다중 에이전트 강화 학습을 위한 상태 공간 타일링과 확률적 행동 선택

권기덕° 김인철
경기대학교 전자계산학과
{kdkwon°, kic}@kyonggi.ac.kr

State Space Tiling and Probabilistic Action Selection for Multi-Agent Reinforcement Learning

Kwon Ki Duk° Kim In Cheol
Dept. of Computer Science, Univ. Kyonggi

요 약

강화 학습은 누적 보상 값을 최대화할 수 있는 행동 선택 전략을 학습하는 온라인 학습의 한 형태이다. 효과적인 강화 학습을 위해 학습 에이전트가 매 순간 고민해야 하는 문제가 탐험(exploitation)과 탐색(exploration)의 문제이다. 경험과 학습이 충분치 않은 상태의 에이전트는 어느 정도의 보상 값을 보장하는 과거에 경험한 행동을 선택하느냐 아니면 보상 값을 예측할 수 없는 새로운 행동을 시도해봄으로써 학습의 폭을 넓힐 것이냐를 고민하게 된다. 특히 단일 에이전트에 비해 상태공간과 행동공간이 더욱 커지는 다중 에이전트 시스템의 경우, 효과적인 강화학습을 위해서는 상태 공간 축소방법과 더불어 탐색의 기회가 많은 행동 선택 전략이 마련되어야 한다. 본 논문에서는 로봇축구 Keepaway를 위한 효율적인 다중 에이전트 강화학습 방법을 설명한다. 이 방법의 특징은 상태 공간 축소를 위해 함수근사방법의 하나인 타일 코딩을 적용하였고, 다양한 행동 선택을 위해 룰렛 휠 선택 전략을 적용한 것이다. 본 논문에서는 이 방법의 효과를 입증하기 위한 실험결과를 소개한다.

1. 서 론

강화 학습은 교사 학습(Supervised Learning)처럼 예제 집합이 주어지지 않고, 성취해야 할 목표(goal)와 행동을 평가하는 보상 함수가 주어지는 학습 방법이다. 강화 학습을 하는 에이전트는 목표에 도달할 수 있는 전략을 찾기 위해서 미지의 환경과 반복적인 상호 작용을 하면서 얻은 경험을 통해 점진적으로 학습한다[1,5]. 에이전트의 학습 목표인 최적의 전략은 상태와 행동의 쌍으로 된 형태이다. 에이전트는 만약 경험한 상태를 다시 만나게 되면, 기억하고 있는 전략을 수행하거나 다양한 전략을 평가하기 위해서 다른 행동을 선택할 수 있다. 강화 학습은 사전 지식을 가지지 않고 미지의 상태 공간을 탐색하므로 학습 속도가 매우 느리고 불안정할 수 있지만, 새로운 환경을 학습할 수 있다는 장점이 있다. 에이전트와 환경의 상호 작용은 실시간으로 이루어진다. 에이전트가 환경의 상태를 인식하고, 행동을 선택하여 수행을 하면 환경은 그 행동의 영향으로 변화가 있게 되고 에이전트는 환경의 변화된 상태를 다음 상태로 다시 인식하는 과정을 반복하게 된다. 강화 학습은 상태 정보를 이용해서 예측을 하므로 에이전트가 정확한 예측을 하기 위해서는 상태가 환경에 대한 모든 것을 함축하고 있어야 한다. 하지만 대부분의 실세계 문제에서 센서는 왜곡된 값을 감지할 수 있고 상태 값의 측정이 지연될 가능성이 있다. 또한, 영향이 있는 모든 요소들이 상태로 구성이 된다면 보다 정확한 예측을 할 수 있지만, 하나의 요소를 추가하면 상태 공간의 크기가 지수적으로 증가해서 강화 학습의 문제점인 거대한 상태 공간의 문제를 해결해야 한다.

본 논문에서는 다중 에이전트 강화 학습을 Keepaway를 이용하여 실험하고자 한다. Keepaway는 실시간 환경을 제공하는 로보컵 축구 시뮬레이션 리그를 이용한 강화 학습 테스트용으로 만들어진 도구이다. Keepaway의 상태 변수는 무수히 많은 실수로 구성된 거대한 상태 공간으로 이루어져 있다. 이를 해결하기 위해 상태 공간을 표현하는 상태 변수를 일정한 기준에 의해 간격을 두고 나누어 상태를 간략화하는 타일 코딩 방식으로 해결하였다. 또한, 행동 선택은 Q학습에 의한 Q값에 의해 최적의 행동 선택만을 하는 것을 지양하고 학습 초기에 선택 가능성은 적더라도 행동의 다양성을 위해 확률적 방법인 룰렛 휠 방식을 이용하여 다양한 행동이 선택될 가능성을 높여 주도록 하였다.

2. 관련연구

2.1 강화 학습

강화학습은 목표 지향적 학습 방법으로 시행착오, 지연된 보상, 그리고 주어진 환경과의 상호작용을 통하여 학습하는 특성을 가지고 있다. 에이전트는 특정 시각의 주어진 상태에서 에이전트 자신의 정책에 따라 행동을 결정하고 자신이 위치한 환경으로부터 스칼라 값의 보상 신호를 받게 되고 새로운 상태로 전이된다. 여기에서 정책이란 주어진 시각에 에이전트가 취할 수 있는 행동 집합에서 특정 행동을 선택하게 되는 기준이다[2,3].

$$V^{\pi}(s_t) \equiv \sum_{i=0}^{\infty} \gamma^i r_{t+i} \qquad [식\ 1]$$

$$\pi^* \equiv \arg\max_{\pi} V^{\pi}(s), (\forall s) \qquad [식\ 2]$$

[식 1]은 행동 전략 π에 따른 누적 보상 값을 나타내고, [식 2]는 이러한 누적 보상 값이 최대가 될 수 있는 최적의 행동 전략을 나타내고 있다. Q-학습과 같은 비 모델 기반의 강화학습은 사전에 환경에 대한 별다른 모델을 설정하거나 학습할 필요가 없으며 다양한 상태와 행동들을 충분히 자주 경험할 수만 있으면 최적의 행동전략에 도달할 수 있어 다양한 응용분야에 적용되고 있다. 일반적인 Q-학습법에서 Q-값 갱신은 [식 3]과 같다.

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r - \gamma \max_{a'} Q(s',a') - Q(s,a)] \qquad [식\ 3]$$

실제 응용분야에서 Q-학습과 같은 강화학습이 겪는 최대의 문제는 큰 상태 공간을 갖는 문제의 경우에 학습에 필요한 Q-테이블의 크기가 너무 커진다는 것이다. 따라서 적절한 시간 내에 가능한 모든 상태와 행동들에 대한 최적의 Q-값을 계산해낼 수 없어 실효를 거두기 어렵다는 점이다.

2.2 Keepaway

Keepaway는 로보컵 축구 시뮬레이션 리그를 이용한 강화 학습 테스트용으로 만들어진 도구이다. 로보컵 축구 서버로부터 Keepaway에 필요한 정보를 획득하여 시뮬레이션한다[4,6]. Keepaway는 keeper와 taker로 구성되며, 로보컵 축구시뮬레이션 리그의 제한된 영역을 사용한다. keeper의 목표는 공을 점유, 유지하는 것이고, taker의 목표는 공을 뺏는데 있다. 파라미터는 영역의 크기, keeper와 taker의 개수이다. Keepaway는 keeper가 공을 빼앗기거나 영역 밖으로 나갔을 때를 하나의 에피소드로 한다.

로보컵 축구 시뮬레이션 리그와의 공통점은 로보컵 축구 시뮬레이션 리그의 환경 특성인 잡음을 허용하며, 각각의 에이전트가 서로 통신하지 않는다는 것이다. 그래서 코치(coach) 에이전트를 두어 독립적인 정책을 가지고 동시에 조절할 수 있는 방책을 마련하였다. 에이전트는 매 150ms마다 공이나 상대편 에이전트 등의 정보를 받아들이며, 매 100ms마다 가장 기본적인 행동인 kick, turn, dash를 수행한다.

[그림 1] Keepaway 실행 화면

로보컵 축구 시뮬레이션 리그와의 차이점은 각각의 에이전트가 가지는 센서 정보 중 기본적인 행동을 하기 위한 각도나 스태미나, 힘에 대해서는 고려하지 않는다. 각도의 경우에는 360도 전체를 볼 수 있다고 가정한다. 또한 환경의 복잡성을 줄이기 위해 로보컵 축구 시뮬레이션 리그 전체 영역을 사용하는 것이 아니라 제한된 영역만을 사용한다.

3. 다중에이전트 강화학습

3.1 상태

본 논문의 실험 환경인 Keepaway는 로봇 축구 시뮬레이션 리그 게임이 제공하는 실시간 환경을 기반으로 강화 학습을 실험하기 위한 환경으로 변형한 환경이다. Keepaway는 keeper와 taker로 구성되는데, 각 에이전트는 자신의 행동을 선택함으로써 전체 팀의 목표를 달성하기 위한 협력이 이루어지도록 하였다. 상태 변수는 다음과 같이 정의한다.

K1과 중심과의 거리, K2와 중심과의 거리, K3과 중심과의 거리, T1과 중심과의 거리, T2와 중심과의 거리, K1과 K2와의 거리, K1과 K3과의 거리, K1과 T1과의 거리, K1과 T2와의 거리, K2와 상대편과의 최소 거리, K3과 상대편과의 최소 거리, K2가 K1을 바라보는 각도의 최소, K3이 K1을 바라보는 각도의 최소

여기에서 K1, K2, K3은 keeper의 번호이며, T1,T2는 taker의 번호이다. 대부분의 경우, 특히 6,7,8,9의 경우 K1이 공을 소유하고 있는 keeper의 번호이기 때문에 K1을 기준으로 거리를 계산한다.

강화 학습에서 타일 코딩은 상태 행동 값들의 선형 간략화(linear approximator) 하는데 사용된다. 각 타일은 불리언(boolean) 값들로 이루어져있으며 연속적인 특징 벡터들의 집합으로부터 높은 차원의 불리언 벡터의 매핑에 의해 얻어진다. 각각의 타일은 중첩되어 있으며 같은 메모리를 공유한다. 또한, 포인터를 주어 현재의 상태를 나타내도록 하였다. 해싱은 일반적으로 타일 코딩에 사용하는 방법인데 여기서 각 타일은 고정된 같은 크기를 가진다. 상태 공간의 각 차원에 타일링이 분배되고 각 타일링은 타일들로 분할된다. 복잡도는 타일링과 타일의 수에 따라 증가한다. 각 타일링에서 그 상태-행동 쌍을 포함하는 타일이 활성화되면 활성화된 타일의 값의 합이 해당 상태-행동 쌍의 Q값이 된다.

타일의 구조와 기본 연산은 다음과 같다.

타일 구조 : N_T개의 타일링 집합으로 구성된다.

$$\{T_i, i=1..N_T\} \qquad [식\ 4]$$

각 타일링 T_i는 n_i개의 타일로 구성된다. 각 타일 $f_{ij}(1 \le j \le n_i)$은 가중치 w_{ij}와 적합도 e_{ij}를 기억한다.

타일의 선택 : 각 타일링 T_i에서 상태 χ_q와 행동 a_q으로 이루어진 질의 (χ_q, a_q)를 포함하는 타일 f_{ij}가 활성화 된다. 활성화된 모든 타일들의 집합 $F(\chi_q, a_q)$는 다음과 같다.

$$F(x_q, a_q) = \{f_{ij} \in T_i \mid (x_q, a_q) \in f_{ij}\} \qquad [식\ 5]$$

질의 평가 : 질의에 대한 Q값은 $F(\chi_q, a_q)$에 포함된 타일들의 가중치 합이다.

$$Q(x_q, a_q) = \sum_{f_{ij} \in F(x_q, a_q)} w_{ij} \qquad [식\ 6]$$

Q값의 갱신 : 모든 타일의 가중치 w_{ij}는 다음과 같이 갱신된다.

$$\Delta w_{ij} = \alpha(r_{t+1} + \gamma Q_{t+1} - Q_t)e_f \qquad [식\ 7]$$

적합도의 갱신 : 모든 타일들의 적합도는 다음과 같이 갱신된다.

$$e_{ij} \leftarrow \begin{cases} \dfrac{1}{|F(x_q, a_q)|} & \text{if } f_{ij} \in F(x_q, a_q) \\ \lambda\gamma e_{ij} & otherwise \end{cases} \qquad [식\ 8]$$

Q값의 갱신과 적합도의 갱신은 타일의 수만큼 이루어진다. 타일과 타일링의 수가 많을수록 좋은 결과를 얻을 수 있지만 기억 장소도 비례해서 증가한다.

3.2 행동

본 논문에서 제안한 다중 에이전트 강화 학습을 이용한 행동 선택을 하는 에이전트는 공을 점유, 유지하기 위해 3개의 keeper 에이전트가 각자의 행동을 자율적으로 선택하여 실행하는 에이전트 특성을 가지고 있다. Keepaway에서 강화 학습을 사용하기 위해 학습자는 로보컵 축구 시뮬레이션 리그의 기본적인 행동 뿐 만 아니라 CMUnited99 팀에서 사용한 상위 레벨의 매크로 행동을 선택한다[3]. 상위 레벨의 매크로 행동에는 *HoldBall()*, *PassBall(k)*, *GetOpen()*, *GoToBall()*, *BlockPass(k)*가 있다

- *HoldBall()*: 가능한 상대편으로부터 멀리 떨어져서 공을 점유, 유지하는 동안 움직이지 않고 남아 있는 행동
- *PassBall(k)*: keeper k 앞으로 직접 공을 차는 행동
- *GetOpen()*: 공이나 상대편으로부터 자유로운 위치로 이동하는 행동
- *GoToBall()*: 공을 가로채기하거나 정지된 공 앞으로 이동하는 행동
- *BlockPass(k)*: 공을 가지고 있는 keeper와 keeper k 사이로 이동하는 행동

다중 에이전트의 상태 공간은 <S, a, h, r>로 정의한다. S는 상태 공간, a는 에이전트가 취할 수 있는 행동(a_t)들의 유한 집합으로 keeper 에이전트 수 -1개의 행동을 의미한다. h는 각 에이전트가 현재 상태에서 다음 상태로 전이하기 위한 상태전이 함수, 그리고 r은 에이전트가 선택한 행동이 얼마나 좋은가에 대한 정도를 나타내는 보상 값이다. 각 에이전트

(A_i)는 공을 점유, 유지하기 위해 상태 공간으로부터 에이전트가 선택한 행동(a_t)에 대한 보상으로 보상 값(r^i_{t+1})과 다른 에이전트들에 대한 상태 정보(s^i_t)를 입력 받아 상태 전이 함수(h)에 의해 최적의 행동을 선택하여 다음 상태(s^i_{t-1})로 이동한다. 각 에이전트는 현재 상태(s^i_t)에서 선택 가능한 모든 (상태-행동)쌍에 대한 평가 값을 [식 9]와 같이 갱신한다.

$$For\ all(s_t, a_t) \qquad a \in A(s_t) \qquad [식\ 9]$$
$$Q(s_t, a_t) = (1-\alpha)Q(s_t, a_t) + \alpha[r_{t-1} + \gamma \max Q(s_{t-1}, a_{t-1}) - Q(s_t, a_t)]$$

[식 9]에서 α는 학습 율, γ는 할인 율, 그리고 r은 에이전트가 선택한 행동에 대한 평가를 나타내는 보상 값이다. 행동의 선택은 [식 10]과 같은 확률에 의해 확률적으로 결정된다.

$$P(a^i_t | s_t) = \frac{Q(s_t, a^i_t)}{\sum_{i=0} Q(s_t, a^i_t)} \qquad [식\ 10]$$

[식 10]의 $P(a^i_t|s_t)$는 상태 s_t에 행동a^i_t를 선택할 확률을 나타내며, 이 확률은 이 상태에 적용 가능한 모든 행동들의 Q값의 합에 대한 행동a^i_t의 Q값의 크기에 비례하여 결정된다. 강화학습 에이전트는 매 순간 적용 가능한 행동들에 대한 Q값을 기초로 이와 같은 확률을 계산하고 이 확률에 따라 행동을 선택한다. 이러한 방법은 일반적으로 룰렛 휠 방식이라 부르며, 단순히 Q값이 최대인 행동을 선택하는 방법보다 경험해보지 않은 낮은 Q값의 행동도 선택될 수 있도록 해주는 것이 특징이다.

3.3 보상 값

보상 값의 경우 keeper의 목표가 공을 점유, 유지하는데 있으므로, 나에게 유리하고 적에게 불리한 점을 고려하여 목표를 달성하기 좋은 기준으로 정한다. 보상 값의 우선순위는 다음과 같다.

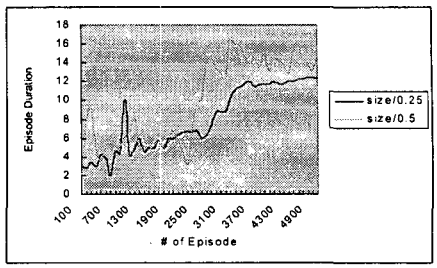
1. 일정 시간 동안 공을 우리 편이 점유하고 있을 경우
2. 공이 밖으로 나갔을 경우
3. 공을 상대편에게 빼앗겼을 경우

보상 값은 하나의 에피소드가 끝나면 주도록 하였다. 1의 조건일 경우 1을, 3의 경우는 0을 주도록 하였으며, 2의 경우에는 공을 상대편에게 빼앗기지 않았으므로 0.5를 주도록 하였다. 1의 경우 일정 시간은 12ms로 하였다. 이는 기존의 Keepaway에서 실험한 결과, 학습을 통한 에피소드의 평균 시간이기 때문이다.

4. 실험

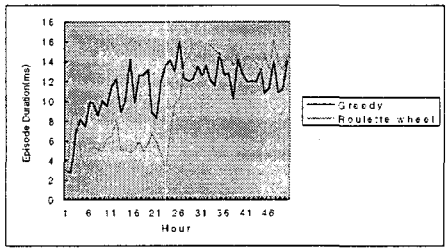
3대2 Keepaway의 경우에는 13개의 상태 변수를 가진다. 하나의 타일은 하나의 상태 변수를 가지고 구성하게 되며, 하나의 상태를 표현한다. 본 논문의 실험 환경인 keepaway의 Featurc Sct은 3대2 Keepaway의 경우, 13개의 상태 변수와 앞에서 언급한 행동 5개의 불리언 조합인 13 X 2^5 = 416개의 타일로 구성된다. 중심을 기준으로 각 keeper간의 거리를 상태 변수로 하였기 때문에 각 타일의 크기는 고정된 값으로 전체 상태 공간 20X20의 중간인 10으로 고정하였다.

행동 선택은 하나의 상태에서 각 keeper가 취할 행동들에 대한 Q값들을 계산하고 그 값들을 기준으로 비율을 정하며 이를 바탕으로 행동이 랜덤하게 선택될 수 있도록 하였다. Keepaway의 경우 에피소드를 단위로 에피소드의 기간이 얼마나 지속되는가를 학습의 측정치로 사용하였다. 총 50시간씩 10번에 걸쳐 실험하였으며, 에피소드는 약 100만개 정도이다. Q학습의 학습율은 0.125, 할인율은 0.9로 하여 그리디 행동 선택과 룰렛 휠 방법을 비교 실험하였다.



[그림 2] 타일 크기에 따른 에피소드의 변화

그림2에서 볼 수 있듯이 타일의 크기가 작을 경우 상태 공간의 크기가 커지기 때문에 오랜 경험을 필요로 하는 것을 볼 수 있다. 상대적으로 타일 크기를 최대로 선정하여 실험하였을 경우 더 효율적인 Q 학습이 이루어짐을 볼 수 있었다.



[그림 3] 시간에 따른 에피소드 기간의 변화

그림3에서 볼 수 있듯이 최대 Q값에 의한 최적의 행동 선택을 하는 그리디 행동 선택(greedy action selection)의 경우 에피소드의 기간(duration)이 짧은 시간에 최적의 행동만을 선택하여 반복적으로 수행하는 것을 볼 수 있다. 그러나 본 논문에서 제시한 룰렛 휠 방식을 채택하여 행동을 선택한 경우에는 그리디 방식보다 좀 더 오랜 시간 후에 학습에 의한 최적의 행동을 수행함을 볼 수 있다. 이를 통해 다중 에이전트 강화 학습을 이용한 행동 선택 전략에 자율성을 보장하였다.

5. 결론

본 논문에서는 강화 학습을 이용한 다중 에이전트의 자율적 행동 선택을 실험하기 위해 Keepaway를 이용하여 성능을 평가하였다. 상태 변수를 간략화하기 위해 타일 코딩 방법을 사용하여 상태를 일반화하였다. 탐험과 탐색의 균형 문제를 해결하기 위해 확률적 선택 방법인 룰렛 휠 방식을 채택하였다. 그리고 실험을 통해 탐색의 기회가 많은 행동 선택 전략도 긴 시간 후에는 최대 Q값에 의한 행동 선택과 마찬가지로 최적의 행동을 선택함으로써 효과적인 강화학습이 이루어졌음을 확인 할 수 있었다. 향후 연구로는 강화 학습의 다른 유형과의 비교 실험과 행동 선택에 있어서 확률적 방법 이외에 다른 방법 등을 적용해 보고자 한다.

참고문헌

[1] Bertsekas, Dimitri P., Dynamic Programming and Optimal Control, Athena Scientific, Belmont, Massachusetts, Vol.1 and 2, 1995

[2] Gregory Kuhlmann and Peter Stone, Progress in Learning 3 vs. 2 Keepaway, In RoboCup-2003: Robot Soccer World Cup VII, 2004

[3] Junling Hu, Michael P. Wellman, Multiagent Reinforcement Learning:Theoretical Frmaework and an Algorithm, In Proceedings of the 15th International Conference on Machine Learning, pp.242-250, Madison, WI, USA, July 1998

[4] Peter Stone, Richard S. Sutton, and Gregory Kuhlmann, Reinforcement Learning for RoboCup-Soccer Keepaway, Adaptive Behavior, 2005

[5] Sutton, R.S., Barto, A.G. Reinforcement Learning: An Introduction. MIT Press, 1998