KIPS Trans. Softw. and Data Eng.
Vol.11, No.8 pp.331~338
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2022.11.8.331

Evaluating SR-Based Reinforcement Learning Algorithm Under the Highly Uncertain Decision Task

Kim So Hyeon† · Lee Jee Hang††

ABSTRACT

Successor representation (SR) is a model of human reinforcement learning (RL) mimicking the underlying mechanism of hippocampal cells constructing cognitive maps. SR utilizes these learned features to adaptively respond to the frequent reward changes. In this paper, we evaluated the performance of SR under the context where changes in latent variables of environments trigger the reward structure changes. For a benchmark test, we adopted SR-Dyna, an integration of SR into goal-driven Dyna RL algorithm in the 2-stage Markov Decision Task (MDT) in which we can intentionally manipulate the latent variables – state transition uncertainty and goal-condition. To precisely investigate the characteristics of SR, we conducted the experiments while controlling each latent variable that affects the changes in reward structure. Evaluation results showed that SR-Dyna could learn to respond to the reward changes in relation to the changes in latent variables, but could not learn rapidly in that situation. This brings about the necessity to build more robust RL models that can rapidly learn to respond to the frequent changes in the environment in which latent variables and reward structure change at the same time.

Keywords : SR Based Reinforcement Learning Algorithm, 2-Stage Markov Decision Task, State Transition Probability, Reward Function

불확실성이 높은 의사결정 환경에서 SR 기반 강화학습 알고리즘의 성능 분석

김 소 현† · 이 지 항††

요 약

차기 상태 천이 표상(Successor representation, SR) 기반 강화학습 알고리즘은 두뇌에서 발현되는 신경과학적 기전을 바탕으로 발전해온 강화학습 모델이다. 해마에서 형성되는 인지맵 기반의 환경 구조 정보를 활용하여, 변화하는 환경에서도 빠르고 유연하게 학습하고 의사결정 가능한 자연 지능 모사형 강화학습 방법으로, 불확실한 보상 구조 변화에 대해 빠르게 학습하고 적응하는 강인한 성능을 보이는 것으로 잘 알려져 있다. 본 논문에서는 표면적인 보상 구조가 변화하는 환경뿐만 아니라, 상태 천이 확률과 같은 환경 구조 내 잠재 변수가 보상 구조 변화를 유발하는 상황에서도 SR-기반 강화학습 알고리즘이 강인하게 반응하고 학습할 수 있는지 확인하고자 한다. 성능 확인을 위해, 상태 천이에 대한 불확실성과 이로 인한 보상 구조 변화가 동시에 나타나는 2단계 마르코프 의사결정 환경에서, 목적 기반 강화학습 알고리즘에 SR을 융합한 SR-다이나 강화학습 에이전트 시뮬레이션을 수행하였다. 더불어, SR의 특성을 보다 잘 관찰하기 위해 환경을 변화시키는 잠재 변수들을 순차적으로 제어하면서 기존의 환경과 비교하여 추가적인 실험을 실시하였다. 실험 결과, SR-다이나는 환경 내 상태 천이 확률 변화에 따른 보상 변화를 제한적으로 학습하는 행동을 보였다. 다만 기존 환경에서의 실험 결과와 비교했을 때, SR-다이나는 잠재 변수 변화로 인한 보상 구조 변화를 빠르게 학습하지는 못하는 것으로 확인 되었다. 본 결과를 통해 환경 구조가 빠르게 변화하는 환경에서도 강인하게 동작할 수 있는 SR-기반 강화학습 에이전트 설계를 기대한다.

키워드 : SR기반 강화학습 알고리즘, 2단계 마르코프 의사결정 과제, 상태 천이 확률, 보상함수

1. 서 론

강화학습(Reinforcement learning; RL) 에이전트는 환경과 상호작용하면서 보상을 최대화하는 전략을 학습하고 최적 의사결정을 도모한다[1]. 최근 빠르게 발전한 딥러닝이 강화학습 알고리즘에 적용되고, 강화학습에 대한 신경과학적인 증거들이 계속 보고되면서, 신경과학-인공지능 융합형 심층 강화학습 에이전트는 게임, 자율주행, 로봇 등 다양한 분야에서 전문가의 능력을 넘어서는 성취를 보이고 있다[2-4]. 그럼에

※ 이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2020R1G1A1102683). 본 연구는 삼성미래기술육성센터의 지원을 받아 수행하였음(No. SRFC-TC1603-52).
※ 이 논문은 2021년 한국정보처리학회 ACK 2021의 우수논문으로 "2-stage 마르코프 의사결정 상황에서 Successor Representation 기반 강화학습 알고리즘 성능 평가"의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 상명대학교 지능정보공학과 석사과정
†† 비 회 원 : 상명대학교 휴먼지능정보공학과 조교수
Manuscript Received : December 30, 2021
First Revision : February 15, 2022
Accepted : February 22, 2022
* Corresponding Author : Lee Jee Hang(jeehang@smu.ac.kr)

도 불구하고, 심층 강화학습 에이전트는 인간의 강화학습에 비해 환경 변화에 대한 유연성과 적응성 측면에서 아쉬운 성능을 보이고 있다[1, 5].

계산신경과학 연구에 따르면, 인간의 강화학습은 모델 기반 강화학습과 모델 프리 강화학습 두 가지 시스템을 이용하여 행동을 결정하고 제어한다고 알려져 있다. 이 두 시스템은 전두엽 메타 제어를 통해 중재되어, 환경 변화에 대해 빠르게 대응하고 적응하여, 새로운 환경에 처해서도 높은 성능을 일정하게 유지하도록 한다[5, 6].

모델 기반 강화학습 전략은 에이전트가 환경에 대해 먼저 학습을 수행하고, 학습한 환경 정보를 활용하여 계획을 수립한 후 문제를 해결하는 방식이다. 환경 구조와 상태 천이 모델을 먼저 학습하고 전략을 수립하는 방식을 취하는 바, 빠르게 변화하는 불확실한 환경에 처했을 때, 환경 변화에 기민하게 대응할 수 있고 더 주의 깊고 정확한 예측이 가능하다. 다만, 연합, 학습 및 계획을 위한 인지 부하가 요구되는 바, 계산량이 많고 복잡한 특징이 있다[5].

이에 반해 모델 프리 강화학습 전략은 환경 정보를 고려하지 않고 학습하고 행동을 제어하는 강화학습 전략이다. 어떤 상황에 처해있든지, 모델 프리 강화학습 에이전트는 환경에 대해 학습하는 과정 없이, 행동에 대한 보상 신호만을 통해 문제 해결을 위한 최적 전략을 수립한다. 따라서, 이를 학습하는 데 많은 경험이 필요한 바 시간이 오래 걸리고, 환경이 바뀌는 경우에는 학습한 전략이 유효하지 않아 환경 변화에 유연한 대처가 어렵다는 단점이 있다[7]. 그러나, 한 번 학습한 전략은 계획 과정 없이 빠른 실행이 가능하므로, 익숙한 환경에서는 빠른 속도로 일정한 성능을 유지할 수 있는 장점이 있다[5].

인공지능 에이전트를 설계할 때, 일반적으로 에이전트는 환경에 대한 정보를 모두 접근할 수 없는 상황을 가정한다. 따라서, 기존 강화학습 알고리즘 연구는 환경 정보가 부족하더라도 경험에 대한 보상만으로 최적 전략을 학습할 수 있는 모델 프리 강화학습 위주로 진행되었고, 제한적으로 모델 기반 강화학습을 차용하는 접근을 취해왔다. 그러나 최근 신경과학의 활발한 연구를 통해, 인간의 강화학습에 대해 많은 발견이 이루어지고, 인간과 기계의 강화학습 방법의 차이가 빠르게 규명되면서, 뇌과학적 발견을 적극적으로 반영한 메타 강화학습 알고리즘[8], 혹은 강화학습의 중추 도파민과 연관된 두뇌의 정보처리 기전을 모사한 IQN[9] 등 신경과학-인공지능 융합형 강화학습 알고리즘 연구들이 활발히 진행되고 있다[10].

본 논문에서는 기존 연구[11]에 이어 조명되는 차기 상태 천이 표상 (Successor Representation, 이후 SR) 기반 강화학습 알고리즘을 이용하여 다양한 환경 변화에 따른 강화학습 에이전트의 성능을 평가하였다. SR은 두뇌 내 해마의 공간 세포가 인지맵을 구성하여 환경, 특히 공간 정보를 학습하고, 이 정보를 바탕으로 변화하는 환경에서도 유연하게 최적 전략을 수립하는 기전을 모사한 강화학습 알고리즘이다. 이전에 학습한 환경/공간 정보와 유사한 환경 구조 내에서, 목표와 보상이 변화하더라도 이전에 학습한 환경 정보를 활용하여 최적 전략을 빠르게 찾아, 모델 프리 강화학습 알고리즘들에 비해 빠르게 보상 변화에 적응하는 것으로 잘 알려져 있다. 다만, 환경 구조 자체에 대한 변화나 환경 내 상태 천이 불확실성에 대응하여 보상 변화가 유발되었을 때, 이에 유연하게 최적 전략을 수립하는 연구는 상대적으로 부족한 상황이다.

따라서 본 연구에서는 환경 내 잠재 변수가 변화하여 보상 구조를 변화하는 환경, 예를 들어 상태 천이 확률이 변화하고, 이와 연계되어 보상 함수가 동시에 변화하는 상황에서 SR 기반 에이전트의 성능을 분석하고자 한다. 또한, 각 환경 내 잠재 변수의 변화 요인을 제어했을 경우 SR 기반 강화학습 에이전트의 성능이 잠재 변수 제어에 따라 변화하는 패턴을 분석하고자 한다. 이는 기존 연구에서 SR 기반 강화학습 에이전트가 해당 환경을 학습하지 못한 것에 대한 추가적인 분석을 실시하여 SR 기반 강화학습 알고리즘의 취약점을 규명할 수 있는 확장 연구라고 볼 수 있다.

기존 연구와 동일하게 SR 기반 알고리즘으로는 SR의 특징을 목적 기반 강화학습으로 통합한 SR-다이나를 사용하였고, 실험 환경은 2단계 마르코프 의사결정 과제를 선택하였다[5]. 본 환경은 보상 함수뿐만 아니라 환경 불확실성을 결정하는 상태 천이 확률 및 목표가 변화하는 환경으로, 보상 함수 변화에 강인한 SR 기반 강화학습 알고리즘이 어떠한 환경 구조의 변화에 강인 혹은 취약한지 분석하는 데 좋은 벤치마크 환경이 될 것으로 예상된다.

2. SR 기반 강화학습 모델: SR-다이나

2.1 SR 개요

동물들의 최적 행동 패턴은 보통 그 동물이 현재 있는 위치(혹은 상태)에 따라 결정되는 경향이 강하다. 해마의 공간 세포는 동물이 경험한 환경에 대한 공간 및 위치 정보를 담고 있는데, 보통 동물들은 해마의 공간적 정보를 미래 상태를 예측하는 데에 사용하여 일련의 최적 행동을 보인다고 알려져 있다.

SR 모델은 최적 행동을 수립하기 위해 참조하는 해마 내 공간 정보를 지칭한다고 볼 수 있다. 특히 SR 모델은, 해마의 공간 세포는 단순히 위치 정보를 인코딩하는 것이 아니라, 현재 상태에서 이동할 가능성이 가장 큰 다음 상태를 표현하는 '차기 상태 예측 표상'을 인코딩 한다는 가정을 바탕으로 한다. 따라서, SR 모델에 따르면, 물리적으로 인접한 두 개의 공간 세포가 각각 다른 차기 상태를 예측하고 있다면, 이 두 공간 세포는 서로 다른 차기 상태 예측 표상을 가지고 있을 것으로 가정한다. 만일 예제의 두 공간 세포가 동일한 차기 상태를 예측하고 있다면, 두 공간 세포는 당연히 비슷한 차기 상태 예측 표상을 보일 것이다[12].

SR 모델은 차기 상태 예측 표상을 결정하기 위해서 강화학습에서 자주 사용되는 가치(V)를 통해 형성된다. 강화학

습에서 현재 상태 s의 가치는 보통 미래에 방문할 모든 상태 s_t에서 예측되는 보상의 총 합으로 정의되는데, 감가율 $\gamma \in [0,1]$를 고려하여 다음과 같이 정의한다.

$$V(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R(s_t) | s_0 = 0] \quad (1)$$

여기서, s_t는 시간 t에서 방문한 상태s를 의미한다.

SR에서 상기 가치 함수는 차기 상태 예측 표상M과 보상 함수의 내적으로 나눌 수 있고 다음과 같이 표현이 가능하다.

$$V(s) = \sum_{s'} M(s,s') R(s') \quad (2)$$

이때, SR은 초기 상태 s로부터 진행된 일련의 상태 방문 과정에서 미래 상태 s'를 방문하여 점유할 기댓값으로 볼 수 있다.

$$M(s,s') = \mathbb{E}[\sum_{t=0} \gamma^t \mathbb{I}(s_t = s') | s_0 = s] \quad (3)$$

여기서 만일 $s_t = s'$이면 $\mathbb{I}(s_t = s') = 1$이고, 그렇지 않으면 0이다.

이로써, 가치 함수 내 감가 예측 보상 기댓값[Equation (1)]은 감가 예측 상태 점유 기댓값과 그 상태에서의 보상으로 분할할 수 있다[Equation (2)]. 이 때 M으로 표시된 SR은 미래에 점유할 상태를 예측하기 위한 정보로써, 현재 상태에서 액션을 수행하여 (i) 다음 상태로 이동했을 때 획득 가능한 보상의 예측 값과 (ii) 점유가 예측되는 다음 상태 값을 바탕으로 시간차(Temporal difference) 학습 알고리즘[13]을 통해 구한다[14].

Fig. 1은 모델 프리 강화학습, 모델 기반 강화학습, 그리고 SR의 가치 추정 업데이트 기전을 통해 SR의 특성을 보여주고 있다. V(s)는 상태 s에 대한 가치 추정치를 의미한다. 모델 프리 강화학습 알고리즘의 가치 함수 추정치는 공간 좌표 x, y에 대한 선형 가치함수 근사로 표현되었고, w_x는 x 좌표일 때의 가중치, w_y는 y 좌표일 때의 가중치를 나타낸다. 모델 기반 강화학습 알고리즘 수식에서 등장하는 T(s,s')은 상태 천이 확률을, γ는 미래의 가치를 현재의 가치로 환산하는

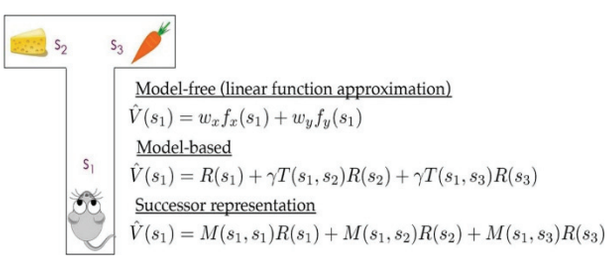


Fig. 1. How Model Free, Model-based and SR-based Reinforcement Learning Algorithms Compute Values in a Simple Maze with Three States(S1,S2,S3) (© Gershman, 2018 [14])

감가율을 나타낸다. SR 알고리즘 수식에서의 M(s,s')은 SR-매트릭스를 의미하고 M(s,s')은 Equation (3)과 같이 정의된다. 마지막으로 아래 두 수식에서 공통적으로 등장하는 R(s)는 각 상태에 대한 보상 값을 의미한다.

모델 기반 강화학습 알고리즘의 경우 에이전트와 직접적으로 연관성이 있는 보상 R과 상태 천이 확률 T를 모두 고려하기 때문에 환경 변화에 대해 유연하게 대처할 수 있지만, 모든 정보를 다루어야 하므로 계산량이 많아 학습 진행이 느릴 수 있다. 반면, 모델 프리 강화학습 알고리즘은 현재의 위치 s_1이 주어졌을 때, 에이전트가 취할 수 있는 가치 값들을 확인하고 이 정보만을 이용해 다음 행동을 결정한다. 여기서도 알 수 있듯이 모델 프리 강화학습 알고리즘은 모델 기반 강화학습 알고리즘에 비해 계산량이 적어 빠르게 학습이 가능하지만, 환경 정보를 반영하는 상태 천이 확률 T에 대한 고려가 없어, 환경 변화에 따른 유연한 대처가 어렵다는 특징이 드러난다. 마지막으로 SR은 여러 가지 다른 조합을 가진 요소들을 그 때의 선호(정책)에 따라 경로를 평가할 수 있다. 부분적으로 계산된 행동 가치 값과 환경 정보를 모두 고려하여 최적 정책을 수립하기에 모델 기반과 모델 프리 강화학습 알고리즘을 모두 고려하여 근사화한 중간 계열의 알고리즘이라고 할 수 있다[15].

2.2 잠재학습 바탕 SR-매트릭스 형성 사례

SR의 특성은 동일한 환경 구조 내에서 보상 함수가 변화하는 격자 구조 환경에서 보다 잘 확인할 수 있다[16]. 본디, 에이전트가 미로를 탐색할 때 설치류의 여러 기본 능력과 유사한 행태를 보이는데, 이 능력은 인지맵에 기인한다고 볼 수 있다. 인지맵은 해마에서 인코딩되며 이는 잠재 학습 문제와 장애물 우회 문제에 모두 빠르게 적응할 수 있도록 도와준다. 잠재 학습이란, 어떠한 보상이 주어지지 않더라도 따라야하는 단계의 반복적인 행동을 통해 잠재적인 학습이 일어난다는 것으로 학습 초반, 보상을 주지 않고 환경을 충분히 탐색하도록 한 뒤 특정한 위치에 보상이 주어졌을 때 약간의 탐색 과정만으로도 목표 위치에 도달할 수 있는지를 보는 문제이다 장애물 우회 문제란 시작 지점으로부터 목표 지점까지의 지름길을 학습한 뒤 해당 경로에 장애물이 생겼을 때 즉, 환경 구조가 변화했을 때 목표 지점까지 도달할 수 있는지를 보는 문제이다[17].

Evan et al. 연구는 SR을 이용하여 잠재학습을 통해 보상 변화에 강인한 SR 기반 강화학습 에이전트를 제안하였다[16]. 모델 기반 강화학습 알고리즘 dyna[18]에 SR 개념을 적용한 SR-다이나를 제안하고, 잠재 학습 문제와 장애물 우회 문제를 해결하였다. Fig. 2는 실험에 사용한 환경 구조를 보여준다. SR-다이나는 Fig. 2-A와 같은 환경구조를 먼저 학습하여 SR을 형성한 다음, 보상 R 위치가 변경되었을 때 기존에 학습한 SR을 바탕으로 변경된 보상을 획득하는 과정을 수행하였다. Fig. 2-B는 장애물 우회 문제 환경이다. 초기 SR-다이나가 그림의 B 위치에 장애물이 없는 상황에서 환경

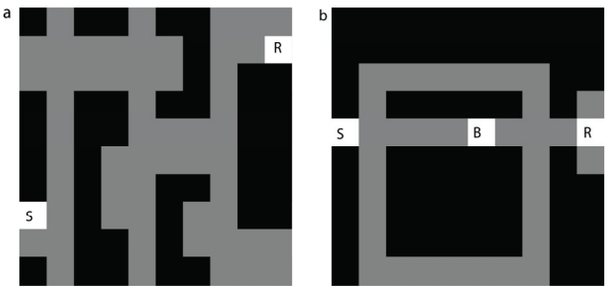


Fig. 2. Grid-world Representation of Tolman's Latent Learning Tasks. A is for Latent Learning, and B is Designed to Solve Detour Tasks after Learning Shortcut. (© RUSSEK, Evan M., et al., 2017[16])

구조를 학습하여 SR을 형성하고, 이후 B 위치에 장애물이 생겼을 때, 이를 우회하여 기존 R위치에 존재하는 보상을 획득하는 과정을 보였다. 종합해보면, SR 기반 강화학습 에이전트는 보상 구조에 영향을 미치는 환경 구조가 미세하게 변하는 경우, 기존 형성한 환경 구조를 바탕으로 약간의 학습만으로도 잘 대처하는 결과를 보였다.

본 논문에서는 이를 고려하여 보상 함수가 변화하는 환경에서도 구조를 잘 배우며, 모델 기반과 모델 프리 강화학습의 행동이 분리되는 시나리오에서도 강인한 성능을 보일 수 있을 것이라고 기대되는 SR-다이나를 본 연구의 모델로 삼았다.

3. 환경: 2단계 마르코프 의사결정 과제

보상 함수 변화에 강인한 SR-다이나 알고리즘이 어떠한 환경 구조의 변화에 성능 차이를 보이는지를 확인하기 위해 2단계 마르코프 의사결정 과제[6] (Fig. 3)에서 시뮬레이션을 수행하였다.

의사결정 신경과학에서 사용한 2단계 마르코프 의사결정 과제 환경은 초기 상태 S_1에서 에이전트가 두 번의 이동 액션(Left/Right)을 통해 최종 상태로 이동하고, 최종 상태(S_6-S_9)와 연계된 동전에 기입된 보상을 받는 환경이다. 성취 조건에 따라 상태 천이 확률이 조정되는데, 특정 성취 조건의 경우 에이전트의 상태 천이 확률은 (0.9, 0.1)의 확률로 결정되어 에이전트가 Left를 선택했을 경우, 선택한 액션은 90%의 확률로 수행되고, 10%의 확률로 반대로 수행됨을 뜻한다. 특정

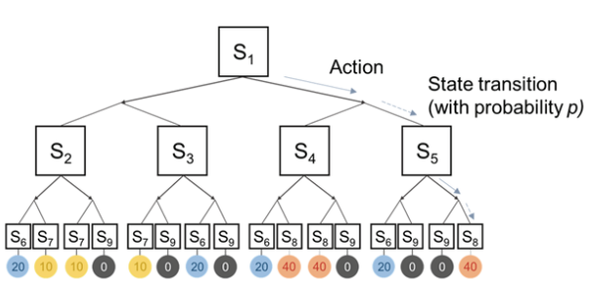


Fig. 3. 2-stage Markov Decision Task[6]

성취 조건에서는 수집할 동전의 색깔이 제시된다. 이때 에이전트는 반드시 제시된 색깔의 동전을 수집해야만 동전에 기록된 보상을 받을 수 있고, 만일 다른 색깔의 동전을 수집하면 동전에 기록된 보상은 주어지지 않는다. 반면, 자유 성취 조건의 경우 어떠한 색깔의 동전을 수집하더라도 동전에 기록된 보상만큼을 획득할 수 있다. 다만, 상태 천이 확률은 (0.5, 0.5)로 설정되어, 에이전트의 이동 액션 선택을 따르지 않고 무작위로 선택된다는 특징이 있다.

일반적으로 성취 조건이 특정 성취 조건인 경우 환경 불확실성이 낮기 때문에 환경 구조를 배우기 용이하고, 상태 예측이 가능한 상황이다. 따라서 이 상황에서는 모델 기반 강화학습 전략이 선호된다. 반면, 자유 성취 조건의 경우에는 환경 불확실성이 매우 높아 환경 구조를 배우기 어려워 상태 예측 또한 매우 어렵다. 이 상황에서는 행동에 따른 보상을 추구하는 방식인 모델 프리 강화학습 전략이 더 적합한 학습 전략이라고 볼 수 있다. 이렇게 환경 변수에 따라 환경 구조가 변화하고 보상 값이 변하는 환경에서 SR 기반 강화학습 알고리즘의 성능 평가를 시도하였다.

4. 실험 결과 및 고찰

4.1 높은 환경 불확실성 설정에서 SR-다이나 성능 평가

여기서는, SR-다이나가 상태 천이 확률과 보상 함수가 동시에 변화하는 환경에서 어떤 요소를 학습할 수 있는지를 확인하고자 한다. 이를 위해, 두 가지 버전의 SR-다이나를 구성하였고, 각각을 실험군/대조군으로 정의하였다. 실험은 SR 기반 강화학습 에이전트가 2단계 마르코프 의사결정 과제를 총 500 게임 수행하도록 하였다. 매 50 게임마다 각 성취 조건(goal condition)이 바뀌도록 설정하였다. 즉 500 게임을 수행하는 동안 10번의 성취 조건 변화가 발생하고 이를 1 세션이라 정의하였다. 에이전트가 획득한 보상을 0-1 사이의 값으로 정규화하였다.

기존 SR-다이나 알고리즘은 표계산 방식으로 Equation (3)에서 보여주는 바와 같이, 하나의 SR-매트릭스 M에서 상태 가치 값이 계산되고, 보상이 주어졌을 때 이 SR-매트릭스를 기반으로 전략을 수립한다. 보상 변화가 발생하거나 상태 변화가 발생해도, 이 하나의 M을 기반으로 적응하는 방식을 취하게 된다. 이를 실험군으로 설정하고, 2단계 마르코프 의사결정 과제를 학습하도록 하였다.

기존 SR-다이나 알고리즘은 표계산 방식으로 Equation (3)에서 보여주는 바와 같이, 하나의 SR-매트릭스 M에서 상태 가치 값이 계산되고, 보상이 주어졌을 때 이 SR-매트릭스를 기반으로 전략을 수립한다. 보상 변화가 발생하거나 상태 변화가 발생해도, 이 하나의 M을 기반으로 적응하는 방식을 취하게 된다. 이를 실험군으로 설정하고, 2단계 마르코프 의사결정 과제를 학습하도록 하였다.

대조군으로는 2단계 마르코프 의사결정 과제를 수행할 때, 각 성취 조건별로 SR-매트릭스를 학습할 수 있는 에이전트를

구현하였다. 다시 말해, 특정 성취 조건 환경에서만 학습하고 사용하는 $M_{specific}$과 자유 성취 조건 환경에서만 학습하고 사용하는 $M_{flexible}$을 준비하였다. 이 에이전트는 모든 환경 변화를 관찰할 수 있고, 그에 맞추어 가장 최적화된 결정을 수행하는 이상적인 에이전트라고 가정하였다. 따라서 2단계 마르코프 의사결정 과제에서 성취 조건에 따른 보상의 지급 방식과 상태 천이 확률이 다른 것을 모두 알고 이에 최적으로 대응하고 학습할 수 있도록 SR-매트릭스를 두 개($M_{specific}$/$M_{flexible}$)로 분리한 에이전트를 사용하였다.

마지막으로 실험군/대조군의 성능을 평가하기 위한 상한 평가 모델로 2단계 마르코프 의사결정 과제에 최적화되어 학습된 포워드(forward) 알고리즘 기반 모델 기반 강화학습 에이전트[6]를 이용하여 최적 보상 값을 구하였다.

Table 1과 Fig. 4에서 보는 바와 같이, 실험군(단일 SR-매트릭스)와 대조군 (이중 SR-매트릭스) 사이에 평균 보상은

Table 1. Average Normalized Reward in the 2-stage Markov Decision Task

	In normal 2-satge MDT[6]		
Chance Level	Combined SR-Matrix	Separated SR-Matrix	Model-based
0.35	0.36	0.34	0.63

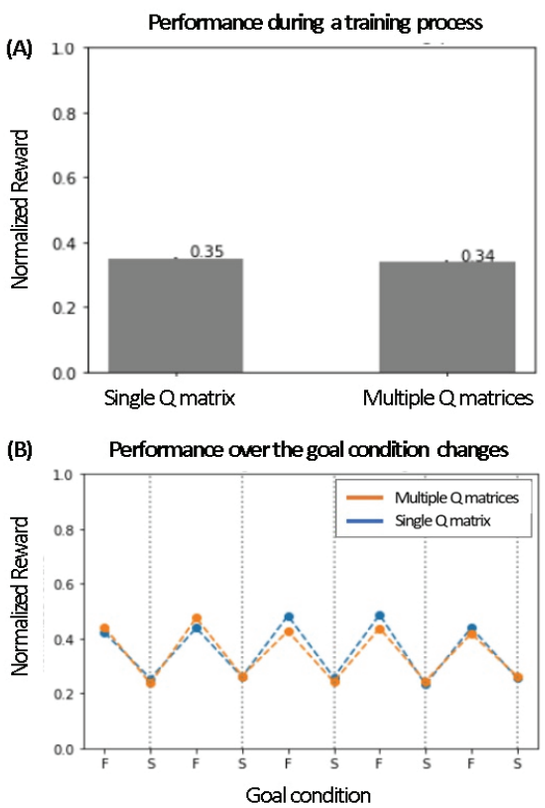


Fig. 4. Changes in Average Normalized Reward in Response to the Changes in Goal Conditions While Performing 2-stage Markov Decision Task

큰 차이를 보이지 않았다. 특히 특정 성취 조건에서의 두 매트릭스를 비교해본 결과, Fig. 4의 오른쪽에서 보는 바와 같이 특정 성취 조건인 경우 (S) 두 그룹 간 평균 보상 차이는 거의 나지 않았으며, 두 그룹 모두 자유 성취 조건에 비해 리턴이 매우 낮은 것을 볼 수 있었다.

2단계 마르코프 의사결정 과제에서 우연 수준(우연 수준) 보상 값은 0.35로 분석되는데[5], 두 그룹 모두 우연 수준 값과 유사한 수준을 보여 적정 학습에 실패한 것으로 사료된다. 뿐만 아니라. 상한에 해당하는 모델 기반 강화학습 알고리즘의 보상 값인 0.63보다 낮은 수치를 기록한 것을 고려했을 때, SR-다이나는 잠재 변수가 모두 변화하여 보상 함수가 변화하는 환경은 효과적으로 학습하지 못한 것으로 보인다.

4.2 환경 불확실성 제어에 따른 SR-다이나 성능 변화 실험

앞선 4.1에서는 SR-다이나가 상태 천이 확률과 보상 함수가 동시에 변화하는 환경, 다시 말해 두 가지 잠재 변수가 동시에 변화하는 불확실성이 높은 환경에서 학습할 수 있는지 확인하고자 하였다. 결과적으로 잠재 변수의 변화로 인한 보상 변화가 일어날 때, SR-다이나는 학습이 어려움을 확인하였다.

따라서, 기존 연구의 실험 설정에서 추가적으로 환경을 변화시키는 잠재 변수(latent variable)을 제어하여, (i) 환경 불확실성이 낮은 경우에서의 학습 성능을 확인하고 (ii) 어떤 잠재 변수에 더 취약한지 살펴보고자 두 가지 시나리오를 고안하였다. 2단계 마르코프 의사결정 과제환경에서 환경을 변화시키는 잠재 변수로는 성취 조건과 state transition probability(상태 천이 확률)가 있다. 특히 이 성취 조건이 수집해야 할 동전의 색깔을 지정해주는 특정 성취 조건으로 지정되어 있는 경우, 50 게임 동안 매 게임에서 무작위로 목표 동전 색깔이 바뀌도록 설정되어 있어서 잠재 변수가 항상 보상 변화를 야기했었다.

본 실험에서는 이 설정 또한 context 변화가 일어나는 것으로 가정하여 이를 단순화하여 더 정적인 환경 상태를 만드는 시나리오 1을 고안하였다. 시나리오 1은 특정 성취 조건에서 수집해야 할 동전의 색깔이 50 게임 동안 동일하게 지속되는 환경이다. 이에 따라 잠재 변수 중 하나가 비활성화되어 4.1에서 제시한 환경보다 환경 불확실성이 더 낮아진, 그래서 더 정적인 상황으로 볼 수 있다.

시나리오 2는 또 다른 잠재 변수 상태 천이 확률을 고정하였다. SR-다이나가 보상 함수가 변화하는 환경에서도 구조를 잘 배운다는 특성을 가정하여, 2단계 마르코프 의사결정 과제 환경에서 보상 함수만 변화할 때의 성능을 보고자하기 위해 상태 천이 확률을 고정하였다. 즉 기존의 (0.9, 0.1) 또는 (0.5, 0.5) 확률로 다음 상태로 이동하는 것이 아닌 무조건 (1.0, 0.0)으로 이동하게 된다. 상태 천이 확률이 위와 같이 설정되는 경우 Fig. 3의 최종 상태(S6-S9)에서 S7을 방문하지 않고 S6을 중복하여 방문하는 문제가 발생하는데, 이 경우 중복되는 S6중 하나를 S7로 바꿔주어 모든 최종 상태(모

Table 2. Normalized Reward at goal condition in Each Scenario

	Specific	Flexible
Scenario 1	0.439375	0.439375
Scenario 2	0.35	0.35

든 동전의 색깔)가 등장하도록 해주었다. 시나리오 1, 2에서, 성취 조건 변화에 따른 보상 기댓값을 계산하여 성능을 평가할 수 있는 지표로 삼았고, 이는 Table 2에 기술되어 있다.

단일 매트릭스 기준 1 세션을 30번 수행(=500 게임 *30 세션 = 15000 게임s)한 것에 대한 정규화된 평균 보상 값은 Fig. 5와 같다. Fig. 5의 위쪽 막대 그래프 (Fig. 5-A상, B상, C상)와 Table 2에서 안내된 기댓값을 비교해보았을 때, 전체적으로 평균 보상 값이 기댓값보다 높은 패턴을 보인다.

4.3 고찰

세부적으로 확인해보면, 시나리오 1에 해당하는 우연 수준 기댓값은 (특정 = 0.44, 자유 = 0.44)로, Fig. 5-A상과 비교해 보았을 때, 근소하게 높은 값을 보인다. Fig. 5-A의 하단 그래프에서 보는 바와 같이 성취 조건 별 평균 보상 값의 간격이 4.1장 실험과 비교했을 때 크게 줄어든 것을 보아, 이 두 가지 증거로 볼 때 SR-다이나가 시나리오 1에 해당하는 환경을 성공적으로 학습했다고 볼 수 있다. 동전 색깔이 고정됨에 따라, 2단계 마르코프 의사결정 과제는 상태 천이 확률과 관계없이 가장 높은 값의 동전을 추구하는 의사 결정 문제로 귀결되어, 상태 천이 확률에 따른 보상 분포 변화를 학습하는 상황이 된 것으로 사료된다. 이에 따라, 보상 분포 변화에 강인한 SR 기반 강화학습 알고리즘인 SR-다이나의 특성이 잘 발휘되었다고 사료된다.

시나리오 2는 성취 조건 잠재 변수는 유효한 상태에서 상태 천이 확률을 고정한 상황이다. 이 경우, Fig. 5-B의 하단에서 보는 바와 같이 자유 성취 조건에서는 매우 높은 평균 보상을 보이고 있다. 이는 상태 천이 확률에 따른 불확실성이 전혀 없는 바, 최대 보상을 획득할 수 있는 동전에 접근하는 것에 불확실성이 전혀 없어 항상 최적 보상을 추구할 수 있는 전략을 학습했다고 볼 수 있다. 다만, 다른 잠재 변수 성취 조건에 따른 불확실성은 존재하는 상황이므로 특정 성취 조건에서는 평균 보상이 크게 떨어지는 것을 볼 수 있다. 특히 특정 성취 조건에서 수집할 동전 색깔을 게임마다 새로이 제시하였을 때, 상태 천이 확률이 고정됨에 따라, 제시된 색깔의 동전에 접근할 수 있는 전략이 확률적으로 존재하지 않을 가능성도 있어 특정 성취 조건에서는 평균 보상이 크게 하락하는 것으로 예상된다. 그림 4의 우측 그래프 특정 성취 조건에서의 리턴 값과 비교해보았을 때, 그 값이 비슷한 것으로 보아 역시나 SR-다이나는 빠르게 변화하는 보상 변화를 학습하지 못하는 것으로 보인다.

마지막으로, 특정 성취 조건에서 수집해야 할 동전의 색깔이 일정 게임 동안 지속되는 시나리오 1과 상태 천이 확률이 (1.0, 0.0)으로 고정된 시나리오 2를 모두 적용한 시나리오 3에서 성능을 확인하였다. 이 환경은 두 가지 잠재 변수를 모두 고정하여 환경 불확실성을 모두 낮춘 경우인데, 이 경우에는 평균 보상도 제일 높은 것을 알 수 있고, 성취 조건에 따른 평균 보상 차이도 줄어드는 경향이 관찰된다. 특히 본 논문의 주요 포인트인 실험군(단일 SR-매트릭스)과 대조군(이중 SR-매트릭스)의 성능 차이를 중점적으로 관찰하였는데, 4.1 설정 및 시나리오 1, 2에서와 달리 평균 보상 값이 비슷하거나 더 높은 수준을 보였다.

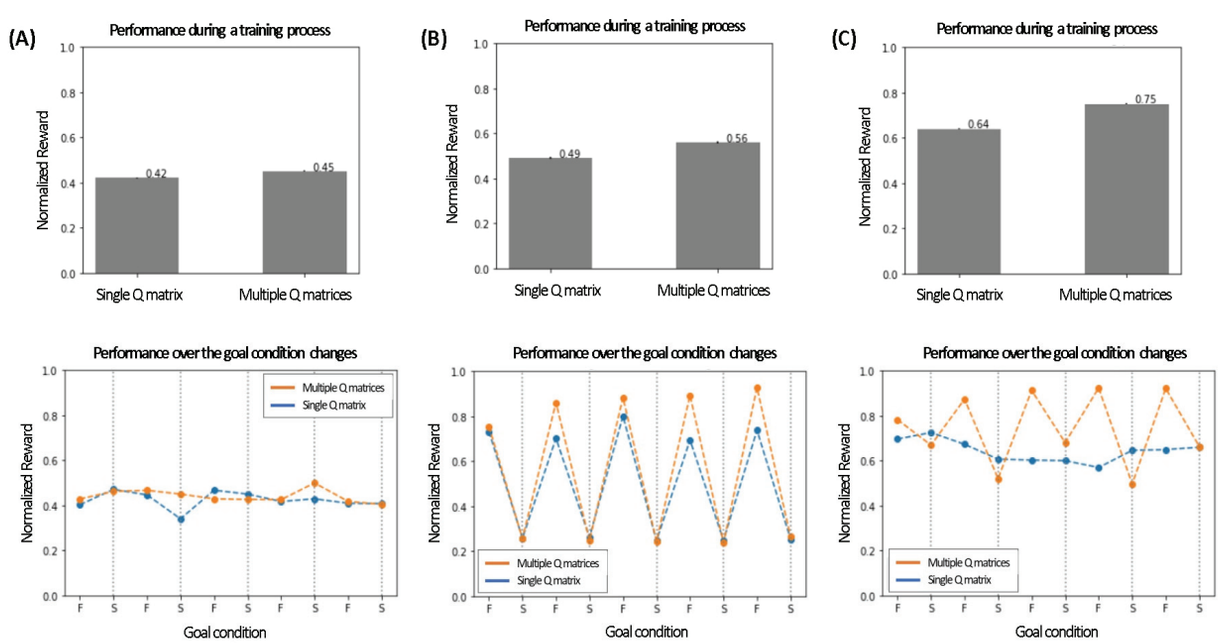


Fig. 5. A: Scenario 1, B: Scenario 2, C: Scenario 1+2. In Each Scenario, the Top Shows the Average Normalized Reward During a Training Process. The Bottom Shows the Average Normalized Reward in Response to the Changes in Goal Conditions.

상기 세 가지 시나리오에 따른 결과를 종합해보자면, SR 기반 강화학습 알고리즘은 상태 천이 확률이 변화함에 따라 보상 함수가 변화하는 환경에서는 학습 가능성이 있으나, 성취 조건이 변화하면서 상태 천이 확률도 함께 변하는 환경에서는 두 잠재 변수가 모두 보상 함수 변화에 영향을 주는 바, 학습이 어려움을 확인하였다. 이에 따라, 상기 두 잠재 변수에 따른 환경 불확실성이 높은 경우에는 상태 변화를 면밀히 관찰하는 동시에 이에 따라 변화하는 목표 변화를 연합하여 일정 수준의 성능을 유지할 수 있는 전략을 학습하고 계획할 수 있는 방안이 필요하다.

5. 결 론

본 연구에서는 모델 기반과 모델 프리 강화학습을 고루 근사화한 SR-기반 강화학습 에이전트가 모델 기반과 모델 프리 강화학습의 행동을 분리하는 환경에서 어떠한 성능을 내는지 분석해보고자 하였다. 환경을 변화시키는 요인을 순차적으로 제어해가며 기존 환경을 단순화한 시나리오에서 추가 분석을 실시하였다. 환경 변화에 따른 보상 변화에 강인한 SR-다이나 모델을 사용하여 성능 평가를 시도하였고, 실험 결과 SR-다이나는 성취 조건별로 분리하여 학습한 경우에서 더 효과적인 학습을 보여주었다. 또한 상태 천이 확률에 따라 변화하는 보상 변화도 학습하는 모습을 보여주었는데, 다만 빠르게 변화하는 보상 변화는 쫓아가지는 못하는 것으로 예상한다. 이를 극복하기 위해 빠르게 변화하는 보상에도 강인한 성능을 낼 수 있도록 학습률을 적응적으로 조절할 수 있는 심층 메타 강화학습[19]의 개념을 적용하거나, 평균 보상을 일정하게 유지하도록 상태 변화와 가치 변화를 동시에 고려하는 자가 항상성 RL 모델[20], 또는 예측 오류 신호 기반 메타 제어 강화학습[5]을 근간으로 하여 추가적 연구를 고려할 수 있다.

References

[1] R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction," MIT press, 2018.

[2] D. Silver, et al., "Mastering the game of go with deep neural networks and tree search," *Nature*, Vol.529, No.7587, pp.484-489, 2016.

[3] D. Silver, et al., "Mastering the game of go without human knowledge," *Nature*, Vol.550, No.7676, pp.354-359, 2017.

[4] J. Schrittwieser, et al., "Mastering atari, go, chess and shogi by planning with a learned model," *Nature*, Vol.588, No.7839, pp.604-609, 2020.

[5] J. H. Lee, B. Seymour, J. Z. Leibo, S. J. Lee, and S. W. Lee, "Toward high-performance, memory-efficient, and fast reinforcement learning-Lessons from decision neuroscience," *Science Robotics*, Vol.4, No.26, pp.eaav2975, 2019.

[6] S. W. Lee, S. Shimojo, and J. P. O'Doherty, "Neural computations underlying arbitration between model-based and model-free learning," *Neuron*, Vol.81, No.3, pp.687-699, 2014.

[7] J. P. O'Doherty, S. W. Lee, and D. McNamee, "The structure of reinforcement-learning mechanisms in the human brain," *Current Opinion in Behavioral Sciences*, Vol.1, pp.94-100, 2014.

[8] J. X. Wang, et al., "Prefrontal cortex as a meta-reinforcement learning system," *Nature Neuroscience*, Vol.21, No.6, pp.860-868, 2018.

[9] W. Dabney, G. Ostrovski, D. Silver, and R.Munos, "Implicit quantile networks for distributional reinforcement learning," In: *International Conference on Machine Learning*, PMLR, pp.1096-1105, 2018.

[10] D. Hassabis, D. Kumaran, C. Summerfield, and M.Botvinick, "Neuroscience-inspired artificial intelligence," *Neuron*, Vol.95, No.2, pp.245-258, 2017.

[11] S.-H. Kim, and J. H. Lee, "Evaluating a successor representation-based reinforcement learning algorithm in the 2-stage Markov decision task," In: *Proceedings of the Korea Information Processing Society Conference*, Korea Information Processing Society, pp.910-913, 2021.

[12] K. L. Stachenfeld, M. M. Botvinick, and S. J. Gershman, "The hippocampus as a predictive map," *Nature Neuroscience*, Vol.20, No.11, pp.1643-1653, 2017.

[13] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine Learning*, Vol.3, No.1, pp.9-44, 1988.

[14] S. J. Gershman, "The successor representation: Its computational logic and neural substrates," *Journal of Neuroscence*, Vol.38, No.33, pp.7193-7200, 2018.

[15] I. Momennejad, E. M. Russek, J. H. Cheong, M. M. Botvinick, N. D. Daw, and S. J. Gershman, "The successor representation in human reinforcement learning," *Nature Human Behaviour*, Vol.1, No.9, pp.680-692, 2017.

[16] E. M. Russek, I. Momennejad, M. M. Botvinick, S. J. Gershman, and N. D. Daw, "Predictive representations can link model-based reinforcement learning to model-free mechanisms," *PLoS Computational Biology*, Vol.13, No.9, pp.e1005768, 2017.

[17] E. C. Tolman, "Cognitive maps in rats and men," *Psychological Review*, Vol.55, No.4, pp.189, 1948.

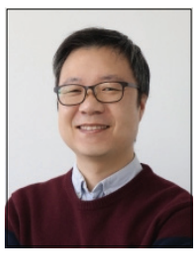
[18] R. S. Sutton, "Dyna, an integrated architecture for learning, planning, and reacting," *ACM Sigart Bulletin*, Vol.2, No.4, pp.160-163, 1991.

[19] J. X. Wang, et al., "Learning to reinforcement learn," *arXiv preprint arXiv:1611.05763*, 2016.

[20] G. Farquhar, et al., "Self-Consistent Models and Values," *Advances in Neural Information Processing Systems*, Vol.34, pp.1111-1125, 2021.

김 소 현

https://orcid.org/0000-0002-1317-4363
e-mail : 202131054@sangmyung.kr
2021년 상명대학교 휴먼지능정보공학과 (학사)
2021년 ~ 현 재 상명대학교 지능정보공학과 석사과정
관심분야 : Reinforcement learning, Decision Making



이 지 항

https://orcid.org/0000-0002-4337-2774
e-mail : jeehang@smu.ac.kr
2015년 University of Bath(박사)
2000년 ~ 2005년 한글과컴퓨터 주임연구원
2005년 ~ 2010년 삼성전자 DMC연구소 책임연구원
2015년 ~ 2016년 University of Bath, Research Associate
2017년 ~ 2019년 한국과학기술원 KI-Postdoc
2019년 ~ 2020년 한국과학기술원 바이오및뇌공학과 연구조교수
2020년 ~ 현 재 상명대학교 휴먼지능정보공학과 조교수
관심분야 : 의사결정, 규범추론, 뇌기반인공지능