2001년도 대한전기학회/대한전자공학회 시스템및제어분야 합동 추계학술대회 논문집 (2001.11.24)

강화학습을 이용한 다개체 시스템의 협조행동 구현

이창길*, 김민수, 이승환, 오학준, 정찬수
숭실대학교 전기공학과

Cooperative Behavior Using Reinforcement Learning for the Multi-Agent system

Chang-Gil Lee*, Min-Soo Kim, Seung-Whan Lee, Hak-Joon Oh, Chan-Soo Jung
Dept. of Electrical Eng. Soongsil Univ.

Abstract - 다수의 자율이동로봇으로 구성되는 다개체 시스템에서의 협조행동을 위해서 각 개체는 주변환경의 인식뿐만 아니라 환경변화에 적응할 수 있는 추론능력이 요구된다. 이에 본 논문에서는 강화학습을 이용하여 동적으로 변화하는 환경 하에서 개체들이 스스로 학습하고 대처할 수 있는 협조행동 방법을 제시한다. 제안한 방법을 먹이와 포식자 문제에 적용하여 포식자 로봇간의 협조행동을 구현하였다. 여러 대로 구성된 포식자 로봇은 회피가 목적인 먹이로봇을 추적하여 포획하는 것이 임무이며 포식자 로봇들 간의 협조행동을 위해 각 상태에 따른 최적의 행동방식을 찾는데 강화학습을 이용한다.

1. 서 론

우리가 어려운 문제에 부딪쳤을 때 그 해결책을 자연계의 생물로부터 찾는 경우가 종종 있다. 이러한 시도는 기존에 개별적으로 이루어지고 있던 자연계 생물에 대한 연구결과를 통합하고, 더 나아가 연구를 보다 활성화하면서 어떤 새로운 연구방법을 모색해보자는 취지중의 하나로 인공생명이라는 하나의 연구분야가 탄생하게 되었다[1]~[6].

본 논문에서는 인공생명을 갖는 로봇을 실현하기 위한 현재까지 연구된 기본모델인 신경회로망, 퍼지시스템, 진화알고리즘, 강화학습 중에 강화학습을 이용하였다. 강화학습은 자신과 환경과의 상호관계와 이에 따른 강화신호를 통하여 자신의 행동을 개선해 나가는 방법으로서 환경에 대한 정확한 사전지식이 없이 학습 및 적응성을 보장하기 때문에 로봇의 학습에 유용하다. 서로 협조하여 사냥하는 동물에서 볼 수 있듯이 각자의 임무를 가지고 목표를 달성하게 된다. 주어진 임무를 수행하는 동안 예측하지 못한 환경에 처하게 되었을 때 각 객체들은 즉각적으로 환경에 대처하여 적응 및 학습해 나가게 된다.

본 논문에서는 강화학습 중에서 Q-학습에 대해 간단히 알아보고, Q-학습을 먹이와 포식자 문제에 적용하여 특정 환경에서 먹이로봇을 포획하는 포식자로봇의 상태와 행동을 정의한다[7]. 먹이로봇을 포획하기 위한 정책과 학습알고리즘, 보상을 정의한다.

2. Q-학습(Q-Learning)

Q-학습은 모델이 알려지지 않더라도 지연된 보상으로부터 최적의 정책을 얻을 수 있는 강화학습의 한 방법이다. 최적의 정책(optimal policy)을 얻기 위해서는 상태전이 함수와 보상함수가 필요하다. 시간 t에서 개체는 상태 s_t에서 행동 a_t를 선택하고 확률적인 보상 r_t를 받는다. 상태 s_t에서 행동 a_t를 수행할 때 다음 상태 s_{t+1}이 될 확률함수를 $T(s_t, a_t, s_{t+1})$로 나타내고, 상태 s_t에서 행동 a_t를 선택할 때 받는 예상되는 보상함수를 $R(s_t, a_t)$로 나타낸다. 감쇄된 예상보상 $R(s_t, a_t)$는 개체가 있는 상태와 선택한 행동에 의존하고 개체의 목표는 감쇄된 예상보상을 최대로 만드는 정책을 찾는다. 감쇄된 예상보상은 아래 식 (1)과 같다.

$$R(s_t, a_t) = E\left[\sum_{i=0}^{\infty} \gamma^i r_{t+i}\right] \qquad (1)$$

여기서 γ는 감쇄계수이며 $0 \le \gamma < 1$의 범위를 가진다. 감쇄계수는 t시간에서 받은 보상이 현재 받은 보상에 비해 γ^i만큼 적다는 것을 뜻한다. Q-학습에서는 정책 π에 대해 Q값을 식 (2)와 같이 정의한다.

$$Q^{\pi}(s) \equiv R(\pi(s), a) + \gamma \sum_{s' \in S} T(\pi(s), a)\ V^{\pi}(s') \qquad (2)$$

$V^{\pi}(s')$은 다음상태의 정책 π에 대한 상태값 함수이다.
Q-학습에서는 최적의 Q값을 찾아내는 것이 목적이다. Q값과 최적의 정책 V^*의 관계식은 아래와 같다.

$$V^*(s) = \max_a Q^*(s, a) \qquad (3)$$

식 (4)는 결정적인 보상과 행동인 경우와 비결정적인 보상과 행동인 경우로 나누어진다.
결정적인 보상과 행동인 경우에 현재의 상태와 행동을 각각 s, a 라하고, 다음 상태와 그 상태에서의 행동을 s′, a′ 라 할 때 Q-학습의 식은 식 (4)과 같다.

$$Q(s, a) \Leftarrow r + \gamma \max_{a'} Q(s', a') \qquad (4)$$

비결정적인 보상과 행동인 경우 Q-학습은 식 (5)과 같다.

$$Q(s, a) \Leftarrow Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)] \qquad (5)$$

여기서, α는 학습률을 나타낸다. 현재 취한 행동에 의해 다음상태가 바뀌었을 때 그 상태에서의 Q값 중에서 최대를 나타내는 값이 있다. 이 값과 함께 전의 Q값, 학습률, 그리고 감쇄율로 새로운 Q값을 정한다.

3. 먹이-포식자 문제

3.1 먹이(Prey), 포식자(Predator) 문제

2대의 포식자로봇이 상대적으로 빠른 먹이로봇을 포획하는 방법을 학습하기 위한 환경이나 먹이로봇, 포식자로봇의 행동양식을 정의한다.

먹이로봇과 포식자로봇이 행동하는 환경은 모든 로봇에 막대한 영향을 줄 수 있다. 포식자로봇이 반복적으로 변화하는 환경에 적응하여 다음행동을 결정 할 수 있도록 한다. 아래의 환경은 로봇축구에 실제로 적용되어진다[8]. 환경은 축구경기장과 로봇이 있는 공간에 해당되고, 먹이로봇은 공에 해당되고, 포식자로봇은 공격수로봇에 해당된다. 공격수로봇은 $150cm \times 130cm$의 축구장에서 자유롭게 공을 향해 달려가는 행동을 학습하게 된다. 각 로봇의 위치는 로봇의 맨 위에 부착된 칼라색종이(Color Patch)를 이용하여 비젼시스템으로 인식한다. 비젼시스템에서 얻어진 위치정보를 가지고 각 공격수로봇들의 다음 동작을 결정하게 된다. 각각의 로봇은 $6.5cm \times 6.5cm \times 6cm$의 크기이고, 2개의 바퀴를 가지고 360° 회전 할 수 있다.

본 논문에서는 로봇축구에 실제로 적용하기 전에 시뮬레이션을 하기 위한 환경을 설정하도록 하였다. 각각의 로봇들의 위치는 비젼시스템에 의해 좌표값이 넘겨진다고 가정하기로 한다.

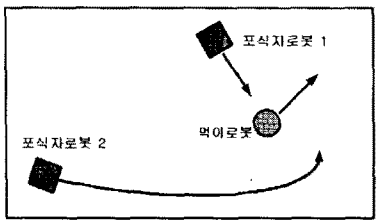

그림 1. 먹이-포식자 문제

먹이로봇은 포식자로봇이 다가오는 반대방향으로 이동을 한다. 먹이로봇은 항상 포식자로봇을 감지하여 잡히지 않는 최선의 행동 방법을 결정하여 행동한다.

먹이의 행동양식은 아래와 같이 나타낸다.

1. 멈추지 않고 임의의 방향으로 계속 움직임
2. 포식자로봇이 멀리 있으면 속도 줄임
3. 포식자로봇을 만나면 유리한 방향으로 이동
4. 바로 움직이거나 바로 멈출 수 없음
5. 장애물을 만나면 가장 빠른 방향으로 회피

본 논문에서는 두 대의 포식자로봇이 각각의 상대로봇과 먹이와의 위치를 고려하여 강화학습을 통해 먹이를 추적하여 포획하는 임무를 수행하도록 협조행동을 학습하게 된다. 각각의 포식자는 강화학습에 의해 학습된 판단력에 의한 제어 알고리즘을 바탕으로 행동한다. 포식자로봇은 먹이로봇보다 상대적으로 느리기 때문에 두 대의 포식자로봇이 서로 협조하여만 먹이로봇을 포획할 수 있다.

효과적인 협조행동을 위하여 포식자로봇은 다른 포식자로봇의 먹이로부터의 상대적인 위치를 알아야한다. 그림 1에서와 같이 포식자로봇1이 먹이에 더 가까이 있으므로 먹이로봇을 구석으로 몰고 가면 포식자로봇2가 뒤쪽으로 돌아 들어가면서 포식자로봇1과 함께 먹이로봇을 잡을 수 있다. 그러나 이렇게 협조행동을 하기에는 포식자로봇이 상대적으로 느리고, 먹이로봇의 행동을 예측할 수가 없기 때문에 먹이로봇을 포획하는데 어려움이 따른다. 이를 해결하기 위해서 시행착오에 대한 평가를 통한 최선의 방법을 찾아가는 강화학습을 이용하여 먹이를 추적한다.

3.2 포식자로봇의 설계

포식자로봇의 구조는 그림 2와 같이 먹이로봇, 포식자로봇1, 포식자로봇2의 거리와 방향을 입력받아서 강화학습에 의해 행동하게 된다.

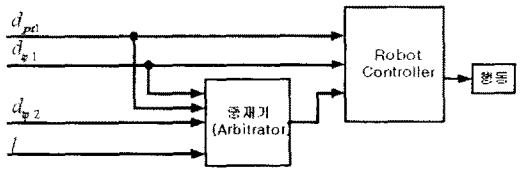

그림 2. 포식자로봇의 구조

Q-학습에 의해 학습된 최선의 행동으로 평가된 행동방식은 중재기내의 $Q(s,a)$ 배열에 저장되고, 이를 이용하여 가장 적은 Q값을 갖는 행동을 선택하여 포식자로봇의 제어기에 이동하는 방법을 알려준다.

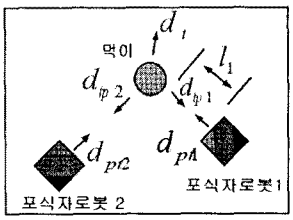

그림 3. 포식자로봇의 먹이로봇 추적

그림 3에서 먹이의 진행방향 d_t, 먹이로봇과 포식자로봇의 거리 l, 포식자로봇1의 진행방향 d_{t1}, 포식자로봇2의 진행방향 d_{t2}, 포식자로봇1이 바라본 먹이로봇의 방향 d_{pt1}, 먹이로봇이 바라본 포식자로봇1의 방향 d_{tp1}, 포식자로봇2가 바라본 먹이로봇의 방향 d_{pt2}, 먹이로봇이 바라본 포식자로봇2의 방향 d_{tp2}이다. 그림 4는 로봇의 방향에 따른 상태를 정의한 것이고, 표 1은 그림 4를 표로 나타낸 것이다.

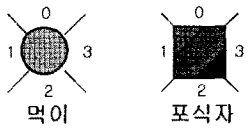

그림 4. 로봇상태의 정의

표 1. 방향에 대한 상태값

방 향	각 도	상태값
앞 쪽	$-\pi/4 \sim \pi/4$	0
왼 쪽	$\pi/4 \sim 3\pi/4$	1
뒤 쪽	$-3\pi/4 \sim 3\pi/4$	2
오른쪽	$-3\pi/4 \sim -\pi/4$	3

표 2는 포식자로봇의 행동을 정의한 것이다. 정의된 행동에 의해 로봇의 두 개의 바퀴를 움직이게 한다.

표 2. 행동에 대한 상태값

행 동	상태값
앞쪽으로 접근	0
왼쪽으로 접근	1
오른쪽으로 접근	2
뒤쪽으로 접근	3

표 3은 각 2bit씩 총 8bit를 가진 포식자로봇1의 상

태정보를 나타내고 있다.

표 3. 포식자의 상태정보

거리(l_1)		다른 포식자의 상태(d_{tp2})		먹이가 보는 방향(d_{tp1})		먹이를 보는 방향(d_{p1})	
d_7	d_6	d_5	d_4	d_3	d_2	d_1	d_0

본 논문에서 사용한 정책은 ***ε-greedy*** 방법으로 다음과 같이 정의된다.

1) 확률 1-ε을 이용하여 $Q(s,a)$값들 중에서 가장 큰 값으로 행동을 선택함.
2) 그렇지 않으면 무작위로 행동을 선택함.

학습알고리즘으로는 SARSA 알고리즘을 사용하였다. SARSA 알고리즘의 진행단계는 임의의 값으로 가치함수와 실행환경 초기화를 한 후, 정해진 정책에 따라 행동을 선택하게 된다. 평가함수를 갱신한다. 평가값 $Q(s,a)$의 갱신은 식 (6)과 같다.

$$Q(s,a) \Leftarrow Q(s,a) + a\,TDerr \qquad (6)$$

여기서 $TDerr$는 순간편차오류, a는 학습상수이다.

$$TDerr = r + \gamma Q(s',a') - Q(s,a) \qquad (7)$$

여기서 γ는 감쇄계수이다. 갱신한 후에 다음상태를 현재상태로 바꾼다. 다음상태를 현재상태가 될 때까지 정해진 정책에 따라 행동을 선택하는 것부터 다시 반복한다. 다음상태가 현재상태가 되면, 정해진 반복횟수 만큼 실행환경 초기화에서부터 지금까지의 과정을 반복한다. 이러한 과정으로 SARSA는 학습된 가치함수 Q를 이용하여 직접적으로 최적의 함수인 $Q^*(s,a)$로 향하게 만든다.

포식자로봇이 먹이로봇을 포획했을 경우 0을 받고, 포획에 실패하면 -1의 벌점을 받는다. 이와 같은 보상은 다음행동에서 선택에 영향을 주게된다.

4. 시뮬레이션

먹이로봇이 포식자로봇보다 상대적으로 빠르다. 시뮬레이션 환경은 먹이로봇의 속도는 $70cm/\sec$이고, 포식자로봇은 각각 $50cm/\sec$이다. 학습상수 $a=0.1$이고, 확률상수 $\varepsilon=0.05$이다. 감쇄계수 $\gamma=0.9$이다.

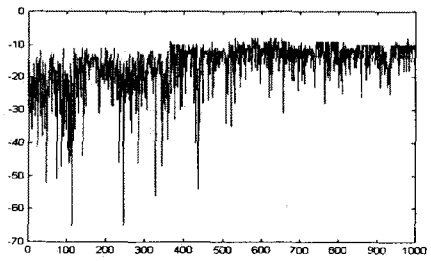

그림 5. 보상값 분포

그림 5에서 보상값의 분포를 살펴보면 총 1000번의 학습을 하면서 점점 해에 근접하는 것을 볼 수 있다. 시뮬레이션 결과에서 볼 수 있듯이 처음에는 의미없는 행동을 많이 하여 보상값이 많이 떨어졌지만, 학습을 통해 점점 더 해에 접근해 간다.

그림 6에서 학습전에는 먹이로봇을 포획하는데 많은 시간과 에너지가 들었지만, 그림 7과 같이 강화학습 후에는 두 포식자로봇이 환경의 변화에 잘 적응하여 서로 협동하여 먹이로봇을 포획하는 것을 볼 수 있다.

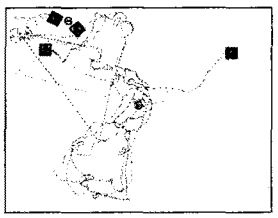

그림 6. 학습전

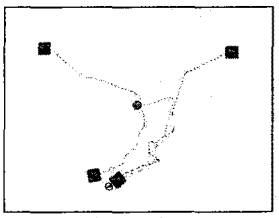

그림 7. 학습후

5. 결 론

군집생활을 하는 생물들의 대표적인 특징인 상호 협조행동을 인공적으로 구현하기 위해 다수의 자율이동로봇으로 구성된 다 개체 시스템을 이용한 연구가 활발하게 진행되고 있다. 이러한 협조행동을 하는 다수의 자율이동로봇들은 자기 자신의 상태뿐만 아니라 주변의 다른 개체나 환경의 상태를 고려하여 다음행동을 할 수 있도록 하는 고도의 추론능력이 요구된다.

이에 본 논문에서는 사전지식이 없는 상태에서 하나의 먹이로봇을 두 대의 포식자로봇이 협조행동을 통해 포획하는 문제에 강화학습을 사용하였다.

일정한 행동을 하는 먹이에 대하여 강화학습의 적용은 강화학습을 적용하기 전의 모습과는 다르게 먹이로봇에 보다 빠르게 접근하도록 하여 효과적으로 포식자로봇의 성능을 개선할 수 있다.

[참 고 문 헌]

[1] M. Sipper, *"An Introduction to Artificial Life"*, Explorations in Artificial Life, pp.4-8, 1995.

[2] J. S. Bay, "Design of the army-ant cooperative lifting robot", *IEEE Robotics and Automation Magazine*, Vol.2, No.1, pp.36-43, March 1995.

[3] K. S. Evans, "A Reactive Coordination Sheme for a many-robot system", *IEEE Trans. On Systems, Man and Cybernetics*, Vol.27, pp.598-610.

[4] L. E. Parker, *"Fault tolerant muti-robot cooperation"*, MIT Artificial Intelligence Lab Videotape AIV-9, 1994.

[5] L. Steels, "Cooperation between distributed agents through self-organization", *Proc. of Workshop on Multi-agent Cooperation*, 1989.

[6] S. K. Kim, "Generation of Rules for Swarm Intelligence of Autonomous Mobile Robots", *AROB2000*, pp.325-328, January. 2001.

[7] M-S Kim, "Avoidance Behavior of Small Mobile Robots based on the Successive Q-Learning", *2001 Int. Conference on Control, Automation and Systems(ICASE)*, Oct. 2001.

[8] K. C. Kim, "Evolutionary Programming and Q-Learning based Controller Design for Soccer Robot", *Proceedings of the 1st Workshop on Soccer Robotics*, pp.59-71.