강화 학습에 의한 로봇축구 에이전트 행동 전략

최소라*, 이승관**, 이영아***, 정태충*
*경희대학교 컴퓨터공학과
**가톨릭대학교 컴퓨터정보공학부
***용인송담대학 컴퓨터게임정보과
e-mail:sora615@gmail.com

Behavior Strategies of Robot Soccer Agent by Reinforcement Learning

SoRa Choe*, SeungGwan Lee**, YoungAh Lee***, TaeChoong Chung*
*Dept. of Computer Engineering, KyungHee University
**School of computer science & Information Engineering, Catholic University
***Dept. of Computer Game & Infomation, YongIn SongDam Collage

요 약

강화 학습이란 개체가 동적인 환경에서 시행착오를 통해 자신의 최적 행동을 찾아내는 기법이다. 특히 Q-learning과 같은 비(非)모델 기반의 강화학습은 사전에 환경에 대한 모델을 필요로 하지 않으며, 다양한 상태와 행동들을 충분히 경험한다면 최적의 행동 전략에 도달할 수 있으므로 여러 분야에 적용되고 있다. 본 논문에서는 로봇의 행동을 효율적으로 제어하기 위하여 Q-learning을 이용하였다. 로봇 축구 시스템은 공과 여러 대의 로봇이 시시각각 움직이는 시변 환경이므로 모델링이 상당히 복잡하다. 공을 골대 가까이 보내는 것이 로봇 축구의 목표지만 때로는 공을 무조건 골대 방향으로 보내는 것보다 더 효율적인 전략이 있을 수도 있다. 어떤 상황에서 어떤 행동을 하여야 장기적으로 보았을 때 더 우수한지 학습을 통해 로봇 스스로가 판단해가도록 시스템을 구현하고, 학습된 결과를 분석한다.

1. 서론

인간은 과거의 경험이나 학습을 통해 어떤 상황에 대한 적절한 판단을 하여 적응된 행동을 나타내는데, 이러한 것을 지능행동이라고 한다. 로봇이 인간형 로봇이 되기 위해 넘어야 할 결정적 기술은 역시 인간과 같은 학습능력일 것이다.

로봇 축구 시스템은 다중 에이전트 지능 제어 시스템이며, 지능 제어 시스템이란 환경에 대한 판단능력, 내부 상황에 대한 조절 능력, 문제 해결 능력 등을 갖춘 시스템을 말한다. 다중 에이전트 시스템을 구현함에 있어서 여러 수의 개체들로 이루어진 지능 시스템의 구현 원리를 정하고, 그 안에서 에이전트들이 독자적인 행동을 수행하면서 그와 동시에 서로 협력적인 행동도 할 수 있도록 하는 동작 메커니즘을 제공해 줄 수 있어야 한다. 로봇 축구는 위와 같은 지능 제어 시스템과 다중 에이전트 시스템의 특징을 모두 가지고 있으며, 그와 관련된 다양한 기법의 적용이 가능하다. 따라서 학문적으로도 연구 가치가 상당히 높은 시스템이라 할 수 있다.

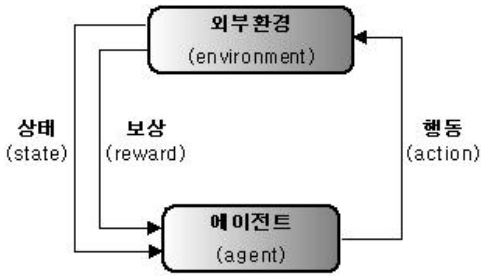
본 논문에서는 로봇이 특정 상황에서 어떤 행동을 해야 장기적으로 더 큰 이익을 얻을 수 있는지 판단할 수 있도록 하기 위해 강화학습을 이용하였다. 특히 Q-learning과 같은 비(非)모델 기반의 강화학습은 환경에 대한 모델 없이도 최적의 행동 전략을 찾는 것이 가능하므로 사전에 정확한 모델이 없는 로봇 축구 환경에 적합하다. 로봇이 실제 경기 경험을 통해 효율적 행동을 찾을 수 있도록 구현하였다.

2. 관련연구

2.1 강화학습(reinforcement learning)의 기본 원리

강화 학습이란 개체가 목적을 이루기 위하여 동적인 환경 속에서 시행착오를 통해 각 상황에서의 행

동을 학습하는 방법이다. 자신의 현재 환경에서 보상을 최대로 할 수 있는 최적의 행동을 선택한다. 보상이란 개체가 수행한 행동에 대해 환경에서 주는 잘하고 못함을 나타내는 값이다. 일반적으로 에이전트와 환경과의 관계는 (그림 1)과 같이 구성된다.



(그림 1) 강화 학습 시스템의 일반적인 구조

2.2 시뮤로솟(SimuroSot)

현재 FIRA Cup 정식 경기에 사용되는 SimuroSot은 실제 이용될 로봇 축구 전술을 사전에 시험할 수 있을 뿐만 아니라, 다중 에이전트 시스템 이론을 적용하여 시험해 보는데 쓰이고 있다. 실제 로봇 시스템을 이용해서 실험을 할 경우, 전략 뿐 아니라 로봇이나 비전 시스템 등의 성능에 크게 좌우되고, 경기를 할 당시의 비전 세팅 정도 등과 같은 경기 외적인 요소에도 상당히 많은 영향을 받는다. 하지만 시뮬레이션 경기의 경우에는 이러한 요소들이 배제되고, 오로지 전략에 따라서만 경기가 좌우되므로 전략이나 이론을 연구하고 테스트하는데 유용하게 사용될 수 있다 [4].

2.3 로봇축구와 강화학습

강화학습을 로봇 축구에 성공적으로 적용한 대표적인 시스템으로 Andhill에이전트가 있다. Andou[1]의 Andhill에서는 자신이 공을 가지고 있지 않은 경우에 정해진 위치에만 머물러 있지 않고, 동적으로 적절한 위치를 찾아 이동하는데 학습을 이용했다.

KAIST의 로봇 축구 팀 NaroSot[2]에서는 효과적인 지역 방어 전략을 수행하기 위해 동료들과 협조하는 방법에 모듈화 Q-learning을 적용하였다.

3. 로봇 축구 에이전트 행동 전략

3.1 문제 정의

로봇 축구에서 가장 기본이 되는 행동은 골대를 향해 직접 공을 차는 것이다. 그러나 축구에는 다양한 전략이 있으며, 어떤 상태에서는 공을 무조건 골대로 보내는 것보다 더 좋은 행동이 있을 수 있다.

로봇 축구 시스템은 동적 환경이기 때문에 분석이 상당히 복잡하고, 사람의 판단만으로는 결정하기 애매한 다양한 상황들이 존재한다. 이러한 환경에서 행동의 좋고 나쁨을 판단하는데 가장 좋은 것은 바로 실제 경기를 통한 경험이다.

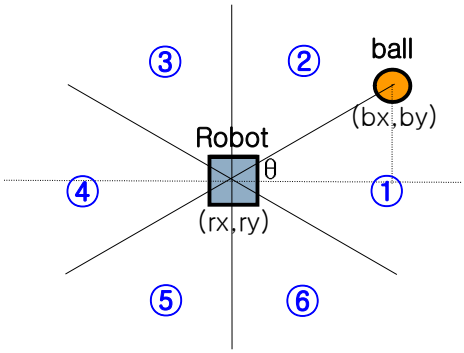
본 논문에서는 다양한 상황에서 로봇이 취할 수 있는 여러 행동 중 어느 것이 가장 우수한지를 판단하기 위해 Q-learning을 이용하였다.

3.2 상태 정의

학습을 하기 위해 다음과 같이 상태를 정의한다.

3.2.1 공과 로봇이 이루는 각도

경기장에서 핵심 지점은 양쪽 골대이다. 따라서 공이 로봇보다 위에 있는지 아래에 있는지 보다는 좌우의 위치가 더 중요하므로 (그림 2)와 같이 각도를 우선 세로선으로 2등분하고, 이등분한 것을 각각 위, 가운데, 아래로 3등분하였다. 상하의 위치를 더 작게 나누어도 전략에는 그다지 영향을 미치지 않았다.



(그림 2) 로봇과 공이 이루는 각도

이 때, 로봇은 공과 가장 가까이 있는 로봇이며, 직접 각을 구하지 않고 식(1)과 같이 좌표 값을 이용해 sinθ값을 구하여 상태를 판단한다.

$$\sin\theta = \frac{by - ry}{\sqrt{(bx - rx)^2 + (by - ry)^2}} \qquad (1)$$

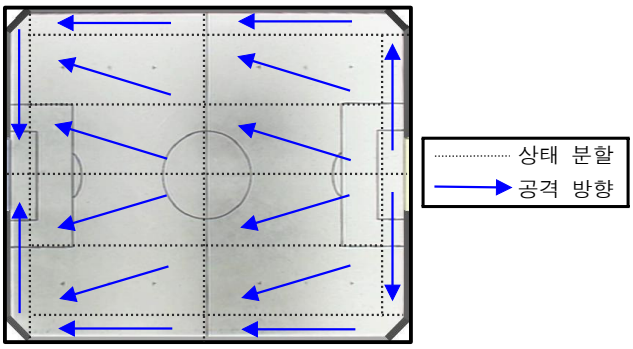
예를 들어, bx<rx이고, sin-30<sinθ<sin30을 만족한다면 1번 상태이다.

3.2.2 공과 로봇 사이의 거리

공과 로봇 사이의 거리는 4가지 상태로 나누었다. 로봇 크기의 3배를 한 상태로 하여 거리에 따라 d≤8.7, 8.7<d≤17.4, 17.4<d≤26.1, d>26.1의 상태를 갖도록 하였다.

3.2.3 공의 위치

공의 위치는 (그림 3)의 점선과 같이 공이 어느 진영에 있는지와 그곳이 골대와의 거리에 따라 나누었다. 경기장 테두리 부분에서는 벽이 있어서 다른 경우와 달리 상황이 특이하므로 좀 더 세분화하였으며, 총 16가지 상태로 정의하였다.



(그림 3) 경기장 상태 분할 및 측면 공격 방향

3.3 행동 정의

로봇이 취할 수 있는 행동은 4가지로 정의한다.

A1: 직접 골대를 향해 공격

직접 골대를 향해 공을 차는 경우의 목표점은 골대 입구의 가운데 점이다. 경기장의 어느 위치에서든 그 점을 목표로 하여 공을 차게 된다.

A2: 다른 로봇을 향해 패스

다른 로봇에게 패스할 때는 자신보다 골대에 가깝게 있는 로봇 중 가까운 로봇을 향해 공을 보낸다. 이때의 목표점은 공을 받을 로봇의 좌표에서 골대 쪽으로 조금 더 간 지점이다.

A3: 측면 공간을 이용한 공격

실제 축구 경기에서도 쓰이는 방법으로 공을 골대 가운데로 몰아가는 것이 아니고, 측면 공간을 이용해서 공격하는 것이다. 각 상태에서 공을 몰고 가야 하는 방향은 그림3에 화살표로 나타나 있다. 상대방의 전략에 따라 가운데로 공격하는 것이 좋은지 측면 공격이 좋은지가 결정된다.

특히 테두리 부분에서는 벽을 이용해서 공을 골대 쪽으로 몰고 가는 것이 가능한데 이 경우에는 로봇과 벽 즉, 2개의 면을 이용하게 되므로 골을 몰고 가기가 훨씬 수월하다.

A4: 회전을 이용한 패스

회전을 이용한 패스는 기존 로봇 축구에서 보기 힘든 새로운 기술로 원하는 방향으로 공을 차기 위한 각이 나오지 않는 경우나 상대편 로봇이 많은 경우에 매우 유용하다. 회전으로 패스를 하기 위해서는 먼저 공에 가까이 다가간 후, 보내려는 방향으로 회전을 하도록 한다. 아래와 같이 양쪽 바퀴값을 다르게 주면 로봇이 돌면서 그 힘으로 공을 치게 된다.

velocityLeft=-100, velocityRight=120;

이 경우는 오른쪽 바퀴 값이 더 크므로 왼쪽으로 회전을 하게 될 것이다. 값을 반대로 주면 오른쪽으로 회전을 하게 된다. 회전의 단점은 공을 원하는 곳으로 보내지 못할 확률이 다른 전략에 비해 크다는 것이지만 정상적인 패스가 힘든 상황에서는 상당히 효율적인 전략이다.

3.4 행동 전략의 학습

Q-learning은 시간차(temporal difference : TD) 학습 방식 중의 하나로, 에피소드의 끝에 도달할 때까지 기다리는 대신 바로 다음 상태의 예측 값을 이용한다. 하나의 시점과 다음 시점 사이에는 기대 반환의 차이가 있을 수 있다. 이 시간에 따른 기대 반환의 차(Δ_t)를 이용해 예측 값을 갱신한다.

$$\Delta = r + \gamma V(s') - V(s) \quad (2)$$

$$V(s) \leftarrow V(s) + \alpha_s \Delta_t \quad (3)$$

α_s는 학습률이며, γ는 절감계수로 상태가 멀어짐에 따라 주어지는 보상을 감소시키는 역할을 한다. 현재의 상태와 행동은 각각 s, a이고 다음 상태와 행동은 s′, a′이다. 이 때 Q값을 위한 식은 다음과 같다.

$$Q(s,a) \leftarrow Q(s,a) + \alpha_s (r + \gamma \max_a Q(s',a') - Q(s,a)) \quad (4)$$

경기가 처음 시작될 때의 상태가 시작점이고, 이 상태는 여러 상태들을 거쳐 목표점에 도달하게 되는데 이것이 한 에피소드이다. 공이 골대 안으로 들어간 상태가 목표점이며, 목표점에 도달하면 경기는 다시 새로운 위치에서 시작되고 이것은 새로운 시작점이 된다.

그리고 슛이 성공한 경우에만 보상을 주도록 하였다. 로봇 축구에서의 궁극적인 목적은 골을 넣는 것이므로 공을 골대 근처에 계속 가까이 보내는 전략이 있다 해도 골대 안에 넣지 못한다면 좋은 전략이 아니다. 실제 구현 결과 자기편 골대에 공이 들어가는 경우가 가끔 발생하여 이 경우에는 벌점을 주어서 전체적으로 좀 더 빠른 학습이 이루어지도록 하였다.

상태가 변화할 때마다 선택했던 행동의 Q값이 식(4)와 같이 갱신된다. SimuroSot에서는 슛이 성공하면 프로그램이 완전히 다시 시작되기 때문에 파일 입출력을 통해 Q테이블을 지속적으로 관리하였다.

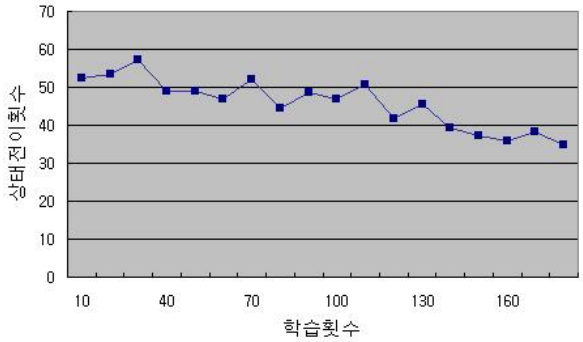
4. 실험 및 결과분석

가장 기본적인 상황에서의 전략을 찾기 위해 상대편에는 골키퍼 로봇만 두었다. 언제나 공과 가장 가까운 로봇만이 공격을 하고, 공격하는 로봇 외에 다른 로봇들은 자신이 맡은 위치에서 대기하도록 한다. 여러 대의 로봇이 한꺼번에 공을 치면 제대로

각 행동의 성능을 평가할 수 없기 때문이다.

학습률(α)과 절감계수(γ)는 각각 로봇축구에서 잘 알려진 값 0.8, 0.5을 이용하였다 [4]. Q값은 0으로 초기화하였으며, 각 행동의 Q값이 모두 0인 첫 번째 학습 때는 A1을 선택하도록 하였다. 다음에서의 행동 선택은 ε-greedy 정책에 따라 결정된다.

시작지점에서 목표지점, 즉 골이 들어가기 전까지 지나는 상태전이 횟수를 통해 학습의 성능을 알아보았다. (그림4)는 10회 학습이 이루어질 때마다의 평균 상태전이 횟수로 150회 이후에서 수렴하였다. 무조건 공을 골대 쪽으로 보내는 것보다 다른 방향으로 공을 보내는 전략들을 함께 학습함으로써 더 빠른 시간 내에 골을 넣을 수 있음을 알 수 있다.



(그림 4) 학습 성능 평가

다음 <표1>은 위의 170회 학습을 거친 Q테이블을 분석한 결과이다.

<표 1> 각 지역에서의 최적 행동 비교

	A1 (%)	A2 (%)	A3 (%)	A4 (%)
전체	18.2	18.2	29.2	34.4
테두리지역	23.3	23.3	19.2	34.2
안쪽지역	13.4	14.6	37.8	34.1
우리진영	22.6	18.7	25.3	33.3
상대진영	13.6	19.7	32.1	34.6

A1 직접 골대를 향해 공격 A3 측면 공간을 이용한 공격
A2 다른 로봇을 향해 패스 A4 회전을 이용한 패스

각 지역 또는 진영에서의 상태들마다 가장 큰 Q값을 갖는 행동의 수를 세어 비교한 표이다. 로봇과의 각도나 거리를 제외하고, 공의 위치만으로 본 상태이다. 전체적으로 A3과 A4에서 높은 비율을 보이고 있다. 실험 전에는 공을 골대를 향해 보내는 A1이 가장 효율적일 것이라고 생각했으나, 회전과 측면 공격이 오히려 더 좋은 결과를 가져왔다. 안쪽지역에서는 테두리지역에 비해 A1과 A2의 비율이 더 낮고, 그만큼 A3의 비율이 높았는데 이것을 통해 안쪽에서는 A1이나 A2같은 중앙 공격보다는 측면 공격이 더 효과적임을 알 수 있다. 테두리 지역에서는 A1이나 A2보다도 A3의 비율이 더 낮았다. A4 즉, 회전 기술은 공의 위치 뿐 아니라, 여러 상태에서 모두 높은 비율을 보였다.

우리진영과 상대진영에서의 비교 결과, 우리진영보다 상대진영에서 A1이 더 유리할 것이라는 일반적인 생각과는 달리 상당히 낮은 비율이 나왔다. 상대진영에서 골대를 향한 공격은 골키퍼에 의해 막혀서 성공률이 떨어지기 때문으로 분석된다. 결론적으로 골의 성공률을 높이기 위해서는 기본적인 패턴으로 공격하기 보다는 A3과 A4 같은 전략이 필요하다.

5. 결론 및 향후 연구과제

강화학습은 에이전트가 자신의 행동을 게임 세계에 점진적으로 적응시켜 나가도록 한다. 본 논문에서는 로봇이 특정 상황에서의 최적 전략을 스스로 찾아가도록 하는데 중점을 두었다.

실험에서 작성한 Q테이블을 분석해 경기에 효율적으로 사용할 수 있는 휴리스틱을 얻을 수 있었으며, 로봇축구에 적용하여 긍정적 효과를 볼 수 있는 Q-learning을 구현하였다. 이미 기본 상황에서 학습된 테이블을 초기값으로 하여 다른 상황에서 학습을 한다면 빠른 수렴 결과를 얻을 것이다.

본 전략은 SimuroSot 뿐 아니라 MiroSot에도 이용할 수 있다. 단, MiroSot은 시뮬레이션이 아닌 실제 환경이므로 위에서 다룬 요소들 외에 좀 더 다양하고 복잡한 요소들이 고려되어야 하기 때문에 이에 맞는 여러 가지 연구가 이루어져야 한다.

참고문헌

[1] Andou, T, "Refinement of soccer agents positions using reinforcement learning.", Robocup-97 : Robot Soccer World Cup 1, 1998.

[2] J.H. Kim, "Modular Q-learning based multi-agent cooperation for robot soccer", Robotics and Autonomous Systems, Vol. 35, 2001.

[3] Steve Rabin. “AI game programming wisdom 2” Charles River Media, 2003.

[4] 김종환 外 8인, “로봇축구공학”, 브레인코리아, 2002. 11, pp.393-431.

[5] 김인철, “강화 학습에 기초한 로봇 축구 에이전트의 설계 및 구현”, 한국정보처리학회 논문지B 제9-B권 제2호, 2002.

[6] 도현호, “다중에이전트 행동기반의 강화학습에 관한 연구” 한국정보처리학회 VOL. 09 NO. 02, 2002.