2020년 한국항해항만학회 추계학술대회 및 ICMASS 2020
2020년 11월 11일(수)~13(금) 온라인 웨비나(한국해양대학교)

선박항해 에이전트 학습을 위한 보상설계 방안에 관한 연구

† 박세길 · 오재용*

*,† 선박해양플랜트연구소

A Study on the Development of Learning Environment for Ship Navigation Agents

† *Sekil Park · Jaeyong Oh**

*,† *Korea Research Institute of Ships & Ocean Engineering (KRISO)*

요 약 : 본 논문은 선박항해 에이전트가 개발 의도와 부합되도록 학습시키는데 있어 가장 중요한 역할을 수행하는 보상설계에 대해 소개한다. 보상설계는 먼저 학습 대상이 무엇인지 명확히 정의하는 것이 중요하며, 보상이 상황에 따라 다른 목적으로 활용되지 않도록 하고 에이전트에게 너무 드물게 주어지지 않도록 보상 형태화를 적용하는 등의 방법을 사용할 필요가 있다. 또한 보상을 구성하는 요소가 많아지는 경우에는 의도가 명확하게 전달이 되지 않을 수 있으므로 문제를 작은 문제들로 나누어 접근하는 계층적 강화학습 방법 등을 적용할 필요가 있다.

핵심용어 : 선박항해, 에이전트, 강화학습, 인공지능, 자율운항

1. 서 론

강화학습은 기계학습의 한 영역으로 에이전트가 주어진 환경의 현재 상태를 인식하고, 이를 바탕으로 선택 가능한 행동들 중 보상을 최대화하는 행동을 결정해 나가는 방법이다. Fig. 1에서 보는 것과 같이 보상은 현재 상태에서 수행한 행동의 결과에 따라 다음 상태에서 환경으로부터 받게 된다. 에이전트는 이러한 과정을 통해 경험을 축적해 가면서 점차 최적의 행동 정책을 찾아낸다.

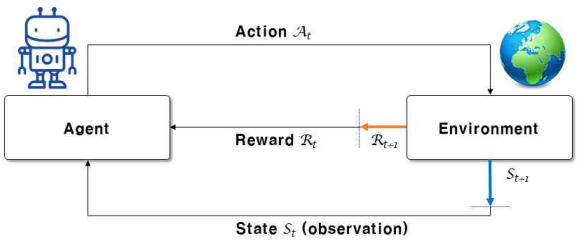

Fig. 1 강화학습의 개념

강화학습은 이처럼 비교적 간단한 원리로 문제 해결이 가능하나 관측해야할 상태와 선택해야할 행동의 개수가 많아질 경우 차원의 저주 문제가 발생하여 적용이 어렵다. 다행히 전통적인 강화학습에 심층 신경망을 적용한 심층 강화학습의 등장과 발전으로 게임과 같은 복잡한 규칙과 상태를 가지는 문제들을 해결할 수 있게 되었고, 실세계 문제 해결을 에이전트와 에이전트의 학습을 위한 다양한 환경들이 연구되고 있다 (Mnih, 2013, Park, 2019). 이러한 환경의 구조 자체는 OpenAI Gym의 사례에서 보듯이 매우 간단하다 (Brockman, 2016). 그러나 에이전트가 해결할 문제의 여러 상황들을 어떠한 상태들로 표현하고 각 상태들에서 선택 가능한 행동들 중에서 어떤 행동이 좋고 나쁜지에 대해 계획하는 것은 매우 논리적인 사고가 필요하며 난해하다. 단순하게 생각했을 때 좋은 행동과 나쁜 행동을 구분하고 그에 따라 보상을 주면 되지만 체계적이고 치밀하게 검토되지 못할 경우 에이전트에게 잘못된 자극이 주어져 코브라 효과를 경험할 수 있다.

본 논문에서는 선박항해 에이전트의 개발 초기 과정에서 경험한 보상 설계 방안과 결과에 대해 소개하고자 한다.

2. 선박항해 에이전트 보상설계

보상 함수는 풀고자하는 문제에 맞춰 수치적(단일 스칼라 값)으로 표현되어야 하며, 에이전트는 수많은 시행착오를 통해 보상을 최대화하는 정책을 찾아야 한다. 보상설계는 에이전트에게 학습 시키고자 하는 올바른 행동을 말이나 글로써 알려주는 것이 아니라 보상이라는 단일 스칼라 값을 통해 에이전트에게 알려줄 수 있도록 환경의 일부로서 설계하는 것을 말한다. 올바른 보상설계를 위해서는 우선 학습 대상을 의도한 목적에 맞게 명확하게 정의하는 것이 필요하다. Fig. 2에서와 같이 선박이 ⓧ로 표시한 목적지에 도달하도록 학습을 시키려 할 경우 왼쪽과 같이 목적지에 맞춰 헤딩을 정확하게 맞추는 것에 대해 보상을 줄 수도 있고 오른쪽과 같이 외력이 존재하여 헤딩과 코스가 달라지는 것을 고려하여 코스를 정확하게 맞추는 것에 대해 보상을 줄 수도 있다. 두 가지 경우 모두 목적지에 도달하도록 학습이 되기는 하나 전자와 같이 학습할 경우는 외력이 존재하면 실제 항해사들의 운항 궤적과 달리 나선형을 그리며 목적지에 도달하는 비효율적인 학습이 이뤄진다. 따라서 보상설계 시 전문가가 어떻게 행동하는지를 면밀히 분석하고 에이전

트가 그에 맞게 학습하도록 보상을 설계하는 것이 필요하다.

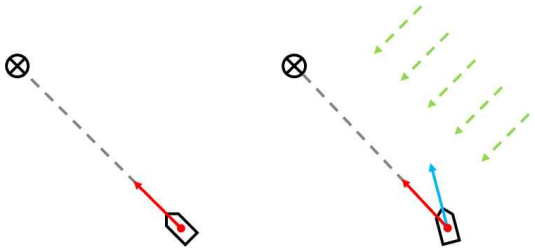

Fig. 2 학습 대상에 따른 결과 차이

다음으로 보상설계 시 반드시 고려해야할 사항으로 드문 보상에 대한 부분이 있다. 드문 보상은 에이전트가 학습을 수행하는 동안 보상이 너무 드물게 발생하여 학습이 제대로 이뤄지지 않는 결과를 만든다. 이를 해결하기 위해서는 Fig. 3에서 보는 것과 같이 보상 형태화(shaping) 기법을 도입할 필요가 있다 (Ng, 1999).

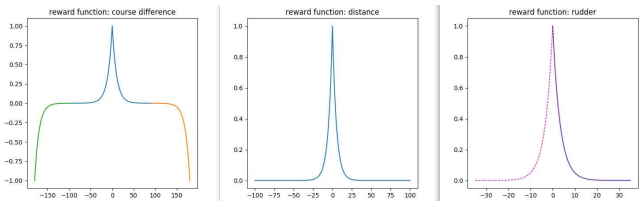

Fig. 3 보상 형태화 예시

보상 형태화는 Fig. 2의 사례에서 선박의 헤딩이 목적지 방향과 일치할 때 한 번의 보상을 주는 것보다 선박의 헤딩이 목적지 방향과 적당히 가까워지면서부터 보상을 주기 시작하여 가까워질수록 보다 높은 보상을 주도록 설계하는 것이 좋은데, 이는 드물게 발생하는 보상을 특정 형태의 함수로 표현하여 보다 자주 발생하도록 하는 방법이다.

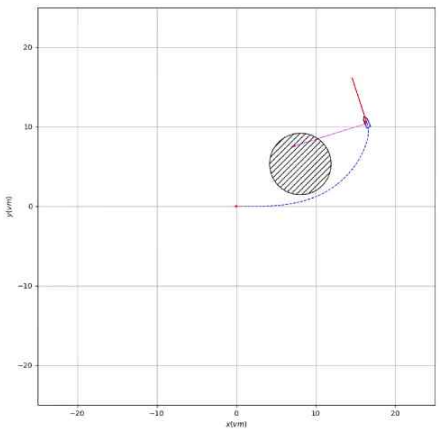

Fig. 4 장애물 회피 시나리오 예시

학습 결과를 검토하다보면 의도치 않게 학습이 된 경우를 자주 경험하게 된다. 선박항해 에이전트의 경우 단순하게 목적지에 도달하는 것이 목적이 아니라 선박의 운동특성도 학습해야 하고 Fig. 4에서와 같이 장애물도 회피해야 한다면 보상 함수는 점점 복잡해지게 된다. 또한 보상 함수를 구성하는 요소가 많아지면 요소별 특성과 요소 간 연관 등의 문제로 논리적인 오류가 발생하여 의도하지 않은 결과가 나오기 쉽고 상황에 따라 학습이 전혀 이뤄지지 않기도 한다. 이 경우에는 해결해야할 문제를 보다 작은 문제들로 나누어 계층적 강화학습을 적용하는 것을 고려해 볼 수 있다. 계층적 강화학습에서의 보상설계는 상, 하위 보상들을 어떻게 설계하고 환경에 반영할지에 대한 또 다른 형태의 어려움은 있으나 드문 보상과 보다 복잡한 형태를 가지는 문제를 해결할 수 있도록 도와준다.

3. 결 론

보상설계는 에이전트가 사용자가 의도한대로 학습이 되도록 하는데 목적이 있으며, 환경 개발에 있어 가장 중요한 부분이다. 본 연구에서는 선박항해 에이전트를 위한 보상 설계 방안에 대해 초기 개발 과정에서의 결과와 함께 소개하였다. 향후 전문가의 행동으로부터 학습이 가능하도록 지도학습이나 역강화학습 기법을 적용하는 방법에 대해 연구를 진행하고자한다 (Ng, 2000).

후 기

본 논문은 선박해양플랜트연구소의 주요사업인 “해상교통 분석을 위한 에이전트 모델링 및 연동 기술 개발(2/5)”에 의해 수행되었습니다(PES3600).

참 고 문 헌

[1] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602.

[2] Sekil Park, Jaeyong Oh, Hye-Jin Kim (2019), The Analysis of Reinforcement Learning Environment for Intelligent Ship Navigation Agents, Proceedings of the Korean Institute of Navigation and Port Research Conference (pp. 3-4)

[3] Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., & Zaremba, W. (2016). Openai gym. arXiv preprint arXiv:1606.01540

[4] Ng, Andrew Y., Daishi Harada, and Stuart Russell. "Policy invariance under reward transformations: Theory and application to reward shaping." ICML. Vol. 99. 1999.

[5] Ng, Andrew Y., and Stuart J. Russell. "Algorithms for inverse reinforcement learning." ICML. Vol. 1. 2000.