Journal of the KIMST, Vol. 23, No. 4, pp. 337-345, 2020
ISSN 1598-9127(Print) · 2636-0640(Online)
DOI https://doi.org/10.9766/KIMST.2020.23.4.337

Research Paper 센서·신호처리 부문

평균 필드 게임 기반의 강화학습을 통한 무기-표적 할당

신민규*,1) · 박순서1) · 이단일1) · 최한림1)

1) 한국과학기술원 항공우주공학과

Mean Field Game based Reinforcement Learning for Weapon-Target Assignment

Min Kyu Shin*,1) · Soon-Seo Park1) · Daniel Lee1) · Han-Lim Choi1)

1) *Department of Aerospace Engineering, Korea Advanced Institute of Science and Technology, Korea*

(Received 14 April 2020 / Revised 10 June 2020 / Accepted 26 June 2020)

Abstract

The Weapon-Target Assignment(WTA) problem can be formulated as an optimization problem that minimize the threat of targets. Existing methods consider the trade-off between optimality and execution time to meet the various mission objectives. We propose a multi-agent reinforcement learning algorithm for WTA based on mean field game to solve the problem in real-time with nearly optimal accuracy. Mean field game is a recent method introduced to relieve the curse of dimensionality in multi-agent learning algorithm. In addition, previous reinforcement learning models for WTA generally do not consider weapon interference, which may be critical in real world operations. Therefore, we modify the reward function to discourage the crossing of weapon trajectories. The feasibility of the proposed method was verified through simulation of a WTA problem with multiple targets in realtime and the proposed algorithm can assign the weapons to all targets without crossing trajectories of weapons.

Key Words : Weapon-Target Assignment Problem(무기-표적 할당 문제), Multi-Agent Reinforcement Learning(멀티 에이전트 강화학습), Mean Field Game(평균 필드 게임)

1. 서 론

유도탄 운용기술의 발달로 교전 중 운용 가능한 유도탄 수가 계속 증가하고 있으며, 다수의 유도탄으로 좁은 영역을 집중공격하는 전략적 운용이 가능해졌다. 방어 요격체계 입장에서는 이러한 공격 형태에 대하여 요격탄을 효율적으로 할당하기 위해 추가로 고려해야 할 제약 조건이 증가하기 때문에 해법의 복잡도 증가한다. 따라서 다양한 제약 조건하에서 최적의 할당 해를 얻기 위해 많은 연구가 수행되고 있다[1-6].

무기-표적 할당 문제(Weapon-Target Assignment, WTA)는 조합 최적화 문제의 분류 중 하나로 방어용 무기 자원을 아군에게 공격적인 표적에 할당하여 아군 자

* Corresponding author, E-mail: mkshin@lics.kaist.ac.kr

산의 피해량을 최소화하거나, 혹은 아군의 무기 자원을 적군의 자산에 할당하여 적군의 피해량을 극대화하는 데에 그 목적이 있다. 일반적으로 무기-표적 할당 문제는 조합 최적화 문제로 NP-완성 문제 형태로 정의되며, 표 최적해를 구하기 위한 계산 복잡도는 무기 자원과 표적의 개수가 증가함에 따라 지수적으로 급격히 증가하게 된다[1]. 이러한 표적-무기할당 문제의 시간 복잡도를 해결하기 위해 많은 연구가 수행되었으며 크게 범주화하여 적절한 수학적 구조로 구성하여 최적화 속도를 증가시키는 혼합 정수 선형 계획법(Mixed Interger Linear Programming, MILP)와 탐색 시간을 줄이기 위한 메타 휴리스틱 기법과 휴리스틱 기법이 있다.

무기-표적 할당 문제가 MILP로 정식화될 경우 비교적 빠른 시간에 최적해를 도출할 수 있다는 장점이 있다. 이를 활용하여 일반적인 무기-표적 할당 문제가 다루는 할당 자원 및 발사 준비시간에 대한 구속조건뿐만 아니라 집중공격에 대한 요격을 수행할 시 발생할 수 있는 유도탄 간 간섭배제를 위한 제약 조건을 포함하여 최적의 할당 해를 얻을 수 있음이 알려져 있다[3]. 하지만 MILP문제는 여전히 NP 문제이기 때문에 문제의 차원이 증가함에 따라 최적해를 찾는데 많은 시간이 소요되어 복잡한 전장 상황에서 실시간으로 활용하기 어렵다. 메타 휴리스틱 알고리즘은 높은 연산 시간 문제를 해결하고 실시간 운용이 가능한 할당 알고리즘을 도출하기 위해 무기-표적 할당 문제에 적용되어 연구되고 있다. 대표적으로 복잡한 전장 상황에서 발생할 수 있는 충돌 방지를 위한 추가 구속 조건이 고려된 문제를 유전자 알고리즘을 이용하여 다룬 연구가 있다[4,9]. 하지만, 메타 휴리스틱 알고리즘의 특성상 할당을 일반화하기 어려우므로, 시시각각 변화하는 교전 상황에서 실시간으로 적용하기 어렵다는 단점이 있다. 휴리스틱 알고리즘은 짧은 시간 안에 근사해를 도출하는 수 있는 특징이 있다. 이러한 알고리즘을 유도탄간 간섭을 배제함과 동시에 밀도 높은 표적을 할당하기 위한 연구가 이전에 수행된 바 있다[5]. 간섭배제에 활용되는 선형 근사 간섭 예측 알고리즘의 정확도를 향상시키기 위해 발사방위각 및 발사 시간을 고려한 할당 점수를 설정하고, 탐욕 알고리즘을 활용하여 간섭의 수를 줄일 수 있는 할당 방법을 제시하였다. 하지만, 이러한 방법론은 문제 설정이 바뀔 때마다 더 좋은 할당을 얻기 위해 가중치의 조율이 필요하다.

기계 학습은 정식화되기 어려운 복잡한 문제를 다룰 수 있다는 장점이 있다. 특히 기계 학습기법의 하나인 강화학습은 에이전트가 상황을 인식하고 다양한 경험에 따른 학습을 통하여 보상을 최대화하는 행동을 선택하도록 한다. 이러한 강화학습의 특징을 활용하여 복잡한 무기-표적 할당 문제 또한 강화학습 문제로 정식화가 가능하다. Mouton[10]은 무기-표적 할당 문제를 강화학습으로 해결하기 위하여 교전 환경을 그리드로 나누어 이산화한 다음 파괴 확률을 고려하여 영역을 방어하는 동적할당 방식의 문제 및 모델을 설계하였다.

본 연구에서는 Mouton[10]과 달리 연속적인 상태 정보를 받으며, 현실적인 제한 조건을 고려함과 동시에 표적을 요격하여 피해를 최소화하는 정적할당 방식의 문제를 다룬다. 추가적으로 좁은 영역에 다수의 표적을 요격해야 하는 경우 발생할 수 있는 유도탄 간 간섭 현상을 배제하기 위해 유사한 시간에 발사된 유도탄의 궤적 교차 여부를 고려한다. 이러한 문제에서는 이전의 할당이 현재의 할당에 복합적인 영향을 미치기 때문에 문제 정의가 복잡하여 휴리스틱한 방법으로 최적해를 찾는 것은 매우 어렵다. 이를 해결하기 위하여 다수-에이전트(Multi-Agent) 강화학습 모델을 정의하고 학습된 모델을 이용하여 실시간 할당이 가능한 모델을 제시하였다. 또한, 에이전트의 수가 증가할수록 학습에 필요한 변수가 급격히 증가하여 발생할 수 있는 차원의 저주를 완화시키기 위해 평균 필드 게임(Mean Field Game) 이론을 적용하였으며 축소된 입력 차원으로 가능한 모든 표적을 할당함과 동시에 교차를 방지한 할당을 성공시켰다. 또한, 실행 시간 측면에서 표적의 개수가 많을수록 높은 효과를 보였다.

2. 무기-표적 할당 문제 정의

무기-표적 할당 문제에서 적의 위협도 최소화 문제는 아래와 같이 교전 후 남은 모든 표적의 잔존 가치의 합을 최소화하는 목적함수로 설정할 수 있다[2].

$$\min J = \sum_{t=1}^{|T|}\left[v_0(t) * \prod_{w=1}^{|W|}(1-p(t,w))^{\sum_{m=1}^{|M_w|}\theta_{w,t}(m)} \right] \quad (1)$$

여기서 T, W, M_w는 각각 표적, 발사대, 발사대의 유도탄 집합을 나타낸다. 표적 t에 유도탄 m이 할당

될 경우 표적의 가치는 초기값 $v_0(t)$로 부터 요격확률 $p(t,w)$에 의해 기하급수적으로 감소한다. $\theta_{w,t}(m)$는 발사대 w에서 발사된 유도탄 m의 표적 t에 대한 할당 여부로 할당될 경우 1, 할당되지 않은 경우 0의 값을 가진다. $p(t,w)$는 주어진 임무에 따라 설계할 수 있으며, 본 연구에서는 경로 교차방지 조건을 반영하여 설계해야 한다.

본 문제에서 고려하고 있는 최적화 문제의 제약 조건은 아래와 같다[3,5].

$$\sum_{t\in T}\theta_{w,t}(m) \le 1, \forall w\in W, m\in M_w \tag{2-a}$$

$$\sum_{w\in W}\sum_{m\in M_w}\theta_{w,t}(m) \le 1, \forall t\in T \tag{2-b}$$

$$\text{if } \theta_{w,t}(m)\equiv 1 \\ then\ \tau_w(m)\in TW_{t,w}, \forall t\in T, \forall w\in W, m\in M_w \tag{2-c}$$

$$\tau_w(m_2)-\tau_w(m_1)\ge \tau^d \\ \forall w\in W, m_1,m_2\in M_w, m_1\le m_2 \tag{2-d}$$

식 (2-a)는 유도탄 한 기가 여려 발의 표적을 요격할 수 없음을 나타내는 일반적인 할당 제약 조건이다. 식 (2-b)는 표적당 최대 하나의 유도탄을 할당할 수 있다는 제약 조건을 나타내며, 식 (2-c)는 할당되는 유도탄의 발사 시간 $\tau_w(m)$을 시 구간(Time Window, TW) 내로 제한한다. 마지막으로 식 (2-d)는 발사대의 운용 제약 조건으로 τ^d는 각 발사대가 가지는 최소 발사 시간 간격을 나타낸다.

무기할당 문제에서 주어진 임무를 반영하여 원하는 할당을 얻기 위해서 요격확률 $p(t,w)$를 알맞게 정의해야 한다. Fig. 1은 Daniel[5]이 사용한 아군 및 적 발사대의 설정에서 가독성을 위하여 아군의 발사대는 2대를 사용하였을 때 지향(왼쪽) 및 회피(오른쪽)하고자 하는 할당 형태이다. 오른쪽 그림과 같이 할당 결과가 유도탄 궤적 간의 교차를 포함하는 경우 유도탄 간 간섭이 발생할 수 있으며 이는 실제 교전 상황에서 요격 성능의 저하 요소로 작용할 수 있다. 하지만 각각의 무기 표적 할당이 전체 할당 결과에 영향을 받기 때문에 교차방지를 고려한 $p(t,w)$를 수식적으로 정의하기 어렵다. 그러므로 본 연구에서는 경로 교차를 방지한 할당을 강화학습을 통하여 해결하고자 한다.

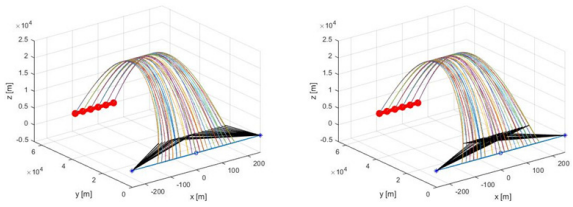

Fig. 1. Safe(left) and unsafe(right) assignment

다음 장에서는 이를 위한 이론적 배경 및 방법론을 기술한다.

3. 이론적 배경

3.1 심층 강화학습

강화학습의 주체인 에이전트는 환경에 대한 사전 지식없이 최적의 행동 정책 $\pi: S\rightarrow A$를 학습한다. 에이전트는 최적의 행동 정책을 찾기 위해 반복을 통하여 주어진 환경 및 상태 $s\in S$에 대해서 특정 행동 $a\in A$를 했을 때 보상 r을 받고, 다음 상태 $s'\in S$로 전환된다. 최종적으로 경험을 통해 구한 최적 행동 정책 $\pi: S\rightarrow A$이 종단 상태에서 보상이 최대가 되도록 한다. 이러한 최적화 문제는 마르코브 결정 과정(Markov Decision Process, MDP)으로 정의되며, 최적 정책을 찾기 위해서 모든 경우의 수를 탐색하는 과정이 필요하다. 따라서 초기 MDP의 해를 구하는 강화학습 방법으로, 모든 상태와 행동에 대해 보상값의 기대치 Q를 저장하는 표를 만들어 훈련하는 동안 에이전트가 행동 후 얻는 보상을 업데이트하는 방식을 사용한다. Q값을 결정할 때 행동 후 즉시 받는 보상뿐만 아니라 앞으로 받을 보상에 감가율 $\gamma\in[0,1]$를 곱하여 (3)과 같이 나타냄으로써 미래에 대한 보상을 같이 고려한다.

$$Q(s,a)=r(s,a)+\gamma\max_a Q(s',a) \tag{3}$$

하지만, 에이전트의 상태 및 행동의 차원이 커짐에 따라 Q값을 표에 저장하는 것에 한계가 발생하기 때문에 이를 해결하고자 Q함수를 심층 신경망으로 근사한 심층 강화학습이 소개되었다[7]. 이러한 기법은 이미지와 같은 높은 차원의 데이터를 상태로 받을 수 있도록 하였고, 추가적으로 경험 리플레이를 도입하여 과거의 행동 선택에 대한 경험을 유지하고, 타깃 네트

워크를 통하여 학습의 안정성을 높였다.

3.2 평균 필드 게임

멀티 에이전트 강화학습은 2개 이상의 에이전트가 환경과 상호작용하여 행동 정책을 찾고 이를 통해 보상을 최대화하는 것을 중점으로 한다. 하지만, 멀티 에이전트 강화학습 환경에서 보상은 개인의 행동뿐만 아니라 다른 모든 에이전트의 행동에도 영향을 받기 때문에 단일 에이전트 강화학습에 비해 문제의 복잡도가 매우 증가한다. 이에 많은 멀티 에이전트 강화학습 연구에서 문제에 따른 게임 이론을 적용하여 에이전트의 전력 선택 문제를 해결하였다. 그 중에서 내쉬 균형(Nash Equilibrium) 이론은 비협력적 게임에서 최적의 행동 정책을 찾기 위하여 중요하게 여겨진다. 내쉬 균형에서 각 에이전트는 다른 에이전트의 행동이 주어졌을 때 다른 에이전트의 전략을 고려하여 최선의 선택을 하는 전략을 이용한다. 하지만 다수의 에이전트 환경에서 고려할 상호작용의 수가 많아 내쉬 균형 전략을 찾기 어려운 문제가 있다. 이를 해결하기 위해 에이전트 주변의 상호작용을 하나의 평균 상호작용으로 고려하여 문제의 복잡도를 줄인 다음 내쉬 평균을 찾는 이론인 평균 필드 게임(Mean Field Game, MFG)이 도입되었다. MFG는 다수의 에이전트가 소규모로 상호작용하는 환경에서 전략적인 의사 결정을 하기 위한 연구로 최근에 에이전트 수가 급증할 때 차원의 저주 문제를 완화하고자 멀티 에이전트 강화학습에 적용되었다[8].

Yang[8]은 확률적 게임(Stochastic Game)에서 Q함수를 전체 에이전트의 행동쌍이 아닌 자신의 행동과 이웃한 에이전트의 평균 행동만을 사용하여 입력 데이터의 차원을 축소한다. 예를 들면, 에이전트 j의 Q함수는 본래 상태 s와 모든 에이전트의 행동쌍 $\boldsymbol{a}$로 나타내지만, 수식 (4)와 같이 자신의 행동과 이웃 에이전트의 행동의 평균으로 나타낼 수 있다.

$$Q^j(s,\boldsymbol{a}) = \frac{1}{N_j}\sum_{k\in N(j)} Q^j(s,a^j,a^k) \approx Q^j(s,a^j,\bar{a}^j) \qquad (4)$$

수식 (4)에서 $N(j)$는 에이전트 j의 이웃 에이전트의 행동 인덱스 집합으로 $N_j = |N(j)|$의 크기를 가진다.

모든 에이전트가 균일할 때 수식 (4)를 이용한 강화학습의 타당성과 수렴성이 증명되었으며, 본 연구에서는 평균 필드 게임 이론을 이용하여 멀티 에이전트 강화학습을 이용한 무기-표적 할당 문제의 차원의 저주 문제를 완화하고자 적용하였다.

4. 강화학습 기반의 무기-표적 할당

본 연구에서는 유도탄간 경로 교차 방지를 고려한 무기-표적 할당 문제의 해를 실시간으로 구하기 위해 멀티 에이전트 강화학습 모델을 설계하였다. 이러한 강화학습 모델에서는 발사대에서 발사할 유도탄을 에이전트로 고려하여 표적 위치와 아군 발사대의 정보를 상태로 받아 할당 결과에 따라 따른 보상을 통하여 교전을 성공시킬 수 있도록 학습한다. 또한, 입력의 차원을 줄이고 학습 속도를 향상시키고자 평균 필드 게임을 도입하여 각 유도탄과 표적 할당쌍 대신 이전 타임 스텝에서 표적에 할당된 유도탄 수를 정규화한 할당 분포를 입력으로 받아 학습을 진행한다.

평균 게임이론을 기반으로 한 멀티 에이전트 강화학습의 각 에이전트의 상태는 모든 표적의 x좌표 위치, 이전 타임 스텝의 행동 분포(표적 할당 분포), 유도탄이 발사된 발사대의 x좌표 위치, 시험에 사용된 발사대의 개수로 정의하였다. 에이전트의 행동은 표적의 색인으로 정의하였으며, 시험에 사용된 표적의 개수보다 높은 색인은 미할당으로 간주한다. 강화학습의 행동 선택은 보상을 통하여 결정되며 최종적으로 에이전트는 높은 보상을 얻도록 행동 정책을 학습하기 때문에 보상 설계가 매우 중요하다. 보상의 경우 다음과 같은 요인을 고려한다.

case1 : 유도탄을 표적에 할당하지 않은 경우
case2 : 유도탄을 표적에 할당하였지만, 먼 표적을 할당한 경우
case3 : 유도탄을 표적에 할당 및 가까운 표적에 할당한 경우

이전의 연구에서는 요격한 표적의 개수에 발사대와 표적간의 거리로 가중치를 두어 가까운 표적을 할당하는 방식을 기본으로 하였다[5]. 다만, 너무 거리를 중요시 할 시 표적이 치우쳐져 있을 때 유도탄 궤적이 교차할 수 있는 문제가 발생할 수 있으며 미할당 또는 중복할당으로 인하여 요격 성능이 저하될 수 있다. 따라서 본 연구에서는 거리를 고려하되 범위의 격차를 줄이고, 전체 요격에 실패했을 시 큰 패널티를 부

여한다. 이를 고려한 보상 함수는 수식 (5)와 같다.

$$R_i = \begin{cases} -0.3 \cdot N_{miss} & \text{if } case1 \\ -0.04 \cdot \sqrt{d_{diff}} - 0.2 \cdot N_{dup} & \text{if } case2 \\ 0.05 + 0.04 \cdot \sqrt{d_{diff}} - 0.2 \cdot N_{dup} & \text{if } case3 \end{cases} \quad (5)$$

N_{miss}는 놓친 표적의 수를 의미하며, case1의 경우 미할당으로 인하여 표적을 맞히지 못할 시 표적을 할당한 경우보다 높은 패널티를 준다. case2와 case3의 경우 할당을 하였으나 올바른 할당을 하였는가를 확인한다. 이를 할당하는 방법으로 비교적 가까운 표적을 할당하였는 지와 다른 유도탄과 같은 표적을 중복 할당하여 놓친 표적이 없는 지를 확인한다. d_{diff}는 이웃 발사대와 표적 간 거리의 차이며, N_{dup}는 중복 할당으로 놓친 표적의 수이다. 최종 할당 결과는 개인의 할당보다 다른 할당과 종합적으로 결정되므로, 에이전트 개인의 보상을 구한 다음에 모든 에이전트에 공통적으로 $R_c = 1.0 - 0.4 \cdot N_{miss}$를 더하여 결과적으로 좋지 못할 경우 다른 에이전트도 행동 정책을 수정할 수 있도록 한다.

5. 임의의 표적 개수 할당을 위한 알고리듬

본 연구에서 사용하는 신경망은 고정된 차원의 입력만 받을 수 있기 때문에 고려할 수 있는 표적의 개수와 발사대 개수에 한계가 있다. 따라서 그 이상의 표적을 다룰 수 있도록 확장하기 위하여 전체 표적을 가장 빠른 요격 시간(EFDT)과 발사준비 시간 간격을 통해 먼저 분할한 다음 발사대 최대 가용 개수를 고려하여 최대 6개씩 분할하였다. 분할한 표적을 각각 학습을 완료한 모델에 적용하여 해를 구한 다음 최종적으로 발사대의 지연시간을 고려하여 발사 시간까지 구하였다. 자세한 알고리즘은 Algorithm 1과 같다. Algorithm 1에서 사용된 T_list와 $candidate$는 분할한 표적의 인덱스와 표적을 저장하는 리스트이며, sep_T에 전체 표적을 각 시행당 할당할 표적의 형태로 분할하여 저장된다. 최종적으로 최대 6개의 표적으로 구성된 sep_T의 원소마다 학습 모델을 통한 할당을 진행하여 그 결과를 res 저장하며, 전체 결과는 $Assignment_pair$에 저장된다.

본 연구의 강화학습 모델은 한번에 최대 6개의 표적을 다루기 때문에 6개 미만의 표적 정보를 받을 경우에도 학습이 가능하도록 설계가 필요하다. 이를 해결하기 위하여 6개 미만의 표적에 대해서 비어있는 표적 정보는 해당하는 신경망의 입력에 –1로 패딩하였다. 또한, 표적의 개수에 따라 문제의 난이도가 달라지기 때문에 균일하게 학습을 한다면 표적의 개수마다 학습 속도의 차이가 매우 클 수 있다. 따라서 표적이 개수가 많을수록 훈련 데이터를 확률적으로 많이 생성할 수 있도록 하여 비교적 학습 속도의 차이를 완화하였다. 학습을 진행할 때는 표적 개수와 발사대 개수를 확률적으로 정한 후 고정한 다음 위치를 달리한 에피소드를 1000번 생성하였다. 이와 같은 방식으로 1000개의 시나리오를 생성하여 총 10^6개의 데이터를 만들어 학습을 진행하였다.

Algorithm 1 Divide and Assign

```
procedure DIVIDEANDASSIGN(T, W, TW, τ^d)
    sep_T ← [];
    T_copy ← T;
    while not empty (T_copy) do
        min_time ← MAX_INT;
        for i ∈ range(T_copy) do
            if min_time > TW[t][1].EFDT then
                min_time ← TW[t][1].EFDT;
        T_list ← [];
        for i ∈ range(T_copy) do
            if TW(t,1).EFDT − min_time| < τ^d then
                T_list.append(i);
        for i ∈ T_list do
            candidate ← [T_copy[i]];
        sep_T.append(candidate);
        for i ∈ reversed(T_list) do
            delete T_copy[i];

    Assignment_pair ← [];
    for i ∈ range(|sep_T|) do
        size ← min(len(sep_T[i], 6);
        res ← Model_Load_and_Assign(sep_T[i][1 : size], W);
        Assignment_pair.append(res);
    return Assignment_pair
```

6. 실험 설계

본 연구에서는 Fig. 1과 같이 적 발사대와 아군 발사대를 y축으로 멀리 떨어진 거리에 최대 6대를 일렬로 배치한 상황을 고려한다. 교전 환경에서 사용한 아군 발사대는 x좌표가 –300~300 m에 배치되며 발사 가능한 유도탄 개수에는 제한이 없으나 발사대가 가지는 최소 발사 시간 간격(τ^d)를 1초로 설정하였다. 한편, 표적은 x좌표 –400~400 m 구간에서 요격이 가능하다고 가정하였다. 설정된 구간 내에서 학습 데이

터를 생성하였으며 최종적으로 학습 후 발사대의 위치와 표적의 위치가 학습 범위 내 무작위로 주어졌을 때 원하는 할당 결과를 도출할 수 있도록 4장에서 설명된 강화학습 모델을 학습시켰다.

고려하고 있는 설정에서는 다음 상태가 없는 단일 스텝 에피소드를 고려하므로, 선택한 행동에 대한 보상 함수와 Q함수를 동일시하였으며, Q함수를 학습하기 위하여 심층 Q 신경망을 이용하였다. 심층 Q 신경망에 상태를 넣기 전 신경망의 안정성을 위하여 각 데이터의 특성에 맞게 각각 0과 1사이로 정규화를 하였으며, Q값을 결정하는 보상도 스케일링하였다. Q값에 따른 행동 선택 방법으로 (6)의 볼츠만 정책을 사용하였다.

$$P_i(a|s) = \frac{\exp(Q_i(s,a)/\tau)}{\sum_{i=1}^{n}\exp(Q_i(s,a)/\tau)} \quad (6)$$

여기서 τ는 볼츠만 온도로, 학습 초기에는 높은 τ를 사용하여 행동을 무작위로 선택하여 넓은 영역에서 가능한 정책을 탐색한 뒤 학습이 진행됨에 따라 감소되어 최종적으로 최적의 행동을 선택하도록 한다. 학습에 사용된 다른 변수는 Table 1과 같다.

Table 1. Learning parameters

학습 변수	값
Learning rate	0.0006
Temperature decay	0.9996
Minimum temperature	0.05
Update period	500
Size of replay buffer	5000

학습을 위해 사용된 인공 신경망 모델은 Fig. 2와 같다. 인공 신경망 모델의 입력으로 표적의 x좌표, 이전 타임 스텝의 표적에 할당된 유도탄 수의 분포, 교전에 사용된 발사대의 x좌표 및 발사대의 개수를 받는다. 모델을 학습하기 위하여 은닉층 4개로 이루어진 완전연결 계층을 사용하였으며 각 은닉층 간에 활성화 함수는 ReLu(Rectifier Linear unit)를 사용하였다. 최종적으로 인공 신경망 모델은 주어진 상태에 대하여 각 표적을 선택하였을 때의 예상되는 보상인 Q값을 출력으로 보내며 이후 행동을 선택하기 위한 수식 (6)의 볼츠만 정책에 사용된다. 학습을 통해 도출된 Q함수는 인공 신경망 내의 가중치에 의하여 결정되기 때문에 가중치는 학습을 통하여 문제에서 찾고자 하는 Q함수에 근접하도록 지속적으로 개선되어야 한다. 이러한 과정을 역전파(Back Propagation)라 하며, 본 연구에서는 이를 위하여 인공 신경망의 Q함수와 수식 (5)를 통하여 도출한 보상 간의 손실 함수로 평균 제곱오차(Mean Squared Error, MSE)를 사용하였으며, 이를 최적화할 옵티마이저(Optimizer)로 Adam 옵티마이저를 사용하였다.

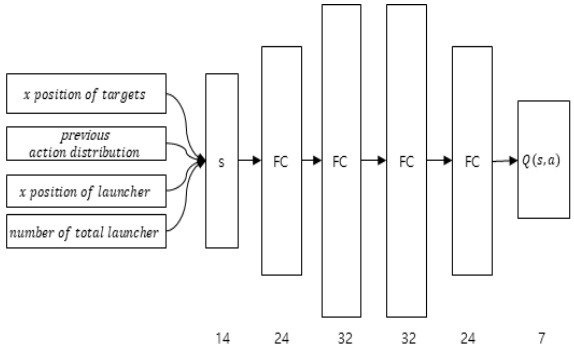

Fig. 2. Network architecture

학습을 한 후 성능 확인을 위해 시험 시나리오를 생성하였다. 시험에서는 발사대와 표적 개수를 각각 2개, 4개, 6개인 경우를 모두 고려하였다. 또한 표적과 발사대의 위치를 설정할 때 학습과 동일한 방식으로 표적의 위치는 무작위로 설정하였으며, 발사대 6대의 경우는 x축 위치를 훈련에 사용한 범위를 6등분 한 뒤 각 범위 내에서 무작위로 설정하여 시험 데이터를 생성하였다.

7. 실험 결과

본 연구에서 제시하는 강화학습 모델이 무기 할당 모델에 부합한지 확인하기 위하여 학습량에 따른 평균 보상을 계산하였다. Fig. 3은 훈련하는 동안 보상 변화를 나타내는 학습 곡선이다. 에피소드가 지남에 따라 선택한 행동에 따른 보상이 증가하는 것을 확인할 수 있다. 하지만 훈련 시나리오마다 표적과 발사대의 거리가 달라지고, 교차 할당을 방지하기 위해서 먼

거리의 유도탄을 할당할 수 있기 때문에 훈련이 끝나도 보상의 변동이 있다.

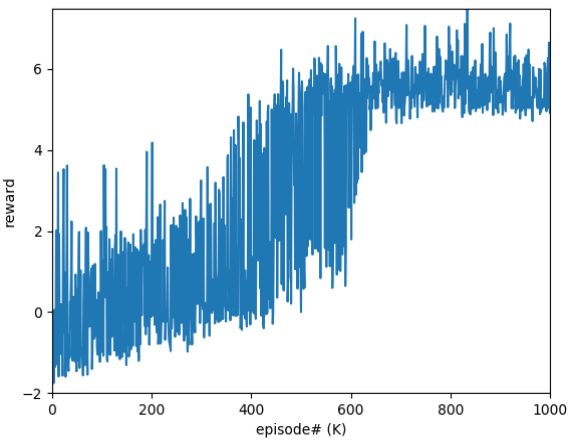

Fig. 3. Learning curve

최종적으로 학습한 모델의 표적 및 발사대에 따른 성능 확인을 위해 발사대 및 표적이 다를 때 학습량에 따른 성능 변화를 분석하였다. 성능 시험을 위하여 각 경우에 대해 50 에피소드마다 1000개의 교전 정보를 생성하였다. 각 교전 정보가 주어졌을 때 할당의 성공 여부는 모든 표적을 할당함과 동시에 유도탄간 경로의 교차 방지 여부로 결정하였다. 표적의 개수가 2, 4, 6개일 때 학습량에 따른 할당의 성공률은 아래 Fig. 4와 같다.

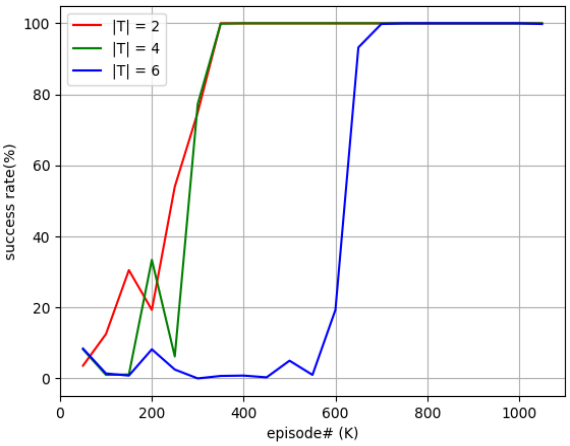

Fig. 4. Assignment success rate from different training stages

표적의 개수가 증가함에 따라 문제의 복잡도가 증가하기 때문에 성공률을 일정 수준 이상으로 확보되기 위해서 학습에 필요한 에피소드가 증가함을 확인할 수 있다. 학습 초기에는 행동을 무작위로 선택하기 때문에 성공률이 낮지만, 특정 학습량 이후에 성공률이 급격히 증가하는 경향을 가진다. 표적이 2개 및 4개 때의 성공률이 6개일 때의 성공률보다 더 적은 학습량에서 급격하게 증가하였는데 이는 표적이 많을수록 학습 데이터를 많이 생성하였으나 표적이 6개 일 경우의 문제가 비교적 복잡도가 높기 때문에 늦게 학습된 것을 확인할 수 있다. 표적이 6개일 경우 700 K 개의 데이터 학습 이후로 최적에 가깝게 수렴하였으며 표적이 2개 및 4개 일 경우 350 K 개의 데이터 학습 이후에 이를 유지하였다. 종합적으로 학습을 할수록 정확도가 높아졌으며, 최종적으로 평균 게임 이론을 통하여 입력 차원을 축소했음에도 최적에 근접한 정확도로 할당할 수 있음을 확인하였다.

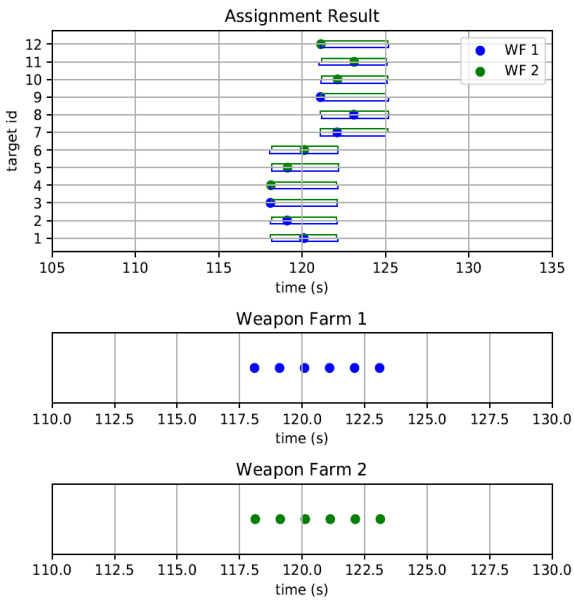

Fig. 5. Assignment results |W|= 2, |T| = 12

Fig. 5는 학습을 수행한 후 할당 결과의 경로교차 정도를 확인하기 위해 Algorithm 1을 시험한 결과이다. 시험은 가시성을 위하여 방어 발사대 2대, 표적 12개로 설정하였으며 표적은 6개의 공격 발사대에서 각각 6개씩 3초 간격으로 동시에 발사된다고 가정하였다. 학습 모델을 통하여 무기- 표적 할당쌍을 구한

후 발사 준비 시간을 고려하여 발사 시간까지 고려한 최종 할당 결과는 Algorithm 1에 맞게 발사 시간이 비슷한 표적을 표적(id 1~6과 id 7~12)끼리 분류한 다음 각각 학습 모델에 적용한 다음 발사 가능시간이 빠른 표적부터 발사 준비 시간을 고려하여 할당한 것을 확인하였다.

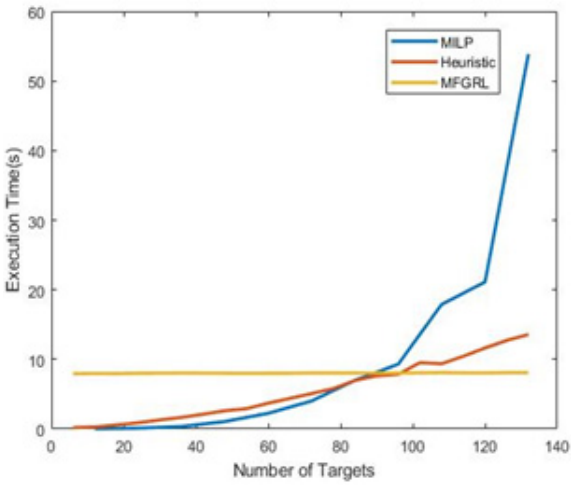

Fig. 6. Execution time with varying number of targets

Fig. 6은 본 연구와 유사한 문제를 해결하기 위한 기존의 할당 기법(MILP[3]와 휴리스틱[5]기법)과 본 연구의 학습 모델 및 알고리즘을 표적을 증가시키면서 시험하였을 때 각 소요한 시간을 나타낸다. 기존의 할당 기법은 적은 표적의 문제에 실행하였을 때 소요 시간은 적으나 표적의 개수가 증가함에 따라 할당에 필요한 시간이 점진적으로 증가하였다. 적은 표적을 다루는 문제에서는 기존의 알고리즘의 속도가 매우 빠르나 제한 조건이 많아질수록 해를 찾기 어려워져 표적이 많아질수록 급격히 증가하는 것을 확인하였다. 이는 기존의 알고리즘이 제한 조건에 크게 영향을 받기 때문이다. 최적해를 찾기 위해 MILP를 사용할 경우 표적 수와 제한 조건이 많을수록 해공간은 커지는데 반해 최적해의 비율은 감소하기 때문에 표적 수가 많아질수록 최적해를 탐색하기 위한 시간이 매우 증가한다. 휴리스틱 기법 또한 교전 정보를 토대로 할당 점수를 계산하여 순차적으로 할당을 하기 때문에 할당을 거듭할수록 제한 조건이 많아져 이전의 할당에 의해 이후의 할당 속도가 감소한다. 반면 본 연구에서 제안하는 알고리즘은 소요 시간이 표적의 수와 관계없이 거의 일정하며 제안된 알고리즘의 소요 시간은 학습된 모델을 불러오는데 사용된 시간이 매우 큰 비중을 차지하며 이후 할당에 필요한 시간은 매우 적은 것을 확인하였다. 따라서 제시된 기법은 기존의 할당 기법과는 달리 많은 표적을 방어해야 하는 교전 상황에서 다른 기법과 달리 빠르고 정확한 성능을 보여준다.

8. 결 론

본 연구에서는 복잡한 교전 상황 중 실시간 무기-표적 할당을 위한 멀티 에이전트 강화학습 모델을 제시하였다. 제시된 모델은 표적과 발사대 수가 증가함에 따라 학습 차원이 증가하는 문제를 평균 필드 게임을 이용하여 완화시켰으며 기존의 할당 기법과 달리 표적 수에 관계없이 일정 시간으로 할당함을 확인하였다. 또한 유사한 시간에 발사된 유도탄 간 경로 교차를 배제하기 위한 목적함수를 설계하여 교차 방지를 고려한 할당이 가능하게 하였다.

표적의 개수와 발사대 개수가 달라질 때 요격 성능과 교차 방지 성능을 비교 분석한 결과, 거의 모든 경우에 모든 표적에 할당됨과 동시에 교차가 방지됨을 확인하였다. 또한, 많은 수의 표적이 주어졌을 때 발사 준비시간에 따라 문제를 분할한 후 할당을 한 것을 확인하였다.

실제 교전 상황에서는 표적 위치나 종류와 같은 정제된 정보가 아닌 교전 상황을 나타내는 이미지와 같은 비정제된 정보를 얻게 된다. 본 연구에서 제시된 기법은 많은 표적의 문제에 대해서도 확장 가능하며, 추후 비정제 정보가 주어진 상황에서의 무기-표적 할당 문제로 확장이 가능할 것으로 기대한다.

후 기

본 연구는 국방과학연구소(계약번호 UD180017CD)의 연구비 지원에 의한 연구 결과임.

References

[1] Lloyd, S. P. and Witsenhausen, H. S., “Weapon Allocation is NP-complete,” Summer Computer

Simulation Conference, 1986.

[2] Ahuja, R. K., Kumar, A., Jha, K. C., & Orlin, J. B., “Exact and Heuristic Algorithms for the Weapon-Target Assignment Problem,” Operations Research, 55(6), pp. 1136-1146, 2007.

[3] Shin, M. K., Lee, D., & Choi, H. L., “Weapon-Target Assignment Problem with Interference Constraints using Mixed-Integer Linear Programming,” 2019.

[4] Lee, Z. J., Su, S. F., Lee, C. Y., “Efficiently Solving General Weapon-Target Assignment Problem by Genetic Algorithms with Greedy Eugenics,” IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), Vol. 33, No. 1, pp. 113-121, 2003.

[5] Lee, D., Shin, M. K., & Choi, H. L., “Weapon Target Assignment Problem with Interference Constraints,” AIAA Scitech 2020 Forum. 2020.

[6] Cho, D. H., & Choi, H. L., “Greedy Maximization for Asset-based Weapon-Target Assignment with Time-Dependent Rewards,” Cooperative Control of Multi-Agent Systems: Theory and Applications, pp. 115-139, 2017.

[7] Mnih, Volodymyr, et al., “Playing Atari with Deep Reinforcement Learning,” arXiv preprint arXiv:1312. 5602, 2013.

[8] Yang, Yaodong, et al., “Mean Field Multi-Agent Reinforcement Learning,” arXiv preprint arXiv:1802. 05438, 2018.

[9] M. Zhang, J. Zhang, G. Cheng, C. Chen and Z. Liu, “Fire Scheduling for Multiple Weapons Cooperative Engagement,” 2016 10th International Conference on Software, Knowledge, Information Management & Applications(SKIMA), Chengdu, pp. 55-60, 2016.

[10] Mouton, H., Roodt, J., & Le Roux, H., “Applying Reinforcement Learning to the Weapon Assignment Problem in Air Defence,” Scientia Militaria: South African Journal of Military Studies, 39(2), pp. 99-116, 2011.