강화학습에 기초한 로봇 축구 에이전트의 동적 위치 결정

권기덕[0], 김인철
경기대학교 전자계산학과
kdkwon@hanmail.net[0], kic@kyonggi.ac.kr

Reinforcement Learning based Dynamic Positioning of Robot Soccer Agents

Ki-Duk Kwon[0] In-Cheol Kim
Department of Computer Science, Kyonggi University

요 약

강화학습은 한 에이전트가 자신이 놓여진 환경으로부터의 보상을 최대화할 수 있는 최적의 행동 전략을 학습하는 것이다. 따라서 강화학습은 입력(상태)과 출력(행동)의 쌍으로 명확한 훈련 예들이 제공되는 교사 학습과는 다르다. 특히 Q-학습과 같은 비 모델 기반(model-free)의 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. 하지만 실제 응용분야에서 Q-학습과 같은 강화학습이 겪는 최대의 문제는 큰 상태 공간을 갖는 문제의 경우에는 적절한 시간 내에 각 상태와 행동들에 대한 최적의 Q값에 수렴할 수 없어 효과를 거두기 어렵다는 점이다. 이런 문제점을 고려하여 본 논문에서는 로봇 축구 시뮬레이션 환경에서 각 선수 에이전트의 동적 위치 결정을 위해 효과적인 새로운 Q-학습 방법을 제안한다. 이 방법은 원래 문제의 상태공간을 몇 개의 작은 모듈들로 나누고 이들의 개별적인 Q-학습 결과를 단순히 결합하는 종래의 *모듈화 Q-학습(Modular Q-Learning)*을 개선하여, 보상에 끼친 각 모듈의 기여도에 따라 모듈들의 학습결과를 적응적으로 결합하는 방법이다. 이와 같은 *적응적 중재에 기초한 모듈화 Q-학습법(Adaptive Mediation based Modular Q-Learning, AMMQL)*은 종래의 모듈화 Q-학습법의 장점과 마찬가지로 큰 상태공간의 문제를 해결할 수 있을 뿐 아니라 보다 동적인 환경변화에 유연하게 적응하여 새로운 행동 전략을 학습할 수 있다는 장점을 추가로 가질 수 있다. 이러한 특성을 지닌 AMMQL 학습법은 로봇축구와 같이 끊임없이 실시간적으로 변화가 일어나는 다중 에이전트 환경에서 특히 높은 효과를 볼 수 있다. 본 논문에서는 AMMQL 학습방법의 개념을 소개하고, 로봇 축구 에이전트의 동적 위치 결정을 위한 학습에 어떻게 이 학습방법을 적용할 수 있는지 세부 설계를 제시한다.

1. 서 론

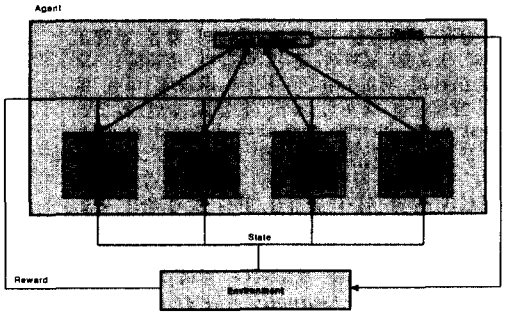
강화학습은 한 에이전트가 자신이 놓여진 환경으로부터의 보상(reward)을 최대화할 수 있는 최적의 행동 전략을 학습하는 것이다. 실제 응용분야에서 Q-학습과 같은 강화학습이 겪는 최대의 문제는 큰 상태 공간을 갖는 문제이며, 이 경우에는 적절한 시간 내에 각 상태와 행동들에 대한 최적의 Q값에 수렴할 수 없어 효과를 거두기 어렵다는 점이다. 이런 문제점에 대해 가장 효과적인 상태공간 축소방법 중의 하나로는 Whitehead에 의해 제안된 *모듈화 Q-학습법(Modular Q-Learning)*이다. 하지만 이 방법은 설계자에 의해 할당된 고정 모듈들을 이용하며, 이들을 결합하는 방식 또한 최대 질량 결합전략(greatest mass merging strategy)이라 불리는 단순하고 고정적인 방식을 사용함으로써 환경이 동적으로 변화하면 효과적으로 적응하기 어렵다는 단점이 있다. 본 논문에서는 종래의 모듈화 Q-학습법을 개선하여, 보상에 끼친 각 모듈의 기여도에 따라 모듈들의 학습결과를 적응적으로 결합하는 *적응적 중재에 기초한 모듈화 Q-학습법(Adaptive Mediation based Modular Q-Learning, AMMQL)*을 제안한다. 이 학습방법은 큰 상태공간의 문제를 해결할 수 있을 뿐 아니라 동적인 환경변화에 보다 높은 적응성을 제공할 수 있다. 본 논문에서는 AMMQL 학습방법의 개념을 소개하고, 로봇 축구 에이전트의 동적 위치 결정을 위한 학습에 어떻게 이 학습방법을 적용할 수 있는지 세부 설계를 제시한다.

2. 로봇 축구 시뮬레이션

로봇 축구, 특히 RoboCup 시뮬레이션 게임리그는 강화학습을 비롯한 다중 에이전트 학습을 위한 좋은 테스트 환경(testbed)을 제공한다. 이 로봇 축구 환경에서는 각 소프트웨어 에이전트들이 축구서버(soccer server)가 제공하는 가상의 필드(field)에서 경기를 해나간다. 즉, 각 선수 에이전트는 축구서버를 통해 자신의 인식 정보(perceptual information)를 받아들이고, 역시 축구서버를 통해 자신의 행동을 수행하게 된다. 일반적으로 이 환경이 갖는 어려운 점들은 다음과 같은 성질로 표현된다 : (1) 다중 에이전트(multiagent) : 각 팀은 11명의 선수로 구성되며, 팀 플레이와 같은 팀 동료간의 협조와 상대 팀과의 경쟁이 요구된다. (2) 불완전한 인식(incomplete perception) : 각 에이전트는 인식범위에 제한이 있으며, 축구서버는 오류가 포함된 센서 정보를 제공한다. (3) 실시간성(real-time) : 경기는 끊어지지 않고 계속되며, 각 에이전트는 이와 같은 연속 환경에서 실시간적으로 자신의 행동을 결정해야 한다. 또한 로봇 축구 환경에서 강화학습을 적용할 때는 다음과 같은 문제점들을 극복해야 한다 : (1) 동적 변화, (2) 큰 상태 공간, (3) 팀원들간의 보상 배분, (4) 부분적으로만 관찰 가능한 Markov 결정 프로세스(Partially Observable Markov Decision Process, POMDP), (5) 탐험(exploration)과 활용(exploitation)의 균형 등이다.

강화학습을 로봇 축구에 적용하는 기존 연구는 많이 찾아볼 수 있다. CMU의 CMUnited 팀 에이전트에서는 공 가로채기(ball interception)와 같은 기본적인 행위를 학습하는 하위 계층에서부터 팀 동료 중 누구에게 패스해야 할지를 학습하는 최상위 계층까지 총 3 개의 계층으로 구성된 계층적 학습(layered learning) 구조를 적용하였으며, 특히 이 중에서 패스선택을 위한 온라인 학습(online learning)에 강화학습을 적용하였다. Andou[1]의 Andhill 팀 에이전트에서는 자신이 공을 가지고 있지 않은 경우에 정해진 원래 위치에

단순히 머물러 있지 않고 동적으로 적절한 위치를 찾아 이동하는데 강화학습을 적용하였다. Andhill에서는 Kimura가 제안한 POMDP에 적용 가능한 Q-학습 형태인 SGA 알고리즘을 로봇축구 환경에 적용하였다. 이 알고리즘에서는 Q-학습과 더불어 대표적인 귀납적 학습(inductive learning) 방법인 신경망(neural network)을 함께 이용함으로써 큰 상태공간 문제에 도움을 주었다. KAIST의 로봇 축구 팀 NaroSot[4]에서는 효과적인 지역방어전략을 수행하기 위해 팀 동료들과 협조하는 방법을 모듈화 Q-학습법을 적용하여 학습하였다.



[그림 1] MQL 구조

3. 강화 학습

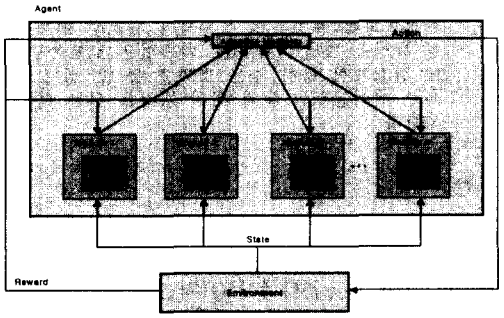
강화학습(reinforcement learning)은 한 에이전트가 자신이 놓여진 환경으로부터의 보상(reward)을 최대화할 수 있는 최적의 행동 전략을 학습하는 것이다. 따라서 강화학습은 입력(상태)과 출력(행동)의 쌍으로 명확한 훈련 예들이 제공되는 교사 학습(supervised learning)과는 다르다. 특히 Q-학습과 같은 비 모델 기반의 강화학습(model-free reinforcement learning)은 사전에 환경에 대한 별다른 모델을 설정하거나 학습할 필요가 없으며 다양한 상태와 행동들을 충분히 자주 경험할 수만 있으면 최적의 행동전략에 도달할 수 있어 다양한 응용분야에 적용되고 있다. 일반적인 Q-학습법에서 Q-값 갱신은 [식 1]과 같다.

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r+\gamma \max_{a'} Q(s',a')-Q(s,a)] \quad [식\ 1]$$

실제 응용분야에서 Q-학습과 같은 강화학습이 겪는 최대의 문제는 큰 상태 공간을 갖는 문제의 경우에 학습에 필요한 Q-표(Q-table)의 크기가 너무 커진다는 것이다. 따라서 적절한 시간 내에 가능한 모든 상태와 행동들에 대한 최적의 Q값을 계산해낼 수 없어 실효를 거두기 어렵다는 점이다. 이런 문제점에 대해 기존의 연구를 통해 다양한 해결책들이 모색되었다. 수렴속도를 높이는 한 가지 방안으로는 실제 경험보다는 환경에 대한 모델을 기초로 좀더 많은 Q 개선 연산을 수행하는 모델기반 강화학습법(model-based reinforcement learning)들이 제안되었고, 또한 상태공간을 줄여보려는 연구 중에는 상태공간 추상화(abstraction)에 대한 연구들도 있었다. 하지만 가장 효과적인 상태공간 축소 방법 중의 하나로는 Whitehead에 의해 처음 제안된 *모듈화 Q-학습법(Modular Q-Learning, MQL)*이다. 이 학습방법은 [그림 1]과 같이 원래 문제의 큰 상태공간에 대해 바로 Q-학습을 전개하지 않고 대신 이 상태공간을 몇 개의 작은 모듈들로 나눈 다음, 모듈별로 개별적인 Q-학습을 전개하고 이들의 Q-학습 결과를 결합함으로써 최적의 행동을 결정하는 방식을 취한다.

$$a^* \leftarrow \arg\max_{a \in A} \sum_{i=1}^{n} Q_i(s,a) \quad [식\ 2]$$

하지만 이 방법은 설계자에 의해 할당된 고정 모듈들을 이용하며, 이들을 결합하는 방식 또한 [식 2]와 같이 최대 질량 결합전략(greatest mass merging strategy)이라 불리는 단순한 고정적인 방식을 사용함으로써 환경이 급격하게 변화하면 효과적으로 적응하기 어렵다는 단점이 있다. 이에 반해, Kohri의 연구[3]에서는 변화하는 환경에 맞추어 에이전트 스스로가 적합한 모듈들의 집합을 만들어 가는 구조인 *자동화된 모듈화 학습법(Automatic Modular Q-Learning, AMQL)*을 제안했다. 이 방법은 3 단계로 자동 모듈 구성을 위한 학습이 진행되는데, 먼저 (1) 상태공간을 표현하는 환경요소(element)들의 부분집합을 임의로 선정하여 각 모듈들을 구성하고, (2) 모듈화 학습을 진행하면서 보상을 받을 수 있는 행동 결정에 기여한 정도에 따라 각 모듈의 적합성을 평가한 뒤, (3) 적합성이 정해진 임계치(threshold)보다 높은 모듈은 잔류하고, 그렇지 못한 모듈들은 새로운 환경요소들을 선택하여 재구성한다. 이 학습방법은 환경변화에 대응하는 에이전트에게 좀더 높은 적응성을 제공할 수 있으나, 자동으로 모듈들을 결정하기 위한 별도의 학습비용과 시간을 부가적으로 필요로 하여 로봇축구와 같이 비교적 시간이 짧은 환경에서는 빠른 효과를 기대하기 어렵다.



[그림 2] AMMQL 구조

4. 적응적 중재에 기초한 모듈화 Q-학습

본 논문에서는 순수 Q-학습이 갖는 큰 상태공간 문제를 해결하기 위해 모듈화 Q-학습(MQL)과 그리고 이것을 확장한 자동화된 모듈화 Q-학습(AMQL) 등과 유사한 접근방법을 제시한다. 하지만 이 방법은 AMQL과 같이 에이전트 스스로 내부 모듈들을 학습하도록 하는 방식 대신에 순수 MQL처럼 설계자가 영역지식을 바탕으로 설계해준 고정 모듈들을 사용하되 급격한 환경변화에 적응할 수 있도록 모듈들의 결합방식을 동적으로 결정한다. 따라서 *적응적 중재에 기초한 모듈화 Q-학습(Adaptive Mediation based Modular Q-Learning, AMMQL)*이라 불리는 이 방법은 부가 노력이 많이 필요하고 학습속도가 느린 AMQL보다는 낮은 수준의 적응성을 제공하나, 순수 MQL에 비해서는 높은 적응성을 적은 비용으로 제공할 수 있다는 것이 큰 특징이다.

[그림 2]는 n 개의 모듈로 구성된 AMMQL의 구조를 보여준다. 에이전트의 행동에 따른 환경으로부터의 보상/강화신호는 개별적인 Q-학습을 전개하는 각 모듈에 전달될 뿐 아니라 각 모듈의 학습결과를 결합하여 최종적으로 실행할 행동을 결정하는 중재모듈(mediation module)에도 전해진다. 그리고 중재모듈은 이와 같은 보상을 이끌어낸 행동을 결정하는데 각 모듈들이 어느 정도 기여하였는지에 따라

행동 결정 함수에 각 모듈의 Q-값을 반영하는 비중을 조정한다. 따라서 AMMQL은 중재모듈에서도 학습이 이루어지는 이중적 학습구조를 취한다. 모듈화 이전 상태공간상의 한 상태 $s_k \in S$에 대응되는 각 모듈의 상태를 각각 s_{1k}, s_{2k}, ..., s_{nk}이라고 하면 중재모듈의 행동결정함수는 각 모듈의 Q-값을 가중치 ω_i에 따라 선형적으로 결합하는 [식 3]과 같다.

$$a^* = \arg\max_{a \in A}(\omega_1 Q_1(s_{1k}, a) + ... + \omega_n Q_n(s_{nk}, a))$$

$$= \arg\max_{a \in A} \sum_{i=1}^{n} \omega_i Q(s_{ik}, a) \quad \text{[식 3]}$$

그리고 이때 가중치 ω_i들의 합은 1로 한다. 즉, $\omega_1 + \omega_2 + ... + \omega_n = 1$.

중재모듈에 의해 선택된 최적 행동 a^*을 수행한 결과 환경으로부터 보상값 R이 주어졌다고 가정하자. 또한 각 모듈 M_i별로 가능한 모든 행동들의 Q-값의 합에 대한 행동 a^*의 Q-값의 비율 r_i을 [식 4]과 같이 계산한다고 가정하자.

$$r_i = \frac{Q_i(s_{ik}, a^*)}{\sum_{a_m \in A} Q_i(s_{ik}, a_m)} \quad \text{[식 4]}$$

그러면 각 모듈 M_i의 임시 가중치 ω_i'는 [식 5]와 같이 계산되며, 이때 계수 β는 학습율(learning rate)을 나타낸다. 그리고 이 임시 가중치들은 [식 6]와 같은 정규화(normalization)를 거쳐 새로운 가중치 ω_i로 갱신된다.

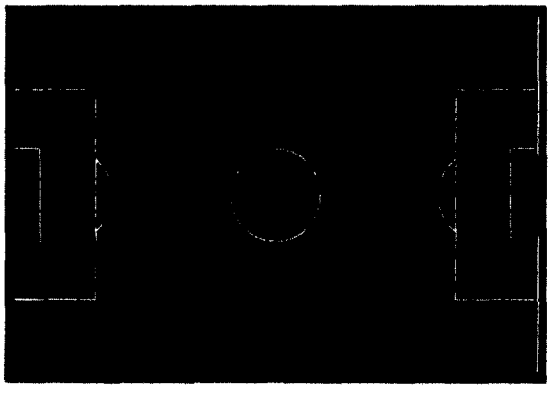
$$\omega_i' \leftarrow \omega_i + \beta \cdot r_i \cdot R \quad \text{[식 5]}$$

$$\omega_i = \frac{\omega_i'}{\omega_1' + \omega_2' + ... + \omega_n'} \quad \text{[식 6]}$$

AMMQL은 이와 같은 과정을 반복함으로써 큰 추가비용을 들이지 않고도 환경변화에 더욱 민감한 모듈화 Q-학습을 진행해갈 수 있다.

5. 구 현

본 논문에서는 로봇 축구 시뮬레이션 환경에서 한 선수 에이전트가 공을 가지고 있는 팀 동료와 협조하여 공격하기에 적합한 위치를 결정하는데 AMMQL 학습법을 적용하였다. [그림 3]은 팀 동료로부터 공을 패스받아 슈팅하기에 좋은 곳으로 이동하는 한 선수 에이전트의 상황을 보여 준다.



[그림 3] 위치 결정 및 이동의 예

위치 결정을 위해 고려해야 할 환경요소로는 (1) 공의 위치를 중심으로 가장 가까운 팀 동료의 상대 위치(relative position), (2) 공의 위치를 중심으로 가장 가까운 적의 상대 위치, (3) 공의 위치를 중심으로 자신의 상대 위치, (4) 목적지를 중심으로 가장 가까운 팀 동료의 상대 위치, (5) 목적지를 중심으로 가장 가까운 적의 상대 위치, (6) 목적지를 중심으로 자신의 상대 위치 등이 있으며, 따라서 경기장 필드를 16X11의 격자로 나누어 위치를 표시하는 경우 순수 Q-학습법을 위한 상태공간 S의 크기는 $(16\times11)^6 = 29721861554176$이나 된다. 본 논문에서는 모듈 당 하나의 환경요소만을 포함하도록 함으로써 각각의 상태공간 크기가 단지 $(16\times11) = 176$인 총 6 개의 모듈로 나누었다. 에이전트가 선택할 수 있는 (위치 이동) 행동(action)은 모두 목적지 위치로 대신 표현할 수 있으므로 각 모듈의 행동공간 A의 크기는 모두 $(16\times11) = 176$로 동일하다. 환경으로부터 주어지는 보상값 R은 공의 위치 변동을 고려하는 [식 7-1]과 [식 7-2]에 의해 계산되는 것으로 정하였다. [식 7-1]은 공이 경기장 안에 있을 경우이고 [식 7-2]는 공이 경기장 밖으로 나갔을 경우이다.

$$R = \frac{R_0}{1 + (\phi - 1) * t_0 / t_{\lim}} \quad \text{[식 7-1]}$$

$$R = \begin{cases} \phi * \dfrac{x_{avg} - x_t}{x_{og} - x_t} (x_{avg} > x_t) \\ -\phi * \dfrac{x_t - x_{avg}}{x_t - x_{lg}} (x_{avg} < x_t) \end{cases} \quad \text{[식 7-2]}$$

각 모듈의 Q-학습을 위한 학습율은 $\beta = 0.1$, 감퇴요소(discount factor)는 $\gamma = 0.9$, 초기 Q-값은 모두 0.1로 하였으며, 중재모듈의 학습율은 $\alpha = 0.1$, 초기 가중치는 모두 동일하게 $\omega_i = 1/6 \approx 0.16$으로 주었다.

6. 결론

본 논문에서는 로봇 축구 시뮬레이션 환경에서 각 선수 에이전트의 동적 위치 결정을 위해 새로운 모듈화 Q-학습 방법인 AMMQL을 제안하였다. 이 학습방법은 큰 상태공간의 문제를 해결할 수 있을 뿐 아니라, 적응적 중재전략을 통해 순수 모듈화 Q-학습에 비해 환경변화에 보다 높은 적응성을 제공한다. 비교적 복잡도가 큰 로봇 축구영역에서의 비교 실험을 통해 AMMQL의 유연성을 입증하는 것이 향후 연구로 남아있다.

7. 참고문헌

[1] Andou, T, "Refinement of soccer agents positions using reinforcement learning.", Robocup-97 : Robot Soccer World Cup I, pp. 373-388, 1998.

[2] Kaebling L. P, Littman M. L, Moore A.W, "Reinforcement learning : A Survey", Journal of AI Research, Vol. 4, pp.237-285, 1996.

[3] Takayuki Kohri et al, "An Adaptive Architecture for Modular Q-Learning", Proceedings of IJCAI-97, 1997.

[4] J.H. Kim, "Modular Q-learning based multi-agent cooperation for robot soccer", Robotics and Autonomous Systems, Vol. 35, pp.109-122, 2001