대규모 가상군의 POMDP 행동계획 및 학습 사례연구*

홍정표 O1 이종민 1 이강훈 1 한상규 1 김기응 1 문일철 2 박재현 3
{1 전산학부, 2 산업 및 시스템 공학과}, 한국과학기술원, 2 국방과학연구소
{jphong, jmlee, khlee, sghan}@ai.kaist.ac.kr, kekim@cs.kaist.ac.kr,
icmoon@kaist.ac.kr, forehand@add.re.kr

A Case Study on Planning and Learning for Large-Scale CGFs with POMDPs

Jungpyo Hong O1 Jongmin Lee1 Kanghoon Lee1 Sanggyu Han1 Kee-Eung Kim1
Il-Chul Moon2 Jae-Hyeon Park3
{1School of Computing, 2Department of Industrial and Systems Engineering}, KAIST,
3Agency for Defense Development

요 약

대규모 가상군의 전투 모델링 및 시뮬레이션은 향후 발생할 전투의 작전을 고도화하고 효율적인 모의 훈련을 가능하게 한다. 이를 위해, DEVS-POMDP 계층적 프레임워크에서는 전투 행동 교범과 그에 따른 구체적인 행동계획을 각각 DEVS와 POMDP로 모델링하여 가상군의 자율적인 행동을 모의하였으나, POMDP 모델에서 최적 행동정책을 계산하는 것은 여전히 많은 컴퓨팅 자원를 필요로 한다. 본 논문에서는 DEVS-POMDP로 모델링된 연평도 대화력전 모의 시나리오의 사례연구를 통해 효율적인 POMDP 트리 탐색 알고리즘 및 적군 행동 양상 모델의 학습을 통한 가상군 전투 개체의 성능 향상을 확인한다.

1. 서론

컴퓨터 모의 가상군 (Computer Generated Forces; CGF)을 이용한 전투모의 모델링 및 시뮬레이션은 국방 분야의 무기체계와 작전 등의 설계, 분석 및 평가에 있어 중요한 기술이다[1,2]. 특히, 이성적이고 지능적인 가상군 전투 개체의 행동 묘사는 지휘관들의 효율적인 전투작전 설정 및 모의훈련을 가능케 한다. 군 전문가가 작성한 전투 행동 교범 (Field Manual)은 전투 개체의 행동을 명시적으로 서술하고 있지만, 전장에서 일어나는 모든 상황을 명시하는 것이 불가능하고, 또한 구체적인 서술이 힘든 행동들이 많다는 단점이 존재한다.

명시적으로 서술될 수 없거나, 서술된 행동 이상의 지능적인 행동을 보여야 할 경우, 전투 개체는 인공지능 의사결정 이론에 기반하여 자율적으로 최적의 행동 정책을 계산하여야 한다. 부분관찰 마코프 의사결정과정 (Partially Observable Markov Decision Process; POMDP)[3]은 개체가 행동하는 주변환경의 불확실성을 확률적으로 모델링하여 순차적인 행동 정책을 계산하는 기계학습 방법론으로, 산업공학, 인지과학 및 인공지능 분야에서 사용되는 의사결정 프레임워크이다. 가상군을 POMD로 모델링 함으로써 명시적으로 서술된 교범 없이, 주변 환경에 대한 확률 모델만으로 지능적인 행동 정책의 계산이 가능하지만[1], POMDP 행동 정책 최적화 알고리즘 (Planning)의 높은 계산 복잡도로 인해 행동 정책 계산이 어렵고, 전투시 지켜야 할 교범을 고려하지 못했다. 따라서, 최근 대규모 가상군 모의 시 교범에 따른 상위 단계 의사결정을 이산 사건 시스템 명세 (Discrete Event System Specification; DEVS)[4]를 이용하고, 교범에 구체적인 서술이 힘든 하위 단계 자율 의사결정은 POMDP를 이용하는 계층적 모델링 기법인 DEVS-POMDP 프레임워크가 제안되었다[2].

본 논문에서는 연평도 일대 대화력전 모의 시나리오 사례연구를 통해 DEVS-POMDP 기반 대규모 가상군의 실시간 전투를 모의한다. 특히, 복잡한 환경 상태를 가지는 POMDP 문제를 효율적인 탐색 알고리즘을 통해 실시간으로 계산하고, 적군 행동 모델에 대한 불확실성을 경험을 통해 스스로 학습 (Learning)함으로써, 가상군 전투 개체의 성능을 향상시킬 수 있음을 보인다.

2. 연평도 대화력전 모의 시나리오

그림 1 은 연평도 일대에서 적군의 포격 도발 상황을 모의한 전투 시나리오이다. 적군은 지도의 북쪽 개머리 진지에서 6개 포대 규모의 갱도포를 통해 포격 도발을 하고, 아군은 지도 남쪽 연평도에서 2개 포대 규모의 자주포로 대응 포격을 하게 된다. 각각의 포대는 6문의 갱도포 및 자주포로 구성되어 있으며, 전투 행동 교범에 따라 같은 포대에 속한 포들은 같은 행동을 수행하게 된다. 이 때, 아군 포대는 적군 포대에 비해 수적 열세인 상황에서 최대한 적은 피해로 많은 수의 적군을 포격해 공격 불능 상태로 만드는 것이 목표가 된다.

* 본 연구는 방위사업청과 국방과학연구소의 지원으로 수행되었습니다(UD140022PD).

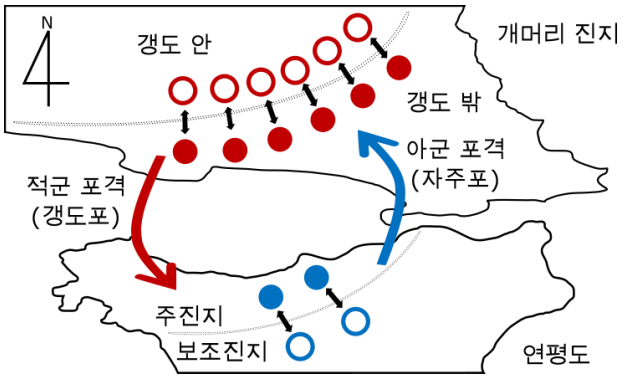

그림 1. 연평도 대화력전 모의 시나리오

모의 시나리오는 적군의 최초 포격을 시작으로, 아군 연대 작전통제소의 지시 하에, 포격한 적군 갱도포대의 표적 획득 및 자주포를 통한 대응 포격을 진행한다. 포격 후, 아군은 주진지와 보조진지 사이를 이동하며 적군이 획득한 표적의 위치를 벗어나야 하고, 적군은 갱도 밖에서 포격 후 갱도 안으로 이동하여 아군의 포격을 방어함과 동시에 다음 사격에 대한 준비를 한다. 이 때, 갱도 밖에서 포격한 적군은 75%의 확률로 갱도 안으로 이동 및 방어 후 다시 갱도 밖으로 나가게 된다.

아군은 적군 행동 양상(즉, 갱도 안팎 이동 확률)을 고려하여 대응 포격 정책을 계획 (Planning)해야 한다. 또한, 적군의 행동 양상에 대한 정보가 불확실할 경우, 적군의 행동 모델에 대한 학습을 동시에 고려할 수 있는 강화학습 (Reinforcement Learning) 알고리즘이 필요하다. 본 논문에서는 적군 행동 양상을 알고 있다고 가정할 때 최적의 포격 정책을 계획하는 방법론과 함께, 적군 행동 양상에 대한 불확실성이 존재할 경우 필요한 강화학습 방법론을 제시한다.

3. POMDP 모델링을 통한 행동계획 및 학습 알고리즘

연평도 대화력전 모의 시나리오의 전투 시뮬레이션은 DEVS를 통해 모의된다. 이 때, 아군 자주포대의 자율적 포격 행동 전략은 POMDP로 계산되고, 이는 DEVS-POMDP 계층적 모델링[2]을 통해 구현가능하다. 다음의 내용에서는 아군 자율 포격 행동의 성능 향상을 위한 전투 개체의 POMDP 모델링, 최적 행동계획 및 학습 알고리즘을 제안한다.

3.1. POMDP 모델링 및 행동계획 알고리즘

POMDP는 <환경 상태, 행동, 관찰값, 상태 전이 확률, 관찰 확률, 보상 함수>로 구성된다. 연평도 대화력전 모의 시나리오의 아군 자주포대 POMDP 모델링은 표 1과 같다. POMDP는 환경 상태를 직접적으로 알 수 없고, 대신 관찰값으로부터 환경 상태들에 대한 확률 분포인 Belief-State를 계산하여 표현한다. 따라서, POMDP의 행동계획 문제란, 현재 Belief-State로부터 미래에 얻게 될 보상값의 합을 최대화하는 순차적 행동들을 효율적으로 계산하는 것이다.

표 1. 연평도 대화력전 모의 시나리오의 POMDP 모델링

환경 상태 (State)	아군 및 적군의 포가 위치한 진지 및 포격, 기동 가능 유무, 보유한 탄약 개수(0~40). (총 가능한 환경 상태의 수 ≈ 10^{34})
행동 (Action)	아군 포대 별 타격할 적군 진지 및 소모할 탄약 수 (0~3)와 이동 대상 진지. (총 가능한 행동의 수 = 2401)
관찰값 (Observation)	아군의 환경 상태와 관측된 적군 포격 유무 및 발사 위치. (총 가능한 관찰값의 수 ≈ 10^{11})
상태 전이 확률 (Transition Probability)	아군 및 적군 포대 이동에 대한 위치 전이. 적군이 아군이 위치한 진지를 포격 시, 70%로 아군 포대의 포격 불능 또는 기동 불능 상태 전이. 적군은 75%로 갱도 안 이동.
관찰 확률 (Observation Probability)	아군 포대 정보는 직접적으로 관찰 가능. 적군 포대 정보는 30%로 관측 실패.
보상 함수 (Reward Function)	공격 불능 상태의 적군 수 + 공격 가능 상태의 아군 수 - 아군 탄약 소모량

POMDP 최적 행동계획을 위해서 Monte-Carlo 트리 탐색 기법을 이용한 효율적 POMDP 탐색 알고리즘인 POMCP (Partially Observable Monte-Carlo Planning) [5] 알고리즘을 적용하였다. POMCP는 복잡한 환경 상태를 가진 문제에서 근사화된 Belief-State를 빠르게 계산할 수 있는 Particle Filtering 기법과 Monte-Carlo 시뮬레이션을 이용한 효율적인 트리 탐색을 적용해 최적의 행동 정책을 실시간으로 빠르게 찾을 수 있다.

3.2. POMDP 모델의 베이지안 (Bayesian) 학습

앞서 POMCP 알고리즘을 이용한 연평도 대화력전 모의 시나리오의 최적 행동계획은 적군의 갱도 안 이동 확률 모델을 아군이 알고 있을 경우를 가정하였다. 하지만, 실제 전투시 적군의 정확한 확률 모델에는 불확실성이 존재하므로, 경험을 통한 강화학습으로 확률 모델의 불확실성을 줄여야 한다.

POMDP 모델의 효과적 학습과 전투 목표의 달성을 동시에 고려하는 방법론으로 베이지안 강화학습이 있다. 베이지안 강화학습은 모델의 불확실성을 환경 상태에 포함시켜, 동일한 문제를 증강 환경 상태 (Augmented State)를 가지는 행동계획 문제로 변환시켜 풀 수 있다. 이러한 베이지안 모델은 모델 학습과 전투 목표 달성의 두가지 행동 균형에 대해 베이지안 입장에서 최적 해를 줄 수 있다. 이에 따라, 연평도 시나리오 POMDP 모델의 강화학습 문제는 Bayes-Adaptive POMDP 모델[6]의 최적 행동계획 문제로 새롭게 모델링하여 앞서 사용한 POMCP 알고리즘을 적용하였다. 베이지안 강화학습 모델링 및 트리 탐색을 이용한 최적 강화학습 행동계획 문제의 적용은 POMCP 알고리즘의 베이지안 강화학습 확장 알고리즘인 BAMCP (Bayes-Adaptive Monte-Carlo Planning)[7]를 참고하였다.

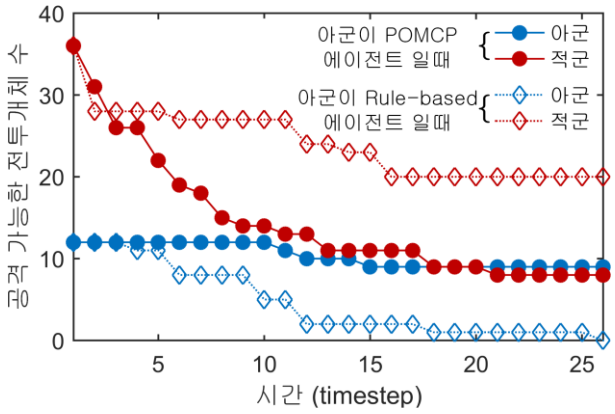

그림 2. 규칙 기반 및 POMDP 행동계획 성능 비교

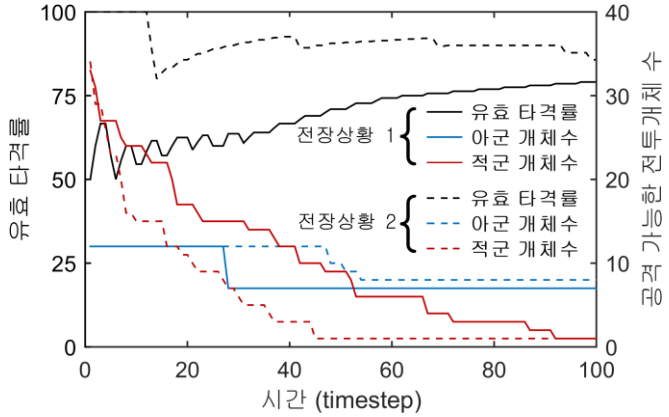

그림 3. 서로 다른 환경 모델의 베이지안 강화학습

4. 연평도 대화력전 모의 시나리오 실험 결과

그림 2 는 연평도 대화력전 모의 시나리오에서 적군 행동에 대한 모델을 알고 있을 경우, 아군의 행동계획에 대한 실험 결과이다. POMDP 모델링을 통한 행동계획과 비교를 위해, 포격 후 진지 이동을 반복하는 규칙을 (Rule-Based) 가지는 모델과의 비교 실험을 수행하였다. 그림에서 공격 가능한 아군 자주포 수는 파란색으로, 공격 가능한 적군 갱도포 수는 빨간색으로 표현하였고, POMDP로 모델링한 아군의 행동계획 결과는 실선으로, 규칙을 기반으로 행동한 결과는 점선으로 표현하였다.

시나리오에서 아군은 수적 열세인 상황이므로 규칙 기반 행동은 25번의 교전 후 아군이 전멸하는 결과를 보이지만, 적군의 행동 양상을 고려한 POMDP 모델의 최적 행동계획은 아군 피해를 최소화하고 적군 포대를 공격 불능 상태로 만든다. 특히, POMDP 모델링을 통한 아군 행동은 협업을 통해, 한 포대가 적군 공격을 유인하는 동안 다른 포대가 적군을 포격하는 전략을 계산함으로써 아군의 피해를 최소화하였다.

그림 3 은 적군 행동 양상의 불확실성이 있을 때, 베이지안 강화학습을 통해 최적 행동 정책을 계산한 결과이다. 실선으로 표시된 [전장상황1]에선 적군 행동 양상으로 90%의 확률로 갱도 안에서 두 차례 휴식기를, 10%의 확률로 즉시 갱도 밖으로 나온다. 점선으로 표시된 [전장상황2]는 반대로 10%의 확률로 갱도 안에 있고, 90%의 확률로 갱도 밖으로 나오게 된다. 아군은 적군의 두 행동 양상에 대해 사전 정보 없이 시작하여 (즉, [전장상황1]일 확률 0.5, [전장상황2]일 확률 0.5), 실제 적군의 행동 양상이 어떤 것인지에 대한 학습과 동시에 적군 포대에 대한 효과적인 포격을 수행한다. 그림에서 검은색으로 표현된 왼쪽 y축은 포탄의 활용 척도를 나타내기 위한 유효 타격률로

$$(\text{유효타격률}) = \frac{(\text{적군이 갱도 밖에 나와 있을 때 포격한 횟수})}{(\text{전체 포격 횟수})}$$

로 정의되고, 파란색 및 빨간색으로 표현된 오른쪽 y축은 아군 및 적군의 공격 가능한 포의 수를 나타낸다.

검은 실선으로 표현된 [전장상황1]의 유효 타격률은 적군이 높은 확률로 갱도 안에 있기 때문에 초반 유효 타격률이 낮지만, 적군이 갱도 밖으로 나올 확률을 학습함에 따라, 유효 타격률이 점점 높아짐을 알 수 있다.

5. 결론 및 향후 연구계획

본 논문은 인공지능 의사결정 프레임워크 POMDP 모델링의 행동계획 및 베이지안 강화학습의 효과를 연평도 대화력전 모의 시나리오 사례연구를 통해 보여주었다. POMDP 행동계획 알고리즘을 이용한 전투개체는 적군의 행동 양상을 효과적으로 고려하여 전투 시뮬레이션의 성능을 향상시켰고, 또한 적군 행동 양상에 대한 불확실성이 존재하는 전장 상황에서는 적군의 행동 양상을 학습함으로써 전투의 불확실성에 강인하게 행동할 수 있음을 알 수 있었다. 향후 연구로 현재 수행된 포대급 모의 시나리오를 확장하여 연대급 전투 시나리오의 효과적 자율 행동계획과 학습을 위한 POMDP 모델링 및 알고리즘 연구가 필요하다.

참 고 문 헌

[1] K. Lee et al., A Case Study on Modeling Computer Generated Forces based on Factored POMDPs, In Proc. of Korea Computer Congress, 2012

[2] J. Bae et al., Modeling Combat Entity with POMDP and DEVS, Journal of the Korean Institute of Industrial Engineers, 2013

[3] E. J. Sondik, The optimal control of partially observable Markov processes, Ph.D. thesis, Stanford University ,1971

[4] B. P. Zeigler et al., Theory of modeling and simulation: integrating discrete event and continuous complex dynamic systems. Academic press, 2000

[5] D. Silver et al., Monte-Carlo planning in large POMDPs, Advances in Neural Information Processing Systems, 2010

[6] S. Ross et al., Bayes-Adaptive POMDPs, Advances in Neural Information Processing Systems, 2007

[7] A. Guez et al., Efficient Bayes-Adaptive Reinforcement Learning using Sample-Based Search, Advances in Neural Information Processing Systems, 2012