# 적대적 멀티 에이전트 환경에서 효율적인 강화 학습을 위한 정책 모델링

권기덕○ 김인철
경기대학교 전자계산학과
kdkwon@kyonggi.ac.kr, kic@kyonggi.ac.kr

# Policy Modeling for Efficient Reinforcement Learning in Adversarial Multi-Agent Environments

Kiduk Kwon○ Incheol Kim
Department Computer Science, Kyonggi University

멀티 에이전트 시스템에 강화 학습을 적용해보려는 멀티 에이전트 강화 학습(multiagent reinforcement learning)에 관한 연구가 최근 들어 관심을 모으고 있다. 멀티 에이전트 강화 학습에서 해결해야 할 중요한 문제는 자신의 작업 성능에 영향을 미칠 수 있는 다른 에이전트들이 존재하는 동적 환경에서 한 에이전트가 시행착오적 상호작용을 통해 어떻게 자신의 최적 행동 정책을 학습할 수 있느냐 하는 것이다. 멀티 에이전트 강화 학습이 적용될 수 있는 분야로는 로봇 축구[1], 컴퓨터 게임[2], 전자 상거래[3], 분산 웹 검색, 자가-관리 컴퓨터시스템, 반-테러 응용시스템 등 다양한 응용분야들이 존재한다. 전통적인 강화 학습방법들은 하나의 마코프 결정 문제(Markov Decision Problem, MDP)로 정의되는 단일 에이전트 환경에서 개발되어왔다. MDP에서 환경 상태 전이는 시간이 흘러도 변하지 않는 하나의 전이 확률 함수(transition probability function)로 정의된다. 그러나 에이전트들이 자율적으로 학습할 수 있는 멀티 에이전트 환경을 생각한다면, 각 에이전트의 행위는 학습이 진행됨에 따라 변화하게 된다. 따라서 학습이 가능한 다수의 에이전트들이 공존하는 멀티 에이전트 환경에서는 시간이 경과함에 따라 상태 전이 함수도 변할 수 있다. 즉, 그러한 환경은 하나의 MDP로 정의할 수 없다. 그럼에도 불구하고 그동안 많은 멀티 에이전트 강화 학습연구들에서는 MDP 기반의 강화 학습기법들이 큰 변화 없이 그대로 적용되어 왔다[4]. 그동안 다른 에이전트의 존재를 명시적으로 고려하는 멀티 에이전트 강화 학습 연구들도 있었다. 대표적인 연구들로는 두 명의 제로-합 게임을 위한 Littman의 Minimax-Q 학습 알고리즘[5], 두 명의 일반-합 게임을 위한 Hu와 Wellman의 Nash-Q 학습 알고리즘[6], Minimax-Q를 일반-합 게임으로 확장한 Littman의 FFQ 학습 알고리즘[7] 등이 있다. 이들은 대부분 적용 가능한 멀티 에이전트 시스템의 유형이 제한적이거나 다른 에이전트에 관해 요구되는 정보나 가정(assumption)이 비현실적이라는 한계점을 가지고 있다.

본 논문에서는 과거에 관찰된 상대 에이전트의 행동들을 기초로 상대 에이전트의 행동 정책 모델을 학습하고, 이 모델을 바탕으로 다시 자신의 최적 정책을 학습하는 강화 학습방법을 제시한다. 이 멀티 에이전트 강화 학습방법은 두 명의 에이전트로 구성된 적대적 멀티 에이전트 환경을 가정하고 있으며, 두 에이전트는 동시에 행동을 수행함으로써 자신의 행동을 결정하기 전에 미리 상대 에이전트의 행동을 알 수는 없으나 일단 동시에 행동을 수행하고 나면 상대 에이전트가 수행한 행동을 관찰할 수 있다. 하지만 두 에이전트 간에는 행동 결정에 영향을 미치는 어떠한 통신도 가능하지 않다고 가정한다. Q 학습 알고리즘을 확장한 이 멀티 에이전트 강화학습 방법은 상대 모델을 이용하는 기존의 멀티 에이전트 강화 학습 연구들에서 주로 시도되었던 상대 에이전트의 Q 평가 함수 모델 대신 상대 에이전트의 행동 정책 모델을 학습하며, 표현력은 풍부하나 학습에 시간과 노력이 많이 요구되는 유한 상태 오토마타(DFA)나 마코프 체인(Markov Chain)과 같은 행동 정책 모델들에 비해 비교적 간단한 형태의 행동 정책 모델을 이용함으로써 학습의 효율성을 높였다.

본 논문에서 제안하는 상대방 정책 모델 $PM(s, a_{opponent})$은 상태 s에서 상대 에이전트가 행동 $a_{opponent}$을 수행할 가능성을 0과 1사이의 값으로 추정한 것이다. 그리고 이러한 상대방 정책 모델 $PM(s, a_{opponent})$은 상대 에이전트의 실제 행동을 관찰함에 따라 적응적으로 조정된다. 즉, 상태 s에서 상대 에이전트가 행동 $a^*_{opponent}$을 수행하는 것을 관찰하면, 상태 s에서 수행 가능한 모든 행동 $a_{opponent}$에 대한 상대방 정책 모델 $PM(s, a_{opponent})$은 [식 1]과 같이 갱신된다.

$$PM(s, a_{opponent}) = (1-\Theta) PM(s, a_{opponent}) + \begin{cases} \Theta & (a_{opponent} = a^*_{opponent}) \\ 0 & (otherwise) \end{cases} \quad [식\ 1]$$

이때, $\Theta(0 \le \Theta < 1)$는 실제 상대 에이전트가 수행한 행동인 $a^*_{opponent}$의 효과를 조절하는 파라미터(parameter)이다. 임의의 한 상태 $s$에서 자주 반복 관찰되는 상대 에이전트의 행동에 대해서는 $PM$ 함수 값이 증가하고, 그렇지 못한 행동에 대해서는 $PM$ 함수 값이 감소한다. 상대방 정책 모델을 이용한 멀티 에이전트 Q 학습은 다음과 같은 절차대로 진행된다.

**단계 1:** 에이전트는 현재 상태 $s$와 상대방 정책 모델 $PM(s, a_{opponent})$을 기초로 상대 에이전트가 수행할 행동 $a_{opponent}$들을 예측한다. 그리고 [식 2]와 같이 계산되는 자신의 확률 정책에 따라 자신이 취할 행동 $a_{self}$을 결정한다.

$$\pi[a_i \mid s] = \frac{e^{\overline{Q}(s, a_i)/\tau}}{\sum_{a_j \in A_{self}} e^{\overline{Q}(s, a_j)/\tau}} \quad [식\ 2]$$

이때, $\overline{Q}(s, a_{self}) \equiv \sum_{a_{opponent} \in A_{opponent}} PM(s, a_{opponent}) Q(s, a_{self}, a_{opponent})$는 상대방 정책 모델을 토대로 계산된 자신의 행동 $a_{self}$에 대한 Q-함수 기대 값이다.

**단계 2:** 에이전트는 단계 1에서 선택된 자신의 행동 $a_{self}$을 실행한다. 이와 동시에 상대 에이전트도 행동 $a^*_{opponent}$을 선택하고 실행한다. 그 결과, 환경은 새로운 상태 s'로 변경되고 에이전트는 환경으로부터 보상 값 $r$을 받는다. 그러면 에이전트는 [식 1]에 따라 상대방 정책 모델을 갱신하고, 또 [식 3]과 같이 Q 함수 값도 갱신한다.

$$Q(s, a_{self}, a^*_{opponent}) \leftarrow (1-\alpha) Q(s, a_{self}, a^*_{opponent}) + \alpha (r + \gamma \max_{a'_{self} \in A_{self}} Q(s', a'_{self}, a'_{opponent})) \quad [식\ 3]$$

여기서 $a'_{opponent} = \arg\neg @_{a'_{opponent} \in A_{opponent}} PM(s', a'_{opponent})$이다.

**단계 3:** 만약 새로운 상태 $s'$가 종결 조건을 만족하면, 게임종료와 더불어 학습을 위한 한 번의 에피소드가 끝나게 된다. 그렇지 않으면 $s' \rightarrow s$ 로 상태를 변경한 다음, 단계 1부터 다시 반복한다.

본 논문에서는 대표적인 적대적 멀티 에이전트 환경인 고양이와 쥐(Cat and Mouse) 게임을 소개하고, 이 게임을 테스베드삼아 수행한 비교 실험 결과들을 설명함으로써 본 논문에서 제안하는 정책 모델 기반의 멀티 에이전트 강화 학습의 효과를 분석해본다.

## 참고 문헌


[1] Yang E. and Gu D., "Multiagent Reinforcement Learning for Multi-Robot Systems: A Survey," University of Essex Technical Report CSM-404, 2004

[2] Tesauro G., "Multi Agent Learning: Mini Tutorial," IBM T.J.Watson Research Center, 2000

[3] Rahimi K.A., Tabarraei H., Sadeghi B., "Reinforcement Learning Based Supplier-Agents for Electricity Markets," Proceedings of the IEEE International Symposium on Control and Automation, pp.1405-1410, 2005

[4] Shoham Y., Powers R., and Grenager T., "Multi-Agent Reinforcement Learning: A Critical Survey," Technical Report, Stanford University, 2003

[5] Littman M.L., "Markov Games as Framework for Multi-Agent Reinforcement Learning," Proceedings of the 11th International Conference on Machine Learning, pp. 157-163, 1994

[6] Hu J. and Wellman M.P., "Nash Q-learning for General-Sum Stochastic Games," Journal of Machine Learning Research, vol. 4, pp. 1039-1069, 2003

[7] Littman M.L., "Friend-or-Foe Q-learning in General-Sum Games," Proceedings of the 18th International Conference on Machine Learning, Morgan Kaufman, pp. 322-328, 2001