https://doi.org/10.7840/kics.2024.49.7.1002

다중 무인항공기의 오프라인 강화학습 데이터셋에 따른 플라잉 애드혹 네트워크 구축 성능 연구

이 동 수◆, 권 민 혜°*

The Impact of Dataset on Offline Reinforcement Learning of Multiple UAVs for Flying Ad-hoc Network Formation

Dongsu Lee◆, Minhae Kwon°

요 약

플라잉 애드혹 네트워크(Flying ad-hoc network; FANET)는 독자적으로 통신이 가능한 공중 이동성 노드로 이루어진 네트워크로, 재난이나 전쟁과 같은 위기 상황에서 기존 통신망이 손상을 입었을 때 대체 네트워크로 활용될 수 있다. 통신 및 라우팅 기능을 갖는 다수의 무인항공기(Unmanned Aerial Vehicles)는 통신이 불가능한 지역으로 이동하여 FANET을 구축할 수 있다. 본 연구는 오프라인 강화학습을 통해 학습된 다수의 무인항공기가 중앙 제어 없이 FANET을 구축하는 시나리오를 고려한다. 본문에서는 데이터셋 및 오프라인 강화학습 알고리즘별 다중 개체의 네트워크 구축 성능 비교 실험을 수행하였으며, 데이터셋과 알고리즘의 특징에 따라 달라지는 학습 양상을 분석하였다.

키워드 : 인공지능, 다중 개체 강화학습, 오프라인 강화학습, 애드혹 네트워크

Keywords : Artificial Intelligence, Multi-agent Reinforcement Learning, Offline Reinforcement Learning, Ad-hoc Network

ABSTRACT

A Flying Ad-hoc Network (FANET) is a network of airborne mobile nodes that can communicate independently, and can be utilized as a fallback network during a crisis, such as a disaster or war, when existing infrastructures are damaged. Several Unmanned Aerial Vehicles (UAVs) with communication and routing capabilities can travel to areas where communication is not possible and establish a FANET. In this study, we consider a scenario where multiple UAVs trained through offline reinforcement learning establish a FANET without a centralized control. We conducted experiments to compare the performance of multi-agent's network construction by datasets and offline reinforcement learning algorithms, and analyzed the learning aspects that depend on the features of datasets and algorithms.

※ 이 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단(RS-2023-00278812) 및 정보통신기획평가원의 지원을 받아 수행된 연구임(No. 2021-0-00739, 분산/협력 AI 기반 5G+ 네트워크 데이터 분석 기능 및 제어 기술 개발)

◆ First Author : Soongsil University Department of Intelligent Semiconductors, movementwater@soongsil.ac.kr, 학생회원

° Corresponding Author : Soongsil University Department of Intelligent Semiconductors and School of Electronic Engineering, minhae@ssu.ac.kr, 종신회원

논문번호 : 202403-041-0-RE, Received February 15, 2024; Revised March 27, 2024; Accepted March 27, 2024

Ⅰ. 서 론

재난이나 전쟁과 같은 위기 상황에서 기존 통신망이 동작 불능일 때 네트워크를 구축 및 복구하는 기술이 주목 받고 있다. 특히, 무인항공기와 같은 이동성 노드를 기반으로 ad-hoc 네트워크를 구축하는 방식은 지형에 따른 제약이 적기에 위기 상황에서의 대체 통신망으로서 적합하다[1,2]. 다수의 무인항공기는 편대를 이루어 작동 불능 상태가 된 통신 시설의 위치에 배치되어 기존 통신망을 복구하거나 새로운 네트워크를 형성하기 위해 긴급 통신이 필요한 지역으로 이동하는 것을 목표로 한다.

네트워크 복구를 위해 다수의 무인항공기를 제어하는 방식은 크게 중앙 제어와 자율적 의사 결정 방식으로 구분할 수 있다. 중앙 제어 방식의 경우 복구 지점을 정확히 판단할 수 있는 경우 중앙 시스템의 직접적인 제어를 통해 이동하므로 무인항공기를 정확한 지점에 위치시킬 수 있다[3]. 그러나 중앙 제어 방식은 모든 노드의 위치와 지형지물의 위험성 등 수많은 변수를 실시간으로 파악해야 하기에 노드의 수와 주변 환경의 복잡도에 따라 구현이 어렵거나 큰 비용을 소모할 수 있다[4]. 따라서 본 연구에서는 무인항공기의 자율적 의사 결정 방식을 통해 효율적으로 FANET을 구축하는 방안을 고려한다.

여러 대의 무인항공기가 협력하여 FANET을 구축하기 위해서는 다수의 강화학습 개체를 동시에 학습시키는 다중 개체 강화학습(multi-agent reinforcement learning) 방법이 고려될 수 있다. 다중 개체 강화학습은 중앙 집중형(centralized) 방식과 완전 분산형(decentralized) 방식으로 구분할 수 있다[5]. 중앙 집중형 방식의 경우 개체 간 지속적인 통신 및 전체 개체를 관리하는 중앙 네트워크의 운용이 포함되기 때문에 실용적이지 못하다는 한계가 있다. 반면 완전 분산형 방식의 다중 개체는 모든 개체가 과제 수행을 위한 독립적인 정책을 학습한다. 본 연구에서는 개체마다 개별적인 정책 네트워크를 관리하며 지속적인 통신을 통한 정보의 공유가 요구되지 않는다는 점에서 완전 분산형 방식을 채택한다.

일반적인 강화학습 방법은 개체가 환경과 실시간으로 상호작용하며 과제 수행 방법을 학습하기에 온라인 강화학습(online reinforcement learning)이라 불린다. 해당 방식은 탐험을 통한 시행착오를 기반으로 정책을 최적화한다. 이와 같은 방법론은 사회적 및 금전적 문제를 야기할 수 있으며 데이터 샘플을 수집하기 위해 큰 비용이 필요하다는 점에서 실용성이 떨어진다는 한계를 갖고 있다[6]. 반면, 오프라인 강화학습(offline reinforcement learning)은 환경과의 상호작용 없이 사전 수집된 데이터셋을 통해 정책을 학습한다. 따라서 오프라인 강화학습은 데이터셋의 품질, 즉 데이터셋이 유의미한 정보를 포함하는 정도에 따라 학습 성능이 변화한다는 특징이 있으며 부적합한 행동들을 직접 경험할 필요가 없기에 온라인 강화학습에 비해 안정적인 학습이 가능하다는 장점을 갖는다[7-8].

본 연구에서는 FANET을 구축하기 위한 자율적 의사결정 방식으로 완전 분산형 다중 개체 강화학습을 고려하며, 안정적인 구현을 위해 오프라인 강화학습 방안을 차용한다. 이와 같은 현실적인 강화학습 문제 해결을 위해 분산형 부분 관측 마르코프 의사결정 모델(decentralized partially observable Markov decision process; Dec-POMDP)을 제안한다. 제안하는 방식의 실용성을 살펴보기 위해 다양한 알고리즘 및 데이터셋에 따른 성능 비교 시뮬레이션을 진행한다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 선행 연구를 제시한다. Ⅲ장에서는 본 연구에서 고려하는 FANET 구축 시나리오의 부분적 관측 마르코프 결정과정 모델을 소개한다. Ⅳ장에서는 시뮬레이션 결과를 통해 FANET 구축 성능을 비교하며 Ⅴ장에서는 결론을 맺는다.

Ⅱ. 선행 연구

2.1 사전 수집 데이터셋 기반 강화학습

기존의 강화학습은 개체가 환경과 실시간으로 상호작용하여 누적 보상을 최대화 할 수 있는 최적의 정책을 학습하는 온라인 방식에 기반한다. 기존 강화학습 연구들은 수많은 성공을 보고하였지만 해당 방법론의 실용적인 한계는 분명하여 실제 세계 문제에 적용되는 사례는 미비하다. 구체적으로 개체는 시행착오를 통해 데이터를 수집하여 정책을 개선하며 이를 위해 굉장히 많은 수의 시뮬레이션을 요구한다. 이러한 방식은 실제 세계의 문제의 경우 굉장히 높은 금전적 비용 및 사회적 혼란에 대한 위험성을 포함한다[6].

온라인 방식의 실용적 문제를 해결하기 위해 최근에는 사전 수집된 데이터셋을 기반으로 학습 기반의 제어 모델을 다루는 연구가 활발히 진행되고 있다. 가장 대표적인 방법으로는 모방학습(behavioral cloning; BC)으로, 지도학습과 같이 주어진 관측 값에 대해 데이터셋의 행동과 모델의 행동을 유사하게 만드는 것을 목표로 한다. BC의 경우 주어진 데이터셋의 품질이 전문가에 가깝지 않은 경우 제대로 된 성능을 보장하기 어렵다는

한계가 있다. 다음으로, 기존의 강화학습 방식에서 고정된 데이터셋을 통해 학습할 때 발생할 수 있는 out of distribution (OOD) 문제 해결을 위해 제약을 함께 고려하는 오프라인 강화학습이 있다[9]. 예를 들어, imitative learning(IL)[7]의 경우 강화학습의 정책 학습 방식에 직접적으로 모방학습 방식을 결합하여 데이터 샘플의 가치 평가와 모방 사이에 균형을 고려하는 방식을 이용한다. 반면, implicit Q-learning(IQL)[10]의 경우 OOD 문제의 발생 원인인 네트워크 업데이트 과정에서 데이터셋 내에 존재하지 않는 행동을 선택하는 과정의 필요를 제거하여 암묵적으로 규제하는 방식을 이용한다.

언급된 모든 알고리즘들은 본래 하나의 강화학습 개체만을 학습하기 위한 방법론이다. 본 연구에서는 다중 개체를 고려하기 때문에, 해당 알고리즘을 단일 개체가 아닌 다중 개체 학습을 위한 환경에서 확장하여 적용 및 평가한다.

2.2 다중 개체 강화학습

다중 개체 강화학습은 다수의 개체가 환경과 상호작용하며 획득한 보상을 토대로 과제 수행 방식을 습득하는 기계학습 방식이다. 각 개체는 관측(observation)한 정보를 토대로 행동(action)을 취하며, 팀 보상(team reward)을 최대화 할 수 있는 정책을 학습하는 것을 목표로 한다[11].

다중 개체 강화학습은 학습 과정에 있어 중앙 집중형 방식과 분산형 방식으로 구분할 수 있다[5]. 중앙 집중형 방식의 경우 개체의 의사결정 및 정책 학습을 위해 외부 정보를 이용할 수 있다는 특징이 있다. 대부분의 선행 연구는 협력 문제 해결을 위해 이와 같은 중앙 집중형 방식을 이용한다[12-15]. 하지만, 이러한 설정은 각 개체 간 지속적인 통신이 가능해야 한다는 가정을 포함하고 있어 실용적으로 한계가 있다. 한편, 완전 분산형 학습 방식의 경우 개인의 관측 정보 만을 이용하여 주어진 문제 상황에 대한 최적의 의사결정 학습을 목표로 한다. 완전 분산형 학습 방식은 중앙 집중형 방식에 비해 개체의 정책 최적화가 어렵다는 한계가 있지만, 보다 실용적으로 이용될 수 있다. 본 연구에서는 오프라인 강화학습 알고리즘을 이용하여 다중 개체를 완전 분산형으로 학습시키는 방법을 채택한다.

2.3 강화학습 기반 애드혹 네트워크 구축 연구

애드혹 네트워크 구축 연구는 군사 및 재난 상황 등에서 기존 네트워크가 손실된 경우 네트워크 복구를 위해 큰 관심을 받고 있다[16]. 애드혹 네트워크 구축을 위해서 이동기기는 자율적인 의사결정 능력이 필수적으로 요구되기 때문에 심층강화학습 기반의 다양한 연구가 수행되고 있다. 애드혹 네트워크 구축을 위해 다양한 기기들이 고려될 수 있으며, 이때 2차원 이동기기를 활용할 경우 mobile ad-hoc network (MANET), 3차원 이동기기를 활용할 경우 FANET으로 분류될 수 있다. [17]에서는 2차원 기기들을 기반으로 다수의 에이전트를 커리큘럼 방식으로 학습하는 방식을 채택하였다. [18]에서는 3차원 기기를 이용하며 커리큘럼 방식으로 단일 에이전트를 학습한다. [19]에서는 해당 문제를 오프라인 강화학습 문제로 확장하여 연구를 수행하였다. 또한 [20]에서는 3차원 이동기기가 유선통신에 비해 비교적 불안정하며 보안이 취약함에 주목하여, 주파수 공유 참여 여부를 고려하며 FANET을 구축하였다.

하지만 기존 선행 연구들의 경우 다수의 3차원 기기를 제어하여 애드혹 네트워크를 구축하는 연구는 굉장히 미비하다. 또한, 일부 연구를 제외하고는 복구 행동이 위기 상황에서 이루어질 수 있다는 주목도 역시 낮은 상태이다. 따라서 본 연구는 다수의 3차원 기기를 이용하여 FANET을 구축하는 연구를 고려하며, 동시에 고정된 데이터 셋을 통해 정책을 구축하는 오프라인 강화학습을 고려한다.

Ⅲ. 다중 개체 오프라인 강화학습 기반 무선 애드혹 네트워크 구축

본 장에서는 문제상황을 정의하고 그 해결방법으로 분산형 다중 개체 오프라인 강화학습 방법을 제안한다.

3.1 고려하는 FANET 환경

본 연구는 그림 1과 같이 여러 대의 무인항공기가 $M \times M \times M$ 크기의 환경에서 서로 직접적인 통신이 불가능한 위치에 놓인 두 개의 기반시설을 연결하도록 FANET을 구축하는 상황을 고려한다. FANET을 구축하기 위해 각각의 무인항공기는 자율적인 의사결정을 통해 애드혹 네트워크를 구축하는 것을 목표로 한다.

학습 환경 내에 존재하는 모든 통신 노드들은 집합 $E=\{e_s,e_1,e_2,\cdots,e_N,e_d\}$로 정의된다. 집합 E 에서 e_s와 e_d는 소스노드(source node)와 목적노드(destination node)로, 각각 정보를 송신하거나 수신하는 기반시설을 의미한다. 두 노드 사시에 위치하여 FANET을 구축하는 N개의 다중 무인항공기는 $e_1,e_2,\cdots,e_N$로 표시된다. 모든 노드의 경우 이중통신(full duplex)를 가정하며, 반경 δ 이내에 존재하는 다른 노드와 무선 통신 및 라우팅이 가능하다.

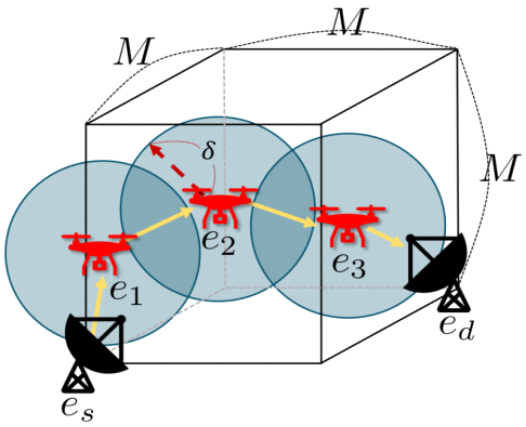

그림 1. 고려하는 FANET 환경
Fig. 1. Considered FANET environment

3.2 분산형 부분 관측 마르코프 의사결정 모델

본 연구는 현실적인 강화학습 문제 해결을 위해 개체별 통신 범위를 제한하여 분산적으로 개별 개체가 제한적인 관측 정보를 통해 의사결정을 수행하는 Dec-POMDP로 문제를 정의한다. Dec-POMDP 설정 아래에서 다중 개체는 독립적인 의사결정을 기반으로 하여 협업을 위한 문제를 해결한다.

Dec-POMDP는 다중 개체의 의사결정 과정을 묘사하는 튜플 $<E,S,O,\{\mathrm{A_i}\},\mathrm{T},\{\Omega_\mathrm{i}\},\mathrm{R},\gamma>$로 정의한다. 개별 개체는 학습과 행동의 주체로 고려되며, 상태 관측 확률 $\Omega_i: S \rightarrow O$을 기반으로 상태 $s \in S$를 관측하여 관측 정보 $o_i \in O$를 얻은 다음, 개별 행동 $a_i \in A_i$을 결정한다. 이와 같은 다중 개체 환경에서 공동행동공간을 $A := A_1 \times A_2 \times \cdots \times A_N$로 정의할 때 공동행동은 $\boldsymbol{a} = \{\mathrm{a_1},\mathrm{a_2},\cdots,\mathrm{a_N}\}$와 같이 정의할 수 있다. 이때, 상태 전이 확률은 $T: S \times A \rightarrow S$로 정의할 수 있으며, 보상의 경우 보상함수 $R(s,\boldsymbol{a},s')$를 통해 얻을 수 있다. 여기서 s'은 상태 및 공동행동의 결과인 다음 상태를 의미한다. 본 연구의 경우 완전한 협력적 업무(fully cooperative task)를 가정하기 때문에 모든 개체의 팀 보상은 동일하게 주어진다.

강화학습은 관측 및 행동 쌍 및 획득한 보상을 바탕으로 최적의 정책을 찾는 것을 목표로 한다. 여기서 최적의 정책이란 누적 보상을 최대화 할 수 있는 행동을 출력할 수 있는 정책을 의미한다. 구체적으로, 개별 개체의 목적은 $J(\pi) = E_\pi[\sum_k \gamma^k R(s,\boldsymbol{a},s')]$와 같이 정의할 수 있다. 이때, 해당 문제를 해결하기 위해 누적 보상을 근사하는 $Q = E_\pi[\sum_k \gamma^k R(s,\boldsymbol{a},s')]$ 함수를 이용한다.

3.2.1 상태 정보(state)

상태 정보는 환경 내에 존재하는 모든 정보를 묘사한다. 본 실험의 FANET 구축 여부는 소스노드와 목적노드 사이에 위치한 중계 노드들의 위치만으로 파악할 수 있기에, 상태 $s \in S$는 통신 노드들의 절대 좌표 정보로 정의한다.

$$s = [\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}]^\top$$

축별 벡터 $\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}$는 E에 속한 모든 노드들의 좌표 정보 $\boldsymbol{x} = [x_s,x_1,\cdots,x_N,x_d]^\top$, $\boldsymbol{y} = [y_s,y_1,\cdots,y_N,y_d]^\top$, 그리고 $\boldsymbol{z} = [z_s,z_1,\cdots,z_N,z_d]^\top$로 구성되어 있다. 각각의 정보는 축별 위치 정보를 포함하고 있다.

3.2.2 관측 정보(observation)

이동성 노드의 자율적 의사 결정을 통한 FANET 구축 상황에서는 중앙 제어 방식과 달리 다중 개체가 네트워크를 구축하는 특정 장소를 알 수 없다. 그렇기에 다중 무인항공기는 소스노드와 목적노드를 연결하는 특정 위치를 발견하기 위해 다른 노드와의 통신을 생성하고 유지하는 것을 목적으로 학습해야 한다.

개체 e_i가 특정 시점에서 관측 가능한 노드 $e_j \in E$와의 축별거리 $l_{\{\mathrm{x,y,z}\}}(\mathrm{i,j})$를 기반으로 유클리디안 거리 $d(i,j) = \sqrt{l_x(i,j)^2 + l_y(i,j)^2 + l_z(i,j)^2}$를 측정한다. 이때, $d(i,j) \le \delta$를 만족하는 e_j를 개체 e_i에 인접하다고 정의할 수 있다. 무인항공기 e_i는 가장 가까운 P개의 노드를 관측 및 통신할 수 있다고 가정한다.

이를 고려할 때, 개체 e_i의 개별 관측 o_i는 다음과 같이 정의한다.

$$o_i = [x_i,y_i,z_i,D_i]^\top$$

개체는 자신의 절대적 위치 및 인접한 P개의 노드와의 상대 거리 집합 D_i을 포함한다. 구체적으로 $D_i = \{\mathrm{d(i,1)},\cdots,\mathrm{d(i,P)}\}$로 정의한다.

3.2.3 행동(action)

개체 e_i의 행동 $a_i \in A_i$는 축별 이동량 정보를 포함하여 다음과 같이 정의할 수 있다.

$$a_i = [\Delta x_i,\Delta y_i,\Delta z_i]^\top$$

e_i는 각 축 방향으로 $\Delta x_i,\Delta y_i,\Delta z_i$ 만큼 이동하며, 개

체의 축별 이동량은 최대 및 최소 행동 범위 $A_{\min}$과 $A_{\max}$ 사이에서 결정된다.

3.2.4 보상(reward)

다중 개체는 다음과 같은 보상 함수를 통해 네트워크 구축 방안을 학습한다.

$$r = \begin{cases} 1, & \text{network is constructed} \\ 0, & \text{otherwise} \end{cases}$$

소스노드와 목적노드 간 모든 중계 노드들의 통신 범위가 연쇄적으로 중첩되었다면 소스노드와 목적노드는 통신할 수 있다. 이와 같은 경우 환경은 다중 개체의 협력적 움직임을 장려하기 위해 수적 보상 1을 제공한다. 그러나 네트워크가 구축되지 않았거나 구축 상태를 유지하지 못하였을 경우 해당 타임스텝의 보상은 0으로 책정된다.

다중 개체는 네트워크를 구축하기 위한 특정 지점을 정확히 알지 못하는 상태에서 학습을 진행하므로 각 개체가 정확한 구축 지점에 도착함에 따른 개별 보상은 책정될 수 없다. 따라서 보상 r은 네트워크가 연결된 시점에 모든 개체가 동일한 값을 얻는 공통 보상(common reward) 방식으로 제공된다.

3.3 독립적 다중 개체 오프라인 강화학습

본 연구에서는 보다 실용적이고 현실적인 다중 개체 강화학습 알고리즘을 고려하기 때문에, 개체의 종합 정보를 이용하는 네트워크가 아닌 다중 개체의 학습을 분산적 및 독립적으로 수행한다. 또한, 온라인 강화학습 환경의 경우 개별 개체 외의 상호작용을 수행하는 외부 개체의 정책이 정적이지 않기(non-stationary) 때문에 독립적인 다중 개체 강화학습의 적용이 어려웠다. 이를 해결하기 위해 온라인이 아닌 오프라인 강화학습 환경에서 독립적 개별 개체를 학습한다.

연속적인 상태 및 행동 공간에서 독립적 다중 개체 오프라인 강화학습 문제를 해결하기 위해 본 연구에서는 액터-크리틱 알고리즘을 고려한다. 액터 네트워크 ϕ는 정책 π을 근사하는 네트워크로 주어진 관측값에 대한 행동 값을 출력한다. 크리틱 네트워크 θ는 정책 학습을 위해 누적 보상을 추정하는 Q 함수를 근사하는 네트워크이다. 문제 해결을 위해 고려되는 N 개의 개체는 개별적인 네트워크를 갖는다. 각 네트워크의 손실함수는 다음과 같이 일반화하여 정의할 수 있다.

$$L(\phi_i) = -Q_{\theta_i}(o_i, \pi_\phi(o_i)) \quad (1)$$

$$L(\theta_i) = (r + \gamma Q_{\theta_i'}(o_i', \pi_{\phi_i}(o_i')) - Q_\theta(o_i, a_i)) \quad (2)$$

액터의 손실함수는 현재 액터가 선택한 행동의 가치가 높을수록 감소하며 크리틱의 손실함수는 현재 얻은 보상과 다음 관측 및 다음 행동에 대한 감가된 가치의 합인 $r + \gamma Q_{\theta_i'}(o_i', \pi_{\phi_i}(o_i'))$와 현재 크리틱이 평가한 가치 $Q_\theta(o_i, a_i)$의 차이가 작을수록 줄어든다. 다중 개체의 네트워크 학습 방법은 알고리즘 1을 통해 확인할 수 있다.

이때 오프라인 강화학습에서는 정책이 데이터셋 내에 포함되지 않은 행동을 출력하는 OOD 문제가 발생할 수 있으며, 업데이트 과정에서 오류가 누적되어 정책의 학습이 어려울 수 있다. 해당 문제를 해결하기 위해 다양한 방법론을 적용할 수 있는데 본 연구에서는 다음과 같은 방법론을 이용한다.

Algorithm 1 Independent multi-agent reinforcement learning algorithm

Require: offline dataset $\mathcal{D}$, training iterations K, the number of agent N, target update frequency u, soft update ratio τ

Initialize: actor network parameters $\{\phi_1, \phi_2, \phi_N\}$, critic network parameters $\{\theta_1, \theta_2, \theta_N\}$, target critic network parameters $\{\theta_1', \theta_2', \theta_N'\}$

for $k = 1,\ K$ **do**
  for $n = 1,\ N$ **do**
    Sample $(\boldsymbol{o}, \boldsymbol{a}, \boldsymbol{o}', r) \sim \mathcal{D}$
    Calculate critic loss using (2)
    Calculate actor loss using (1)
    Back-propagate and update ϕ_n and θ_n
    if k mod u then
      Initialize hidden cell ρ_Q
      Update target networks:
      $\theta'_n \leftarrow \tau\theta_n + (1-\tau)\theta'_n$
    end if
  end for
end for

3.3.1 Imitative Learning

IL은 앞서 정리된 일반적인 강화학습 알고리즘에 BC 알고리즘을 결합한 오프라인 강화학습 알고리즘이다. IL의 경우 크리틱 네트워크의 손실함수는 식 (2)와

동일하며, 액터 네트워크의 손실함수는 다음과 같이 정의된다.

$$L(\phi_i) = -Q_{\theta_i}(o_i, \pi_\phi(o_i)) + (a_i - \pi_\phi(o_i))^2 \quad (3)$$

해당 손실함수 값은 정책이 선택한 행동 $\pi_{\phi_i}(o_i)$의 가치가 높을수록, 데이터셋에 포함된 행동 a_i와의 차이가 적을수록 감소한다. 즉, 해당 방법론의 정책은 가치 평가와 데이터셋의 모방을 동시에 수행하여 데이터셋 내에 존재하거나 존재하지 않는 행동 모두 고려하므로 OOD 문제를 완화할 수 있다.

3.3.2 Implicit Q-learning

IQL은 IL과 같이 직접적인 규제(regularization)항 대신 암묵적 방식의 규제 방법을 통해 OOD 문제를 완화한다. 해당 알고리즘의 경우 정책 및 Q 함수 근사 네트워크와 더불어 상태 가치 함수 V에 대한 네트워크 ψ를 추가적으로 고려한다. 각 네트워크의 손실함수는 다음과 같이 정의된다.

$$L(\phi_i) = \exp(\beta(Q_{\theta_i}(o_i, a_i) - V_{\psi_i}(o_i))) \times -\log \pi_{\phi_i}(a_i|o_i) \quad (4)$$

액터 네트워크는 advantage $Q_{\theta_i}(o_i, a_i) - V_{\psi_i}(o_i)$와 정책 네트워크에서 추출되는 행동의 확률적 값을 기반으로 갱신된다. 즉, 주어진 상태 및 행동 쌍의 advantage 값을 고려하여 어느 정도의 크기로 해당 정책을 추출할 지를 결정할 수 있다.

$$L(\theta_i) = (r + \gamma V_{\psi_i}(o_i')) - Q_\theta(o_i, a_i)) \quad (5)$$

크리틱 네트워크의 경우 $Q_{\theta_i'}(o_i', \pi_{\phi_i}(o_i'))$를 $V_{\psi_i}(o_i')$로 대체함으로써 정책이 OOD 행동을 추출하는 문제를 고려하지 않도록 고려하였다.

$$L(\psi_i) = L_2^\tau(Q_{\theta_i}(o_i, a_i) - V(o_i)) \quad (6)$$

상태가치함수 V_{ψ_i}는 Q 함수를 타겟으로 고려하여 네트워크를 업데이트하며, 비대칭 손실함수(asymmetric loss)를 통해 기울기를 택한다.

$$L_2^\tau(u) = |\tau - \mathbb{1}\ (u < 0)|u^2$$

여기서 u는 변수를 의미하며, $\mathbb{1}\ (u < 0)$는 지시함수로 조건을 만족하는 경우 1, 그렇지 않은 경우 0을 반환한다. 해당 함수에서 τ가 1에 가까울수록 u가 양수인 영역, 즉 Q값이 V값 보다 높은 기울기를 가지며 빠른 속도로 최적화를 수행한다. 구체적으로, τ가 클수록 V는 Q의 평균적인 추정치가 아닌 greedy 한 $\max_a Q_{\theta_i}(o_i, a)$값으로 근사된다.

Ⅳ. 시뮬레이션을 통한 알고리즘 성능 분석

4.1 모의실험 설정

제안하는 방법의 애드혹 네트워크 구축 성능을 검증하기 위해 다음과 같은 설정 아래 시뮬레이션을 수행한다. 시뮬레이션은 $1000 \times 1000 \times 1000m^3$의 3차원 공간 상에서[19], 육상에 설치된 소스노드 및 목적노드만이 존재하는 환경을 고려한다. 이때, 목적노드가 소스노드부터 각 노드를 차례로 거쳐 전달되는 패킷을 수신하는 경우 애드혹 네트워크 구축 성공에 대한 신호를 모든 노드에게 브로드캐스팅하는 네트워크를 고려한다. 이를 통해 각 타임스텝마다 복구를 위한 무인이동체는 네트워크 구축 성공 여부에 대해 파악할 수 있다.

기존 통신망인 소스노드와 목적노드의 공간상 좌표는 $[0m, 0m, 0m]$ 및 $[1000m, 1000m, 0m]$로 설정되었다. 본 문제를 해결하기 위해 투입되는 강화학습 개체는 총 3대의 무인항공기로 정의한다. 학습된 정책의 성능 평가를 위해서 단일 에피소드의 길이는 $T_{ep} = 100$ timesteps 으로 고려하며, 강화학습 개체는 무작위 위치에서 에피소드를 시작한다. 개별 개체는 단일 timestep에서 각 축 별로 $[-30m, 30m]$의 범위에서 행동을 수행할 수 있으며, 반경 $\delta = 550m$[21]의 관측 거리를 갖는다. 또한, 강화학습 개체의 주행 제한 높이를 $[200m, 1000m]$로 가정한다.

4.2 성능 평가 방법

모델의 성능 평가는 각 무인 항공기가 복구 지점까지 이동하는 기간을 충분히 고려하여 단일 에피소드의 마지막 ϵ timestep 동안 애드혹 네트워크가 구축 및 유지되었는지 평가한다. 해당 평가 지표는 다음의 식과 같이 정의 가능하다.

$$\text{hit ratio} = \frac{1}{T_{ep} - \epsilon} \sum_{k=T_{ep}-\epsilon}^{T_{ep}} r_k$$

hit ratio가 1에 가까울수록 높은 애드혹 네트워크 구

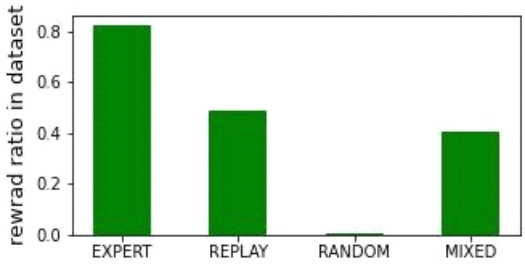

그림 2. 데이터 셋 별 보상 비율
Fig. 2. Reward ratio over each dataset

축 및 유지성능을 갖는다고 해석할 수 있으며, 본 연구에서 ϵ 는 20 timestep으로 고려한다.

4.3 오프라인 데이터셋

본 연구에서는 10^6개 분량의 학습 데이터를 담은 4종류의 데이터셋을 활용하여 오프라인 강화학습 실험을 진행하였다[22-23]. 전문가 데이터셋(EXPERT)의 경우 온라인 학습을 통해 사전 학습된 정책 모델을 기반으로 제작된다. 학습된 정책 모델의 경우 전문가에 가까운 수준으로 문제를 해결할 수 있으며, 이를 환경에 배치하여 상호작용 데이터를 수집한다. 리플레이 데이터셋(REPLAY)은 온라인 학습 과정에서 정책 학습을 위해 리플레이 버퍼에 저장된 데이터 셋을 의미한다. 즉, 온라인 강화 학습을 통해 사전 학습된 정책과 동일한 경험 데이터가 저장된 데이터 셋을 의미한다. 무작위 데이터셋(RANDOM)은 무작위 정책을 통해 제작한 데이터셋을, 합성 데이터셋(MIXED)은 EXPERT와 RANDOM을 50 : 50 비율로 합성한 데이터셋을 의미한다[19,24]. 각 데이터셋이 포함하고 있는 타임스텝 당 보상 비율은 그림 2와 같다. 4쌍의 그래프는 각 데이터셋을 구성하는 튜플 중 보상을 포함하는 튜플($r = 1$)의 비율을 나타낸다.

4.4 비교 알고리즘

1) BC[25]: BC는 주어진 데이터셋에 포함된 관측 행동 쌍을 모방하는 정책 학습을 위한 알고리즘으로 강화학습과 달리 정책 모델만을 포함한다. 해당 정책 모델의 손실함수는 다음처럼 정의된다.

$$L(\phi_i) = (a_i - \pi_\phi(o_i))^2$$

2) Offline Twin Delayed Deep Deterministic Policy Gradient (Offline TD3)[26]: Offline TD3 알고리즘은 온라인으로 고려되는 TD3 알고리즘을 오프라인 방식으로 변형한 알고리즘이다. 정책 및 Q 함수의 손실함수는 식 (1), (2)를 따른다.

3) IL[7]: 본 연구의 프레임워크를 고려하여 다중 개체의 독립적 네트워크 학습을 위해 확장되었으며, 손실함수는 본문에서 확인했던 것과 같이 식 (2)와 (3)을 따른다.

4) IQL[10]: 본 연구의 프레임워크를 고려하여 다중 개체의 독립적 네트워크 학습을 위해 확장되었으며, 손실함수는 본문에서 확인했던 것과 같이 식 (4), (5), (6)을 따른다.

4.5 모의실험 결과

그림 3은 오프라인 강화학습 알고리즘 별 데이터셋의 차이에 따른 다중 무인 항공기의 오프라인 강화학습 성능을 나타낸다. 각 그래프의 실선은 5개 시드의 평균값을 나타내며 음영은 2차 분산을 나타낸다. 데이터셋의 차이에 따라 설정된 색의 구분은 그림의 라벨을 통해 확인할 수 있다.

데이터셋 내 최적의 행동을 모방하는 IQL 및 BC 알고리즘은 EXPERT 데이터셋을 사용한 실험에서 높은 구축률을 보였으나, RANDOM 실험에서는 네트워크 구축에 실패하는 경향성을 보였다. 완전하게 오프라인 방법론에 특화된 알고리즘의 특성상 이러한 현상은 자연스럽다고 볼 수 있다. 반면 기존에 온라인 강화학습을 위해 사용되었던 TD3의 경우 EXPERT와 같이 제한

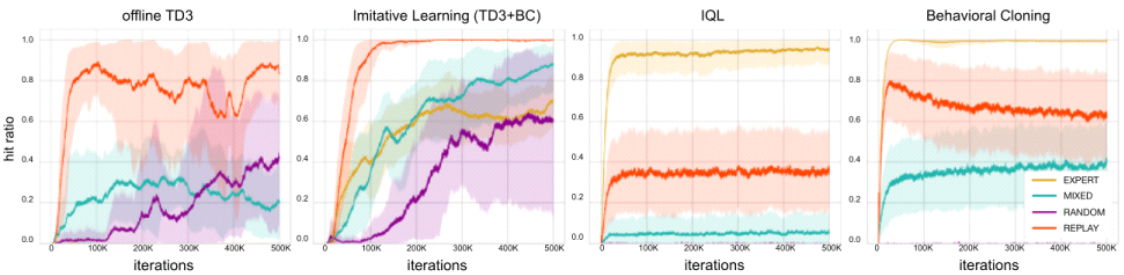

그림 3. 알고리즘 별 데이터셋에 따른 오프라인 강화학습 성능
Fig 3. Performance of offline reinforcement learning over each dataset per algorithm

표 1. 알고리즘 별 데이터 셋에 따른 최종 수렴 hit-ratio 성능
Table 1. Final hit-ratio performance over each dataset per algorithm

	Offline TD3	IL	IQL	BC
EXPERT	0.000 ±0.000	0.665 ±0.021	0.957 ±0.029	0.995 ±0.010
MIXED	0.221 ±0.076	0.905 ±0.066	0.041 ±0.024	0.463 ±0.071
RANDOM	0.442 ±0.178	0.657 ±0.313	0.000 ±0.000	0.000 ±0.000
REPLAY	0.755 ±0.250	1.000 ±0.000	0.426 ±0.047	0.681 ±0.092
Average	0.354	0.807	0.202	0.535

적인 데이터셋 내에서 제대로 성능을 보이지 못함을 확인하였다. 반면, 실제 온라인 강화학습의 탐험 과정 데이터셋을 그대로 포함시킨 REPLAY 데이터 셋에서는 높은 성능을 보였으며 RANDOM 및 MIXED 역시 EXPERT에 비해 개선된 성능을 보였다. 흥미로운 점은 온라인 강화학습에서 고려되던 방법론과 오프라인 학습을 위해 고려되는 방법론이 모두 고려된 IL의 경우 모든 데이터셋에서 준수한 성능을 거두는 것을 확인할 수 있었다. 각 알고리즘 별 최종 수렴 성능에 대한 수치적 값은 표 1을 통해 자세히 살펴볼 수 있다.

4.6 다중 개체의 애드혹 네트워크 정성적 구축 성능

그림 4는 IL 알고리즘과 REPLAY 데이터셋을 통해 학습된 다수의 무인 이동체의 애드혹 네트워크 구축 경로를 나타낸다. 각각의 3차원 공간은 고려하는 3개의 개체의 경로를 나타내며, 빨간색, 파란색, 초록색은 개별 에피소드를 의미한다. 이때 각 에피소드 경로에서 채도가 낮을수록 초기 궤적을 의미하며, 채도가 높을수록 에피소드의 후기 궤적을 의미한다.

그림 4의 결과를 통해 제안된 방식으로 학습된 개체가 시작 위치로부터 가장 가까운 네트워크 구축 지점으로 향하는 것을 확인할 수 있다. 이는 개체들이 특정 역할에 따라 정해진 위치로 이동하는 한정적인 개체가 아닌, 주변 상황에 따라 능동적으로 의사결정이 가능한 개체가 학습되었음을 의미한다. 즉, 본 결과를 통해 무인항공기에게 단순히 특정 위치를 찾아가는 능력이 아닌 상황에 맞는 의사결정 능력을 학습하였음을 확인할 수 있다.

V. 결 론

본 연구에서는 다중 무인항공기의 강화학습을 통한 FANET 구축 시나리오에서 오프라인 강화학습 알고리즘 및 데이터셋에 따른 네트워크 구축 성능 비교를 수행하였다. 네트워크를 구축 및 유지하는 수치인 hit ratio를 기준으로 모의실험을 수행한 결과, 탐험을 필요로 하는 온라인 알고리즘은 리플레이 버퍼 데이터셋을 사용하였을 때 최대 성능을 기록하였으며 오프라인 알고리즘은 전문가에 가까운 데이터셋을 사용하였을 때 높은 구축률을 보였다. 또한 온라인 및 오프라인 방법론을 적절하게 결합한 IL (TD3+BC) 알고리즘의 경우 고려된 전체 데이터셋에서 가장 우수한 평균 성능을 보임을 확인하였다. 종합적으로, 오프라인 강화학습 방법론은 온라인 강화학습 방법론에 비해 학습 측면에서 효율적이라고 할 수 있지만, 데이터 셋의 품질에 의해 성능이 크게 좌우될 수 있다.

References

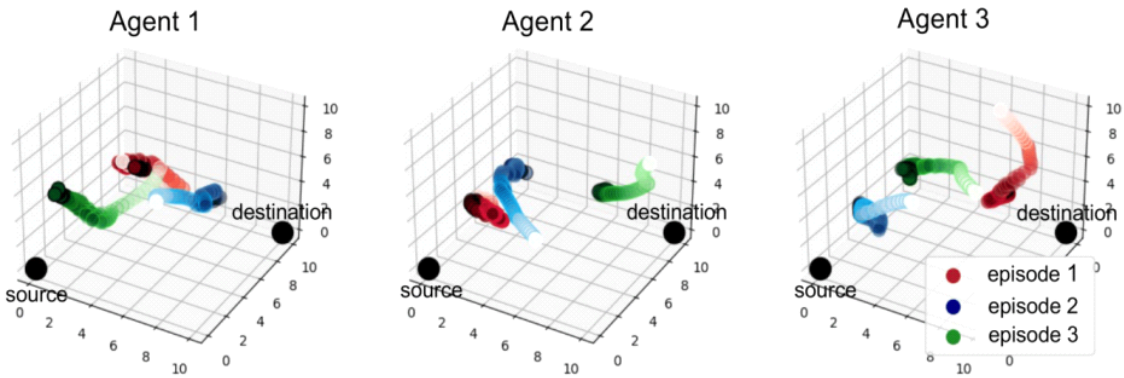

그림 4. 다중 개체의 애드혹 네트워크 구축 궤적
Fig. 4. Ad-hoc network building trajectory of multi-agent

[1] K. Sohrabi, et al., "Protocols for self-organization of a wireless sensor network," *IEEE Pers. Commun.*, vol. 7, no. 5, pp. 16-27, 2000. (https://doi.org/10.1109/98.878532)

[2] A. Chriki, et al., "FANET: Communication, mobility models and security issues," *Comput. Netw.*, vol. 163, 2019. (https://doi.org/10.1016/j.comnet.2019.106877)

[3] K. A. Awan, et al., "StabTrust－A stable and centralized trust-based clustering mechanism for IoT enabled vehicular ad-hoc networks," *IEEE Access*, vol. 8, pp. 21159-21177, 2020. (https://doi.org/10.1109/ACCESS.2020.2968948)

[4] F. Li, et al., "Hierarchical routing for vehicular ad hoc networks via reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1852-1865, 2018. (https://doi.org/10.1109/TVT.2018.2887282)

[5] X. Lyu, et al., "Contrasting centralized and decentralized critics in multi-agent reinforcement learning," *AAMAS*, 2021. (https://dl.acm.org/doi/10.5555/3463952.3464053)

[6] R. F. Prudencio, et al., "A survey on offline reinforcement learning: Taxonomy, review, and open problems," *IEEE Trans. Neural Netw. and Learn. Syst.*, 2023. (https://doi.org/10.1109/TNNLS.2023.3250269)

[7] S. Fujimoto, et al., "A minimalist approach to offline reinforcement learning," *NeurIPS*, vol. 34, pp. 20132-20145, 2021.

[8] A. Kumar, et al., "Conservative Q-learning for offline reinforcement learning," *NeurIPS*, vol. 33, pp. 1179-1191, 2020.

[9] S. Levine, et al., "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," *arXiv preprint arXiv: 2005.01643*, 2020. (https://doi.org/10.48550/arXiv.2005.01643)

[10] I. Kostrikov, et al., "Offline reinforcement learning with implicit Q-learning," *ICLR*, 2022.

[11] S. Gronauer, et al., "Multi-agent deep reinforcement learning: A survey," *Artificial Intell. Rev.*, vol. 55, no. 2, pp. 895-943, 2022. (https://doi.org/10.1007/s10462-021-09996-w)

[12] J. Foerster, et al., "Counterfactual multi-agent policy gradients," *AAAI*, vol. 32, no. 1, 2018. (https://doi.org/10.1609/aaai.v32i1.11794)

[13] C. Yu, et al., "The surprising effectiveness of PPO in cooperative multi-agent games," *NeurIPS*, vol. 35, pp. 24611-24624, 2022.

[14] R. Mu, et al., "Certified policy smoothing for cooperative multi-agent reinforcement learning," *AAAI*, vol. 37, no. 12, pp. 15046-15054, 2023. (https://doi.org/10.1609/aaai.v37i12.26756)

[15] Q. Tian, et al., "Learning from good trajectories in offline multi-agent reinforcement learning," *AAAI*, vol. 37, no. 10, pp. 11672-11680, 2023. (https://doi.org/10.1609/aaai.v37i10.26379)

[16] X. Liu, et al., "Reinforcement learning in multiple-UAV networks: Deployment and movement design," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8036-8049, 2019. (https://doi.org/10.1109/TVT.2019.2922849)

[17] D. Lee, et al., "Online curriculum reinforcement learning based UAV training for disaster network recover," *J. KISC*, vol. 49, no. 1, pp. 12-22, 2024. (https://doi.org/10.7840/kics.2024.49.1.12)

[18] N. Kim, et al., "Curriculum reinforcement learning for cohesive team in mobile ad hoc networks," *IEEE Commun. Lett.,* vol. 26, no. 8, pp. 1809-1813, 2022. (https://doi.org/10.1109/LCOMM.2022.3179235)

[19] J. Eo, et al., "The impact of dataset on offline reinforcement learning performance in UAV-based emergency network recovery tasks," *IEEE Commun. Lett.*, 2023. (https://doi.org/10.1109/LCOMM.2023.3339478)

[20] A. Shamsoshoara, et al., "Distributed cooperative spectrum sharing in UAV networks using multi-agent reinforcement learning," *CNCC*, pp. 1-6, 2019. (https://doi.org/10.1109/CCNC.2019.8651796)

[21] D. Aouladhadj, et al., "Drone detection and tracking using RF identification signals," *Sensors*, vol. 23, no. 17, pp. 7650-7673, 2023.

(https://doi.org/10.3390/s23177650)

[22] J. Fu, et al., “D4RL: Datasets for deep data-driven reinforcement learning,” *arXiv preprint arXiv:2004.07219*, 2020. (https://doi.org/10.48550/arXiv.2004.07219)

[23] D. Lee, et al., “AD4RL: Autonomous driving benchmarks for offline reinforcement learning with value-based dataset,” *IEEE ICRA*, 2024.

[24] K. Schweighofer, et al., “Understanding the effects of dataset characteristics on offline reinforcement learning,” *NeurIPS Deep RL Workshop*, 2021.

[25] F. Torabi, et al., “Behavioral cloning from observation,” *IJCAI*, 2018.

[26] S. Fujimoto, et al., “Addressing function approximation error in actor-critic methods,” *ICML*, pp. 1587-1596, 2018.

이 동 수 (Dongsu Lee)
한국통신학회논문지 vol 48, no 11 참조
[ORCID:0000-0002-9238-4106]

권 민 혜 (Minhae Kwon)
한국통신학회논문지 vol 48, no 11 참조
[ORCID:0000-0002-8807-3719]