(accumulating eligibility)와 대체 기여도(replacing eligibility)가 있으며 이들의 개념을 그림으로 표현하면 다음과 같다[8][9].

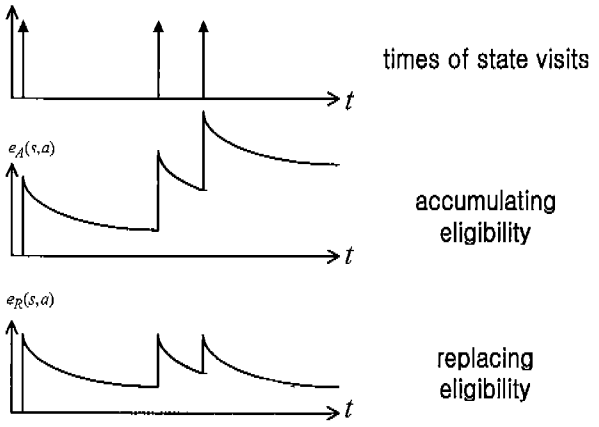

그림 1. 누적 기여도와 대체 기여도
Fig. 1. Accumulating and replacing eligibilities.

그림 1에서 보듯이, 누적 기여도는 어떤 상태를 방문해서 어떤 행위를 수행하게 되면 일정한 값이 현재의 기여도에 더해지고, 그 이후에 시간에 따라 점차 감쇠된다. 즉, 직접 방문한 상태 행위 쌍의 현재 기여도가 이전의 기여도에 누적되는 방식으로, Q(λ) learning 알고리즘과 퍼지 Q-learning 방법에 사용되었다. 이를 수식으로 표현하면 다음과 같다.

$$\dot{e}_A(s,a) = \begin{cases} e_A(s,a) + \triangle e(s,a), & \text{if } s = s_t,\ a = a_t \text{ or } a = a_t^i \\ \lambda\gamma e_A(s,a) & \text{, otherwise} \end{cases} \tag{1}$$

여기서, γ는 보답 감쇠율(discount rate)이며 $0 \le \gamma \le 1$이고, λ는 기여도 감쇠율(eligibility rate)이며 $0 \le \lambda \le 1$ 이다. γ는 추후 식 (13)에서 사용되는 값으로 다음 상태의 최적 행위에 대한 Q-함수값을 얼마나 인정할 것인가를 의미하며, λ는 시간에 따라 기여도를 어떤 비율로 감쇠시킬 것인가를 뜻한다. s_t, a_t, a_t^i는 각각 시간 t 에서의 상태, 행위, 그리고 퍼지 Q-learning의 경우 i번째 상태에서의 행위를 나타낸다. 또한, 더해지는 값 $\triangle e(s,a)$ 는 불연속적인 상태에 대한 기여도인지 연속적인 상태에 대한 기여도인지에 따라 결정된다. 즉, Q-learning 방법은 불연속적인 상태에 대해서

학습하므로, 한 순간에 하나의 상태에 대해서 하나의 행위만이 선택되어 수행된다. 따라서, 한 번에 갱신되는 기여도는 하나이며, 이 때의 기여도 변화량 $\triangle e(s,a)$ 는 1이 된다. 반면에 연속적인 상태에 대한 학습법인 퍼지 Q-learning 알고리즘에서는 매 순간마다 퍼지 규칙의 개수만큼의 행위가 선택되기 때문에 그 개수만큼의 기여도가 갱신되고, 기여도 변화량은 각 규칙의 적합도가 각 규칙의 적합도의 합에서 차지하는 비율로 결정된다. 이상의 기여도 변화량을 식으로 정리하면 다음과 같다.

$$\triangle e(s,a) = \begin{cases} 1 \quad , & \text{if Q-learning} \\ \dfrac{a_i(s_t)}{\sum_{i=1}^{N} a_i(s_t)} , & \text{if fuzzy Q-learning} \end{cases} \tag{2}$$

이러한 누적 기여도와는 달리, 대체 기여도는 현재의 기여도 값이 이전의 기여도 값을 대체하는 방식으로, 다음과 같은 식으로 표현된다.

$$e_R(s,a) = \begin{cases} \triangle e(s,a) \quad , \text{ if } s=s_t,\ a=a_t \text{ or } a=a_t^i \\ \lambda\gamma e_R(s,a) \ , \text{ otherwise} \end{cases} \tag{3}$$

여기서, 대체되는 기여도 값은 누적 기여도의 경우와 같이 식 (2)의 값을 따른다.

식 (1), (3)에서 볼 수 있듯이, 기존의 기여도들은 직접 방문한 상태에서 수행한 행위의 기여도 값을 누적하거나 대체하는 방식으로 현재 선택된 행위의 기여도를 기록하고, 그 값을 시간에 따라 감쇠시키면서 학습에 사용한다.

2.2 분포 기여도

기존의 기여도와 제안한 분포 기여도를 도식적으로 비교하여 보기 위해, 각 기여도를 이용한 강화학습법에 의해서 갱신되는 상태 행위 쌍을 그림으로 나타내면 다음과 같다.

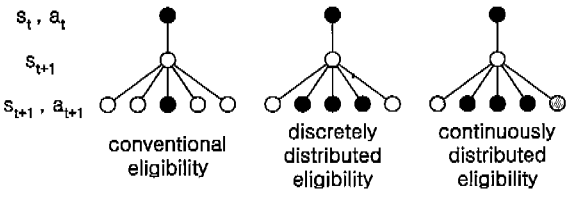

그림 2. 기존의 기여도와 제안한 분포 기여도에 대한 갱신도
Fig. 2. Backup diagrams for the conventional eligibility and the proposed distributed eligibility.

그림 2는 s_t에서 a_t를 수행하여 다음 상태 s_{t+1}으로 상태 천이가 일어나고, s_{t+1}에서 수행될 수 있는 다섯 가지의 행위 중에서 왼쪽에서 세번째 행위를 선택한 경우를 나타낸다. a_{t+1}이 수행된 이후에, 흰색이 아닌 색깔로 표시된 상태 행위 쌍들이 갱신된다. 그림 2에서 보듯이, 기존의 기여도들은 직접 수행한 행위에 대한 기여도만이 갱신되고 나머지 행위들에 대한 기여도는 전혀 갱신되지 않기 때문에, 모든 상태에 대해서 학습을 하기 위해서는 상태 전부를 직접적으로 방문해야 한다. 따라서, 상태의 개수가 많은 경우에는 학습이 거의 불가능하거나, 가능하다 할 지라도 학습의 속도가 느려지게 된다. 또, 상태의 개수가 적은 경우에도 오랫동안 제어 행위를 수행한 후에 하나의 보답이 주어지는 강화학습 문제의 특징을 고려해 볼 때 학습의 효율성에 문제가 있다.

우리는 어떠한 행위를 수행하여 그 결과로 어떠한 보답을 얻었을 때, 그와 비슷한 행위를 수행하면 비슷한 보답을 얻을 것이라고 생각을 가지기 때문에, 사람이 학습하는 과정을 지켜보면 직접 취해 보지 않은 행위에 대해서도 이미 행해본 행위와의 관계에 따라 그 결과를 추측하여 학습함을 관찰할 수 있다. 이러한 관찰을 기반으로 본 논문에서는 직접 수행된 행위에 대해서만 학습을 하는 기존의 기여도들의 단점을 개선하기 위해 분포 기여도라는 새로운 기여도를 제안한다.

그림 2에서는 이산분포 기여도(discretely-distributed eligibility)와 연속분포 기여도(continuously-distributed eligibility) 두가지 분포 기여도의 기여도 갱신 과정을 나타낸다. 이들 분포 기여도는 기존의 기여도와 달리, 직접 수행된 행위뿐만 아니라 수행되지 않은 인접 행위들의 기여도에도 직접 수행된 행위와의 거리에 따른 가중치를 곱한 값을 부여하여, 현재 방문한 상태에서 취할 수 있는 모든 행위에 대해서 기여도가 주어질 수 있도록 한 것이다. 그러므로, 더 많은 상태 행위 쌍들이 학습에 참여하게 되고, 그로 인해 학습 수행 속도가 향상된다. 기존의 기여도와 제안한 기여도를 비교하여 말하면, 기존의 기여도가 보답의 분배를 위해 학습되는 상태 행위 쌍을 시간축으로 확장하기 위해 사용되는 방법이라면, 분포 기여도는 학습되는 상태 행위 쌍을 공간적으로도 확장한 것이라고 할 수 있다.

제안한 분포 기여도를 식으로 표현하면 다음과 같다.

$$e_W(s,a) = \begin{cases} \triangle e(s,a)\cdot w(d) \ , \text{ if } s=s_t, \forall a \\ \lambda\gamma e_W(s,a) \quad , \text{ otherwise} \end{cases} \tag{4}$$

여기서, 기여도 변화량 $\triangle e(s,a)$은 식 (2)와 같으며, $w(d)$는 기여도의 분포를 결정하는 분포함수(distribution function)이다. $w(d)$에 따라 이산분포 기여도와 연속분포 기여도로 나뉜다. 분포 함수는 인접 행위의 거리에 따른 가중치를 결정하는데, 가중치가 이산적인 값을 가지는지 연속적인 값을 가지는지에 따라 이산분포함수(discrete distribution function)와 연속분포함수(continuous distribution function)로 나뉘어진다. 그림 3은 이산분포함수를 표현한 것이다.

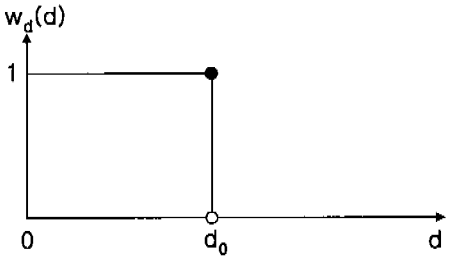

그림 3. 이산분포함수
Fig. 3. Discrete distribution function.

그림 3에서, d는 선택된 행위와 선택되지 않은 인접 행위간의 거리를 뜻한다. 이산분포함수는 d_0에 따라 그 분포 함수값으로 1 아니면 0을 가지며, 이는 다음 식으로 표현된다.

$$w_d(d) = \begin{cases} 1 \ , & \text{if } d \le d_0 \\ 0 \ , & \text{otherwise} \end{cases} \tag{5}$$

여기서, d는 다음 식과 같이 사용되는 알고리즘에 따라 다르게 표현된다.

$$d=\begin{cases}|a-a_t|, & \text{if Q-learning}\\ |a-a_t^i|, & \text{if fuzzy Q-learning}\end{cases} \tag{6}$$

이산적인 가중치를 가지는 이산분포함수에 반해, 연속분포함수는 연속적인 가중치를 가지는데, 다음과 같은 그림으로 표현된다.

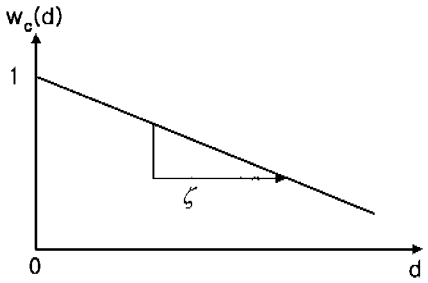

그림 4. 연속분포함수
Fig. 4. Continuous distribution function.

그림 4의 d 역시 강화학습법에 따라 식 (6)처럼 표현되고, 연속분포함수를 수식으로 표현하면 다음과 같다.

$$w_c(d)=1-\frac{d}{\zeta} \tag{7}$$

식 (7)에서 가중율(weight rate) ζ 는 행위간의 거리에 따라 어떤 가중치를 줄 것인지를 결정하는 값으로서 행위 값의 크기와 분포시키고자하는 기여도의 범위에 따라 설정해야할 설계 파라미터이다.

3. 분포 기여도를 이용한 퍼지 Q-learning

본 논문에서는 Glorennec의 퍼지 Q-learning 방법과 Watkins의 Q(λ) learning 알고리즘을 종합하여 정리한 알고리즘에 제안한 분포 기여도를 이용하여 새로운 알고리즘을 제안한다. 제안한 분포 기여도를 이용한 퍼지 Q-learning 제어기의 전체적인 구성은 다음 그림과 같다.

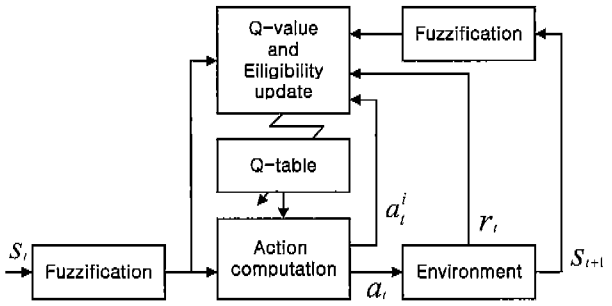

그림 5. 제안한 퍼지 Q-learning 제어기의 구성도
Fig. 5. Block diagram of the proposed fuzzy Q-learning controller.

그림 5에서, 퍼지화부(fuzzification element)는 연속적인 값을 가지는 상태입력에 대해 퍼지화된 상태와 그에 따른 소속도 함수값을 발생시킨다. Q-테이블은 Q-함수값을 저장하는 역할을 한다. 예를 들어, 오차(error)와 오차변화(change of error)를 상태변수로 가지고 각각의 상태 변수는 다섯 개의 퍼지 레이블로 표현하고 각 상태에서 수행할 수 있는 행위의 개수가 5가지라면, Q-테이블은 그림 6과 같이 구성된다.

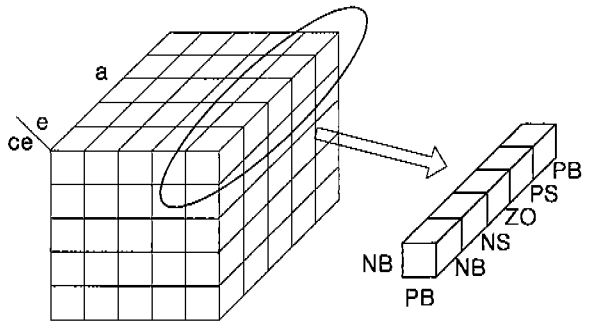

그림 6. Q-테이블의 구성
Fig. 6. Structure of the Q-table.

행위 계산부(action computation element)에서는 퍼지화부를 통해 퍼지화된 상태에 대해서 각 상태에 알맞은 행위를 Q-테이블에 저장되어 있는 Q-함수값을 이용해 결정하고, 각 상태에 따른 행위와 각 상태의 적합도를 이용해 가중합을 구하여 최종 출력을 계산한다. Q-학습자(Q-learner)는 퍼지화된 현재 상태, 퍼지화된 각 상태에 대한 행위, 퍼지화된 다음 상태 그리고 보답을 이용해 기여도와 Q-함수값을 갱신하여 Q-테이블을 변경시킨다. 행위 계산부에서 각각의 퍼지 분할된 상태에 대해 수행될 행위를 선택할 때에는 ε-그리디 방법(ε-greedy policy)[9]을 이용하여 선택하며, 이를 π 함수로 나타낸다. ε-그리디 방법에서는 (1- ε)의 확률로는 최대 Q 값을 가지는 행위를 선택하고, ε의 확률로는 랜덤하게 행위를 선택한다. 즉, ε은 탐색율(exploration rate)의 의미를 가지며 임의의 행위를 선택하는 확률이다.

제안한 알고리즘을 단계별로 정리하면 다음과 같다.

단계 1. $Q(s,a)$와 $e_W(s,a)$, s_t를 초기화한다.

단계 2. 모든 퍼지분할된 상태들에 대해서, ε-그리디 방법으로 a_t^i 를 선택한다.

(단, $a_t^i \in A$, $i=1, 2, \cdots N$ 이며 A는 행위의 집합이고 N 은 퍼지분할된 상태의 개수이다.)

$$a_t^i=\pi(Q(s,a),A) \tag{8}$$

단계 3. 최종적으로 수행될 행위 a_t와 $Q(s_t,a_t)$를 계산한다.

$$a_t=\frac{\sum_{i=1}^{N}a_i(s_t)\cdot a_t^i}{\sum_{i=1}^{N}a_i(s_t)} \tag{9}$$

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

여기서,

$$\alpha_i(s_t) = \prod_{j=1}^{P} \mu_{ij}(s_t) \tag{11}$$

이며, P 는 상태 변수의 수이다.

단계 4. a_t 를 수행하여(즉, 제어 대상 시스템에 인가하여), 다음 상태 s_{t+1} 와 보답 r_t 를 얻는다.

단계 5. $e_W(s,a)$ 를 갱신한다.

$$e_W(s,a) = \begin{cases} \dfrac{\alpha_i(s_t)}{\sum_{i=1}^{N}\alpha_i(s_t)} \cdot w_c(d), & \text{if } s = s_t \text{ and } \forall a \\ \lambda\gamma e_W(s,a), & \text{otherwise} \end{cases} \tag{12}$$

단계 6. 모든 상태와 행위에 대해서 $Q(s,a)$를 갱신한다.

$$Q(s,a) \leftarrow Q(s,a) + \beta[r_t + \gamma Q(s_{t+1}, a^*_{t+1}) - Q(s_t, a_t)]e_W(s,a) \tag{13}$$

여기서,

$$a^*_{t+1} = \frac{\sum_{i=1}^{N}\alpha_i(s_{t+1}) \cdot a^{i*}_{t+1}}{\sum_{i=1}^{N}\alpha_i(s_{t+1})} \tag{14}$$

$$Q(s_{t+1}, a^*_{t+1}) = \frac{\sum_{i=1}^{N}\alpha_i(s_{t+1}) \cdot Q(s_{t+1}, a^{i*}_{t+1})}{\sum_{i=1}^{N}\alpha_i(s_{t+1})} \tag{15}$$

$$\alpha_i(s_{t+1}) = \prod_{j=1}^{P} \mu_{ij}(s_{t+1}) \tag{16}$$

이며, 행위 a^{i*}_{t+1} 는 다음 식과 같이 각 퍼지분할된 상태들에 대해서 최대의 Q 값을 갖는 행위이다.

$$a^{i*}_{t+1} = \arg \max_a Q(s,a) \tag{17}$$

학습율 β는 Q-함수값의 갱신 정도를 결정하는 파라미터이다.

단계 7. $a_t \neq a^*_t$ 이면, $e(s,a)$ 를 초기화한다.

단계 8. s_t 를 s_{t+1}으로 갱신한다.

단계 9. 종료 조건을 만족하지 않으면, 단계 2를 수행한다.

제안한 알고리즘이 Glorennec의 퍼지 Q-learning 방법과 다른 점은 단계 5와 단계 7에서 나타난다. 단계 5에서는 기존의 기여도 대신에 제안된 분포 기여도가 사용된다. Glorennec의 방법에서 기여도는 어떠한 상태에서 어떤 행위를 수행했을 경우에 일정한 값이 더해지거나 시간에 따라 감쇠되지만, 개선한 알고리즘에서는 위의 단계 7에서 보듯이 Watkins의 알고리즘에서처럼 0으로 초기화한다. 이는 학습 과정에서 그리디 방법에 따라 랜덤하게 선정된 행위가 최적이 아닐 경우도 발생하므로 이 경우 Q 값은 유지하고 기여도는 다시 초기화하여 일련의 최적의 행위에 대해서만 기여도를 부여하기 위해서이다.

4. 모의 실험

본 절에서는 퍼지 Q-learning 알고리즘에 제안된 분포 기여도와 기존의 기여도를 사용하여 도립 진자 시스템에 적용했을 때, 각 방법을 사용하여 얻은 결과로부터 두 기여도가 학습 속도에 미치는 영향을 비교하고자 한다. 이 모의 실험에서 제어의 목적은 가능한 한 오랫동안 수레의 위치와 진자의 각도가 모두 일정 범위를 벗어나지 않도록 하는 것이다.

도립 진자 시스템의 동특성은 4개의 상태 변수 θ, $\dot{\theta}$ x, $\dot{x}$으로 표현될 수 있다. θ는 진자가 지면과 수직인 직선과 이루는 각이고, x는 선로의 기준에서 본 수레의 위치를 나타낸다. $\dot{\theta}$ 는 진자의 각속도, $\dot{x}$는 수레의 속도를 나타낸다. f는 수레를 미는 힘으로 제어기가 도립 진자 시스템에 가하는 제어 입력이다. 이러한 상태 변수와 제어 입력으로 수레와 진자의 상태를 수식화하면 다음과 같다.

$$\ddot{\theta} = \frac{g\sin\theta + \cos\theta\left[\dfrac{-f - m_p l\dot{\theta}^2\sin\theta + \mu_c sgn(\dot{x})}{m_c + m_p}\right] - \dfrac{\mu_p\dot{\theta}}{m_p l}}{l\left[\dfrac{4}{3} - \dfrac{m_p\cos^2\theta}{m_c + m_p}\right]} \tag{18}$$

$$\ddot{x} = \frac{f + m_p l[\dot{\theta}^2\sin\theta - \ddot{\theta}\cos\theta] - \mu_c sgn(\dot{x})}{m_c + m_p} \tag{19}$$

여기서, g는 중력가속도, m_c는 수레의 질량, m_p는 진자의 질량, l은 진자의 길이의 반, μ_c는 수레와 지면과의 쿨롱 마찰 계수(Coulomb friction coefficient), μ_p는 수레와 진자와의 점성 마찰 계수(viscous friction coefficient)를 나타낸다. 모의 실험에 사용된 값은 각각 $9.8m/s^2$, $1.0kg$, $0.1kg$, $0.5m$, $0.0005\ kgm/s^2$, $0.000002\ kg/s$이다.

각각의 상태 변수에 대한 제약 조건은 다음과 같다.

$$-12^\circ \le \theta \le 12^\circ,$$
$$-2.4m \le x \le 2.4m,$$
$$-10N \le f \le 10N$$

그리고, 각 상태 변수의 초기값은 $\theta = 1.5$, $\dot{\theta} = 0$, $x = 0$, $\dot{x} = 0$이다. 퍼지 Q-learning 제어기의 각 상태 변수와 행위에 대한 소속도 함수의 모양은 그림 7과 같다. 각 상태 변수는 삼각형 소속도 함수를 이용해 균등하게 분할되었으며, 행위는 5개의 싱글톤 값을 갖도록 하였다.

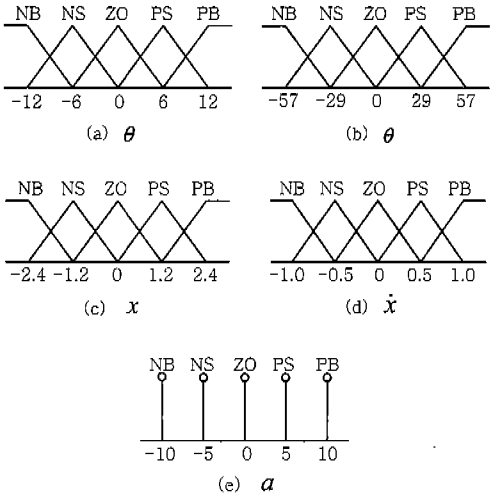

그림 7. 상태 변수들과 행위의 퍼지 레이블
Fig. 7. Fuzzy label of state variables and action.

이상의 수식과 상수값 등의 정보들은 퍼지 Q-learning 제어기는 전혀 알 수 없다. 즉, 제어기의 입장에서는 도립 진자 시스템의 상태를 나타내는 4가지 상태 변수의 값과 보답 신호의 값만을 알 수 있다. 보답 신호는 도립 진자 시스템이 주어진 각도를 벗어나 쓰러졌는지 또는 주어진 선로를 벗어났는지를 알려주는 것으로, 본 실험에서 사용되는 보답 신호는 다음과 같다.

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

퍼지 Q-learning 알고리즘의 각 파라미터들은 β=0.5, γ=0.99, λ=0.3, ζ=25.0, ϵ=0.001을 사용하였다.

학습 결과는 제어기가 행한 10번의 학습 수행(run)의 평균값으로 결정하였다. 1번의 수행은 100번의 시도(trial)로 이루어지는데, 1번의 시도는 도립 진자 시스템의 제어가 성공 또는 실패 여부가 결정되면 종료된다. 즉, 도립 진자의 제어는 진자가 주어진 각도의 범위를 벗어나거나 수레가 주어진 선로 범위를 벗어나면 실패로 간주되고, 제어기가 10000 스텝(step)이상의 제어를 실패하지 않고 수행하면 성공으로 간주된다. 이때, 1 스텝은 0.01초를 의미한다. 또, 매 시도마다 각 상태 변수의 값과 기여도가 초기화된다. 1번의 수행이 끝나면 Q값은 모두 0으로 초기화된다. 학습이 완료된 후에는 더 이상 Q값에 대한 조정이 일어나지 않도록 하여, 실제 시스템이나 시스템 모델에 적용하게된다.

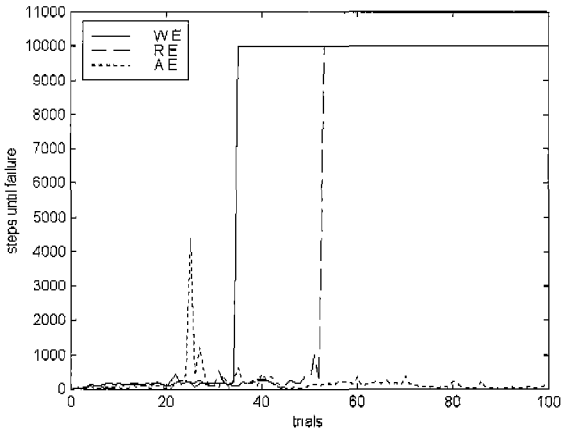

그림 8. 기여도에 따른 학습 속도

Fig. 8. Learning speed with respect to eligibility.

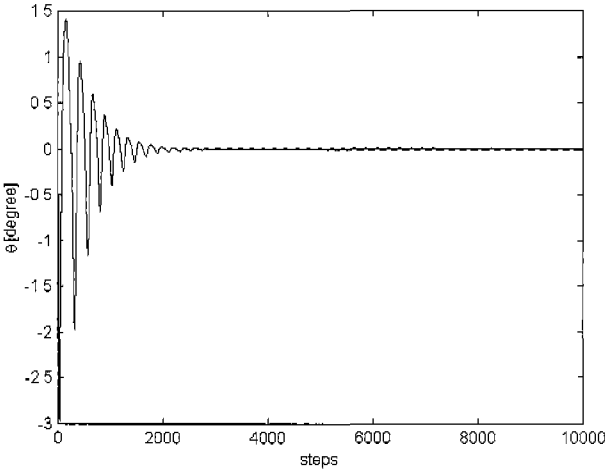

그림 9. 학습 완료 후의 $\theta(t)$ 의 궤적

Fig. 9. Trajectory of $\theta(t)$ after the learning is completed.

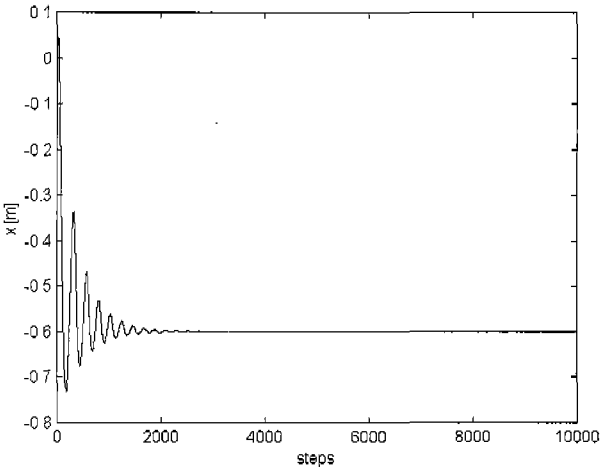

그림 10. 학습 완료 후의 $x(t)$ 의 궤적

Fig. 10. Trajectory of $x(t)$ after the learning is completed.

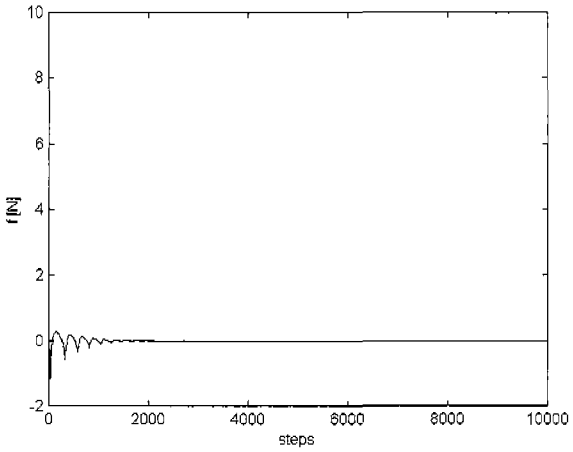

그림 11. 학습 완료 후의 $f(t)$ 의 궤적

Fig. 11. Trajectory of $f(t)$ after the learning is completed.

기존의 기여도와 분포 기여도를 이용한 도립 진자 시스템의 제어 규칙의 학습 결과는 그림 8과 같다. 실선은 연속 분포 기여도, 쇄선은 대체 기여도, 점선은 누적 기여도의 결과를 나타낸다. 누적 기여도를 이용한 경우에는 모든 시도에서 10000 스텝 이전에 도립 진자가 주어진 제한 범위를 벗어났고, 대체 기여도의 경우에는 50번째 시도에서 성공하였다. 이에 비해 분포 기여도를 이용한 경우에는 35번째 시도에서 성공하였다. 이렇듯, 연속 분포 기여도를 이용한 경우에는 기존

의 기여도를 이용한 경우에 비하여 학습 속도가 향상되었음을 알 수 있다. 이는 제안된 기여도를 이용한 경우에 더 많은 상태 행위 쌍이 학습에 참여한 결과로 볼 수 있다.

그림 9~11는 분포 기여도를 이용한 퍼지 Q-learning 알고리즘을 통해 학습을 완료한 후, 학습된 Q-함수에 대해서 그리디 방법을 통해 행위를 선택하여 역진자 시스템을 제어한 결과이다. 그림 9~11에서 보듯이 진자의 각도는 0에 수렴하며 수레의 위치는 특정 위치에서 벗어나지 않는다. 또한 제어입력인 수레를 미는 힘도 0으로 수렴함을 보인다.

5. 결 론

강화학습은 미지의 환경과의 상호작용에 의해 얻은 경험으로부터 제어 규칙을 학습하는 방법으로, 모델이 복잡하거나 모델을 얻기 힘든 실세계의 제어 문제에 보다 효과적으로 활용될 수 있다. 초기의 강화학습은 불연속적인 상태에 대해서 학습을 수행하고, 그 결과로 불연속적인 제어 입력을 출력하였다. 그러나 강화학습법 중에 하나인 Q-learning에 퍼지 이론을 접목시킨 퍼지 Q-learning 알고리즘으로 연속적인 상태공간을 기반으로 학습을 수행할 수 있게 되었고, 연속적인 제어 입력을 출력할 수 있는 방법이 제안되었다.

그러나 기존의 퍼지 Q-learning에서는 학습 과정에서 획득한 경험을 효과적으로 활용하지 못하고 있다. 즉, 현재 방문한 상태 행위 쌍이나 이전에 방문한 상태 행위 쌍에 대해서만 기여도를 부여하여 Q값을 갱신하기 때문에, 모든 상태에 대해서 학습하기 위해서는 모든 상태를 직접 방문해야 한다. 따라서, 학습 시간이 길어지는 단점을 가지고 있다.

이에 본 논문에서는 직접 방문한 상태 행위 쌍뿐만 아니라, 직접 방문한 상태에서 선택되지 않은 행위들의 기여도에도 직접 선택된 행위와의 거리에 따른 가중치를 둔 값을 부여하는 분포 기여도를 제안하였다. 이를 사용함으로써, 직접 방문한 상태 행위 쌍뿐만이 아니라 직접적으로 방문하지 않은 상태 행위 쌍들에도 기여도가 주어져 더 많은 상태 행위 쌍이 학습에 참여하도록 한 것이다.

제안한 분포 기여도와 기존의 기여도를 사용한 퍼지 Q-learning 알고리즘을 도립 진자 시스템에 적용해 얻은 결과를 비교하여 볼 때 제안한 방법이 보답을 더 효율적으로 사용함으로써 학습 속도 향상을 가져옴을 알 수 있었다.

참 고 문 헌

[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, pp. 834-846, 1983.

[2] C. J. C. H. Watkins and P. Dayan, "Technical Note : Q-learning," *Machine Learning*, vol. 8, pp. 279-292, 1992.

[3] L. A. Zadeh, "Fuzzy Sets," *Informat. Control*, vol. 8, pp. 338-353, 1965.

[4] L. J. Lin and M. Mitchell, *Reinforcement learning with hidden states from animals to animats*, The MIT Press, pp. 271-280, 1993.

[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, pp. 594-600, 1996.

[6] T. Takagi and M. Sugeno, "Fuzzy Identification of Systems and Its Applications to Modeling and Control," *IEEE Trans. on Sys., Man, and Cyber.*, vol. 15, no. 1, 1985.

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

[8] S. P. Singh and R. S. Sutton, "Reinforcement Learning With Replacing Eligibility Traces," *Machine Learning*, vol. 22, pp.123-158, 1996.

[9] R. S. Sutton and A. G. Barto, *Reinforcement Learning : An Introduction*, The MIT Press, 1998.

저 자 소 개

정석일 (Seok-Il Jeong)

1999년 : 경북대학교 전자공학과 (학사)
2001년 : 경북대학교 대학원 전자공학과 (석사)
2001년 ~ 현재 : Jatco Korea Engineering 근무

관심분야 : 퍼지제어, Q-learning, 강화학습
E-mail : seokil_jeong@jatco.co.jp

이연정 (Yun-Jung Lee)

1984년 : 한양대학교 전자공학과(공학사)
1986년 : KAIST 전기및전자공학과(공학석사)
1986년 ~ 1989년 : 한국기계연구원 연구원
1994년 : KAIST 전기및전자공학과(공학박사)
1999년 ~ 2000년 : 일본 동경공업대 객원연구원
1995년 ~ 현재 : 경북대학교 전자전기공학부 조교수

관심분야 : 퍼지제어, 학습제어, 지능로보틱스, 보행로봇
E-mail : yjlee@ee.knu.ac.kr