지능형 에이전트의 모호한 목적을 처리하기 위한 FuzzyQ-Learning*

서 호 섭O　　윤 소 정　　오 경 환
(주) 데 이 콤 정보통신연구소　　서 강 대 학 교 컴 퓨 터 학 과
garam910@choollian.net, sjyoun@ailab.sogang.ac.kr, kwoh@ccs.sogang.ac.kr

FuzzyQ-Learning to Process the Vague Goals of Intelligent Agent

Ho-Sub SeoO
Information Communication R&D Center
DACOM

So-Jeong Youn　Kyung-Whan Oh
Dept. of Computer Science and Engineering,
Sogang University

요 약

일반적으로, 지능형 에이전트는 사용자의 목적과 주위 환경으로부터 최적의 행동을 스스로 찾아낼 수 있어야 한다. 만약 에이전트의 목적이나 주위 환경이 불확실성을 포함하는 경우, 에이전트는 적절한 행동을 선택하기 어렵다. 그러나, 사용자의 목적이 인간 지식의 불확실성을 포함하는 언어값으로 표현되었을 경우, 이를 처리하려는 연구는 없었다. 본 논문에서는 모호한 사용자의 의도를 퍼지 목적으로 나타내고, 에이전트가 인지하는 불확실한 환경을 퍼지 상태로 표현하는 방법을 제안한다. 또, 퍼지 목적과 상태를 이용하여 확장한 퍼지 강화 함수와 를 이용하여, 기존 강화 학습 알고리즘 중 하나인 Q-Learning을 FuzzyQ-Learning으로 확장하고, 이에 대한 타당성을 검증한다.

1. 서론

에이전트(agent)는 인공지능 분야에서 다양한 방법으로 연구되어 온 개념으로, 사실상 인공지능 연구의 최종 목표가 사람과 유사한 지적 능력을 소유하고 있는 에이전트의 개발이라고 할 수 있다. 최근 들어 에이전트 기반 시스템(agent based system)기술은 새로운 소프트웨어 시스템의 개념화, 설계 구현을 위한 새로운 패러다임을 제공함에 따라 많은 기대를 모아오고 있으며, 특히 분산적이고 개방적인 인터넷과 같은 환경에서 많은 응용을 보이고 있다[1].

일반적으로 에이전트는 주어진 문제를 해결하기 위해서 주위의 환경으로부터 상태를 인지하고, 자신의 지식을 이용하여 최적의 해를 구하고자 한다. 그러나 에이전트는 주어진 사용자의 문제를 해결하는 도중에 실세계의 다양한 환경에서 발생할 수 있는 불확실성(uncertainty)에 직면할 수 있다. 이런 지능형 에이전트의 불확실성은 일반적으로 "외부 환경에 대한 인지"가 불확실하거나, 부족할 경우, "다중 에이전트 시스템에서 다른 에이전트의 행동이나 그로 인한 결과"를 예측하기 힘들 경우, 또, "지능형 에이전트가 가지고 있는 지식"이 자연적으로 불확실성을 내포하고 있을 경우 발생 가능하다[2][3][4].

또한, 에이전트는 사용자를 대신해서 사용자의 목적을 처리하며, 일반적인 환경에서 사용자의 목적은 언어적인 값(linguistic value)으로 주어진다고 가정할 수 있다. 그 결과 에이전트의 목적은 인간 언어 및 지식이 가지고 있는 불확실성을 포함하게 된다. 그 뿐만 아니라, 에이전트가 문제를 해결하기 위해 이동하는 환경에 대한 인지 역시 불확실성을 포함하게 된다.

따라서, 이런 불확실한 목적과 상태로 이루어지는 결정 과정에서 최적의 정책을 수행하기 위해서 이를 표현하고 해결하는 방법이 필요하다. 본 논문에서는 사용자로부터 주어지는 언어적인 값으로 표현된 지능형 에이전트의 목적을 퍼지 목적(fuzzy goal)으로 나타내고, 지능형 에이전트가 거치는 환경의 상태를 퍼지 상태(fuzzy state)로 인지하여 목적과 환경의 불확실성을 표현하는 방법을 제안한다. 또, 이를 이용하여 퍼지 강화 함수(fuzzy reinforcement function)를 정의하고, 기존 강화 학습 알고리즘인 Q-Learning을 확장한 FuzzyQ-Learning을 제안한다.

2. 퍼지 목적과 퍼지 상태

일반적으로 에이전트는 사용자로부터 언어적인 값으로 주어지는 실세계의 목적과, 환경으로부터 기인하는 불확실한 상태를 통해 최적의 정책을 찾아내어야 한다. 이를 위해 본 논문에서는 퍼지 집합을 이용하여 목적과 환경에 대한 불확실성을 표현한다.

이를 위해서 다음의 수식과 같이 퍼지 집합의 순서쌍을 이용하여 에이전트에게 주어지는 목적을 퍼지 목적으로 표현한다. 이를 위해 사용자의 목적으로

*본 연구는 '97 한국 과학 재단 특정 연구(과제번호 : 97-01-00-08-01-3) 사업지원에 의해서 이루어 졌음

부터 얻는 n개의 구문식 표현에 대한 퍼지 집합을 각각 $\tilde{G}_1,\tilde{G}_2,\ldots,\tilde{G}_n$ 이라고 하고 에이전트의 퍼지 목적 $\check{G}$ 를 다음과 같이 정의 한다.

$$\check{G} = (\tilde{G}_1,\tilde{G}_2,\ldots,\tilde{G}_n) \quad (1)$$

예를 들어 사용자가 어떤 지능형 에이전트에게 "적당한 가격의 자동차에 대한 정보를 최소의 비용을 가지고 적절한 수의 웹 사이트를 검색하여 찾아 올 것"이라는 목적을 주었다고 가정하자. 그러면, 지능형 에이전트가 가지는 퍼지 목적은 다음과 같이 나타낼 수 있다.

$$\check{G} = (MIP, VLC, MIN) \quad (2)$$

여기에서 *MIP, VLC, MIN* 는 "적당한", "최소의", "적절한"등의 구문식 표현하는 퍼지 집합으로 각각 가격(**MIddle Price**), 비용(**Very Low Cost**), 사이트의 수(**MIddle Number**)를 의미한다. 또한, 에이전트가 환경에 대한 불확실성을 표현하기 위해서 퍼지 목적과 마찬가지로 환경으로부터 얻는 n개의 항목에 대한 퍼지 집합을 각각 $\tilde{S}_1,\tilde{S}_2,\ldots,\tilde{S}_n$ 이라고 다음과 같이 퍼지 상태 $\check{S}$ 를 다음과 같이 정의한다.

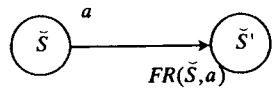
$$\check{S} = (\tilde{S}_1,\tilde{S}_2,\ldots,\tilde{S}_n) \quad (3)$$

만약 에이전트가 주어진 문제 영역(problem domain)에 대한 적절한 해석을 통해 얻을 수 있는 퍼지 집합을 가지고 있다면, 사용자의 목적과 환경으로부터 발생하는 불확실성을 퍼지 목적과 퍼지 상태로 표현할 수 있다. 이 때 에이전트가 가지고 있는 퍼지 집합들은 일종의 지식 베이스(knowledge base)의 역할을 하며, 에이전트의 인지에 따라서 얻는 환경 인자와 언어값에 의한 사용자의 목적은 적절한 퍼지화(fuzzification) 단계를 거쳐 퍼지 목적과 퍼지 상태로 표현된다.

3. 퍼지 강화 함수와 FuzzyQ-Learning

강화 학습(reinforcement learning)은 마코프 결정 과정(Markov Decision Process)과 동적 프로그래밍(dynamic programming)에 기반을 둔 학습 알고리즘으로 동적 환경(dynamic environment)에서 에이전트의 학습을 위한 교사 학습(supervised learning)의 하나로 환경과의 상호 작용을 통해 에이전트의 적절한 행동양식을 학습하는 방법이다[5]. 강화 학습은 에이전트의 행동에 의해서 얻는 보상값, 혹은, 강화값(reinforcement value)의 합이 최대가 되는 방향으로 에이전트의 행동양식을 학습시키는 방법으로, 대표적인 강화 학습 알고리즘으로 Q-Learning이 있다[5]. Q-Learning에서는 에이전트가 어떤 상태 s 에서 행동 a 를 취했을 경우 얻는 감소된 누적 강화값의 최대값을 $Q(s,a)$ 라고 정의하고 이를 최대화하는 방향으로 에이전트의 행동을 학습 시킨다. 본 논문에서는 앞서 제안한 퍼지 목적과 퍼지 상태를 이용하여 새로운 퍼지 강화 함수를 제안하고 기존의 Q-Learning을 이를 이용하여 확장한 FuzzyQ-Learning 알고리즘을 제안한다.

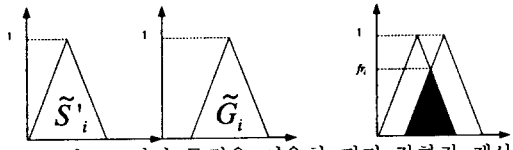
다음 그림과 같이 퍼지 목적 $\check{G}$ 을 갖는 지능형 에이전트가 퍼지 상태 $\check{S}$ 에서 어떤 행동 a 를 선택한 다음, 상태 $\check{S}'$ 으로 전이하고, 이때 퍼지 강화값 $FR(\check{S},a)$ 을 받았다고 가정하자.



<그림 1> 퍼지 상태 전이(fuzzy state transfer)

이 때, 퍼지 상태 $\check{S}$, $\check{S}'$ 와 퍼지 목적 $\check{G}$ 의 i 번째 퍼지 집합을 각각 $\tilde{S}_i$, $\tilde{S}_i'$, $\tilde{G}_i$ 라고 하고 이에 대한 i번째 퍼지 강화값 fr_i 을 다음 수식과 그림과 같이 정의한다.

$$fr_i = \max\{\mu_{\tilde{S}_i'} \wedge \mu_{\tilde{S}_i}\} = \max\{\min\{\mu_{\tilde{S}_i'}, \mu_{\tilde{S}_i}\}\} \quad (4)$$



<그림 2> 퍼지 목적을 이용한 퍼지 강화값 계산

즉, 에이전트가 행동 a 에 의해 이동한 상태 $\check{S}'$ 이 퍼지 목적 $\check{G}$ 와 유사하면 유사할수록 보다 큰 강화값을 얻게 된다. 따라서 수식 (4)를 이용하여 에이전트의 퍼지 강화 함수 $FR(\check{S},a)$ 을 다음과 같이 정의한다. r,m 은 실험에 의해서 결정되는 상수값이다.

$$FR(\check{S},a) = r \cdot \min\{fr_1^m, fr_2^m, \ldots, fr_n^m\} \quad (5)$$

이를 이용하여 새로운 *FuzzyQ* 함수를 다음과 같이 정의하고 수식 (6)을 에이전트의 각 학습 단계에서 반복적으로 적용함으로써 에이전트에게 주어진 사용자의 목적에 대한 최적의 정책에 따른 *FuzzyQ* 함수의 근사값을 구할 수 있으며, 에이전트는 퍼지 상태와 행동, 그리고 *FuzzyQ* 값으로 구성된 FuzzyQ-Table을 각 단계마다 갱신함으로써 FuzzyQ-Learning을 수행한다. γ 는 감소인자를 의미한다.

$$FuzzyQ(\tilde{S},a) \leftarrow FR(\tilde{S},a) + \gamma \max_{a'} FuzzyQ(\tilde{S}',a') \quad (6)$$

FuzzyQ-Learning의 알고리즘은 다음과 같다.

1. 각 퍼지 상태 $\tilde{S}$, 행동 a 에 대해 FuzzyQ-Table의 *FuzzyQ*값을 0으로 초기화 한다.
2. 현재의 퍼지 상태 $\check{S}$ 를 인지한다.
3. 다음의 과정을 무한히 반복한다.
   (1) 행동 a 를 선택하고 이를 수행한다.
   (2) 즉각적인 퍼지 강화값 *FR* 을 얻는다.
   (3) 행동에 따른 새로운 퍼지 상태 $\check{S}'$ 을 인지한다.
   (4) FuzzyQ-Talbe의 *FuzzyQ* 값을 다음에 의해 갱신

$$FuzzyQ(\tilde{S},a) \leftarrow FR(\tilde{S},a) + \gamma \max_{a'} FuzzyQ(\tilde{S}',a')$$

   (5) 새로운 퍼지 상태 $\check{S}'$ 으로 이동한다.

4. 구현 및 실험

본 논문에서 제안한 퍼지 강화 함수와 FuzzyQ-Learning에 대한 타당성을 검증하기 위해서, 격자 공간에서 목적지를 탐색하는 에이전트에 대한 실험을 수행하였다. 또한, 실험할 퍼지 강화 학습 에이전트는 다음의 수식에 의해서 FuzzyQ-Table의 *FuzzyQ*값을 갱신

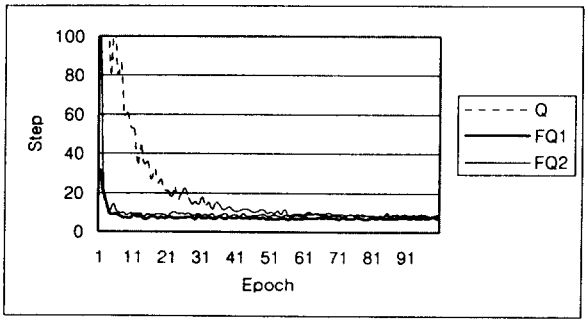
한다. 수식에서 α는 학습률, γ는 감소 인자를 의미한다.

$$FuzzyQ(\tilde{S},a) \leftarrow FR(\tilde{S},a) + \alpha\{FR(\tilde{S},a) + \gamma \max_{a'} FuzzyQ(\tilde{S}',a') - FuzzyQ(\tilde{S},a)\} \quad (7)$$

또한 에이전트의 행동 선택을 위해서 기존 강화 학습 연구에서 가장 널리 사용된 탐색전략 중 하나인 볼쯔만(Boltzmann) 탐색전략을 사용하였다.

1) 실험 I : Q-Learning과 비교

우선 첫번째 실험은 격자 모양의 가상 공간에서 기존 강화 학습 알고리즘 중 하나인 Q-Learning과 FuzzyQ-Learning의 비교 실험으로, 가상 공간은 8x8의 격자로 이루어져 있으며, 1개의 목적지가 특정 위치에 고정되어 있다. 이 공간에 목적지를 찾아 가는 3가지 종류의 에이전트를 설계하였으며, 그 중 하나는 Q-Learning을, 나머지 2개는 퍼지 목적을 이용한 FuzzyQ-Learning을 사용한다. 퍼지 강화 학습 에이전트의 경우, 첫번째 에이전트는 이동 거리를 퍼지화하여 에이전트의 상태로 인지한다. 두번째 에이전트는 보다 많은 정보를 위해서 x 축과 y 축 방향의 거리를 퍼지화 하였다. 두 에이전트 모두 그 거리를 작게 만드는 것을 퍼지 목적으로 삼는다. 3개의 에이전트에 대한 실험 결과는 다음의 그래프와 같다.



<그림 3> Q-Learning과의 비교

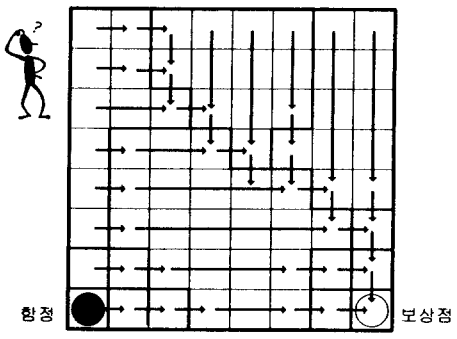
위의 그래프에 따르면 2가지 FuzzyQ-Learning이 Q-Learning보다 빠른 수렴 속도를 보여 주고 있다. FuzzyQ-Learning을 사용하는 에이전트의 경우 퍼지 집합의 수만큼의 상태를 가지고 있기 때문에, Q-Learning의 Q-Table보다 작은 크기의 FuzzyQ-Table을 가지고 있다. 따라서 Q-Learning보다 학습 수렴 속도가 빠르며, 퍼지 집합으로 표현된 모호한 목적에 대해서도 최적의 정책에 수렴함을 보여 주었다.

2) 실험 II – 퍼지 목적을 갖는 문제

첫번째 실험과 마찬가지로 8x8의 공간에 (0,7)의 위치에는 함정이 있고, (7,7)의 위치에는 보상점이 있다. 에이전트는 함정으로 다가갈수록 많은 비용을 소모해야 하며, 보상점으로 다가갈수록 많은 이익을 얻는다. 이렇게 정의된 문제에 대해서 사용자에 의해 에이전트에게 주어진 목적이 다음과 같다고 하자.

$$\tilde{G}_{BandC} = (VL_B, VL_C) \quad (8)$$

즉, 에이전트는 가상 공간을 이동하며 얻을 수 있는 이익을 매우 크게(Very Large Benefit) 하고, 비용은 매우 적게(Very Low Cost) 하는 행동양식을 학습해야 한다. 첫번째 실험과 마찬가지 방법으로 실험한 결과, 퍼지 강화 학습 에이전트는 다음의 그림과 같은 행동양식을 학습하였다.



<그림 4> 이익과 비용을 고려하는 에이전트

에이전트는 가상 공간을 14개의 퍼지 상태로 분할하였고, 하나의 퍼지 상태에서는 일관된 행동을 선택하였음을 알 수 있으며, 함정을 피하고 보상점을 찾아 가는 행동양식을 학습하였음을 알 수 있다. 이에 비해 Q-Learning을 사용한 에이전트의 경우에는 학습 인자를 변화시켜서 수행한 결과 어떤 경우에 대해서도 퍼지 목적에 대한 에이전트의 최적의 정책을 학습하지 못했다.

5. 결론

본 논문에서는 에이전트에게 언어값으로 주어지는 사용자의 목적을 퍼지 집합의 순서쌍으로 구성된 퍼지 목적으로 표현하고, 외부 환경 또한 퍼지 상태로 표현하는 방법을 제안하였다. 이와 함께 환경에 대한 에이전트의 적응성을 위해, 퍼지 강화 함수를 제안하고 기존의 Q-Learning을 FuzzyQ-Learning으로 확장하였다. 또한 실험을 통해서 동일한 문제에 대해 기존 강화 학습 알고리즘 중 하나인 Q-Learning에 비해 그 성능이 우수하다는 것을 보였으며, 퍼지 강화 학습 에이전트가 사용자의 퍼지 목적에 대해 적합한 행동 양식을 학습할 수 있다는 것을 검증하였다.

6. 참고문헌

[1]Katia O. Sycara, "Multiagents Systems", *AI magazine*, Summer, pp.79-92. 1998

[2]Leslie Oack Kaelbling, Michael L. Littman and Antony R. Cassandr, "Planning and Acting in Partially Observable Stochastic Domains", *Artificial Intelligence*, Vol 101, 1998.

[3]Ping Xuan, Victor R. Lesser, "Handling Uncertainty in Multi-Agent Commitments", *Umass Computer Science Technical Report*, 1999-05, Jan 18, 1999.

[4]David Poole, "The independent choice logic for modeling multiple agents under uncertainty", *Artificial Intelligence*, Vol 94, July 1, 1997.

[5]Richard S. Sutton, Andrew G. Barto, "Reinforcement Learning – An Introduction", *The MIT Press*, 1998.

[6]C.J.C.H. Watkins, "Learning with Delayed Rewards", *Phd Thesis, Cambridge University, Psychology Dept.*, 1989.