Journal of Next-generation Convergence Information Services Technology Vol.11, No.3, June (2022), pp.287-298
ISSN 2508-9099(p) 2672-1198(e) *http://dx.doi.org/10.29056/jncist.2022.06.06*

심층 결정론적 정책 경사모형 기반 포트폴리오 연구

A Study on Portfolio based on a Deep Deterministic Policy Gradient

이우식[1]

Woo Sik Lee[1]

요 약

최근 심층 강화학습을 게임, 로보틱스, 자율주행, 데이터 냉방 솔루션 등 여러 분야에 활용하고 있다. 강화학습은 정책을 스스로 결정할 수 있는 알고리즘으로 지속적인 감독 없이 자동화된 방식으로 포트폴리오의 자산 배분을 수행할 수 있다. 본 연구에서는 활성화 함수에 따른 심층 결정론적 정책 경사모형 기반의 포트폴리오 성능을 비교 분석하였다. 본 연구의 분석 결과, 심층 결정론적 정책 경사모형 기반 포트폴리오의 샤프지수가 벤치마크보다 더 높은 수치를 기록했다. 이는 심층 결정론적 정책 경사모형이 액터-크리틱 모형 기반으로 학습하기 때문에 동작 확률분포로 동작을 선택해 보상을 받는다. 그리고 이것을 상태 가치와 비교하여 이익 계산을 하므로 최적의 정책을 학습시킬 확률이 높아졌다고 판단된다. 하지만 정류 선형 단위, 누출 정류 선형 단위 그리고 지수 선형 단위 활성화 함수를 비교한 결과, 대부분의 활성화 함수에서 비슷한 성능을 보였다. 이런 맥락에서 심층 강화학습에 미치는 영향을 검증하는 것은 금융산업에서도 중요한 관심 대상이 된다.

핵심어 : 강화학습, 비즈니스 애널리틱스, 핀테크, 최적화, 계량금융

Abstract

Deep Reinforcement learning has recently been used in a variety of fields such as games, robotics, autonomous driving, data cooling solutions. Reinforcement learning, which can decide on their own policy, is algorithm for performing portfolio allocation in an automated manner without the need for continuous supervision. Some activation functions were used to compare and analyze portfolio performance based on the Deep Deterministic Policy Gradient Algorithm(DDPG). The Sharp Index of the portfolio based on the DDPG recorded a higher value than the benchmark. One reason for this is that understanding action probabilities is required to select an action and receive a reward, which we then compare to the state value to determine an advantage. Furthermore, the probability of learning the optimal policy is thought to have increased because profits are calculated by comparing this to the state value. However, most activation functions performed similarly when ReLU, Leaky ReLU and ELU were compared. In this context, verifying the impact on in-depth reinforcement learning is a hot topic in the financial industry.

Keyword : Reinforcement Learning, Business Analytics, FinTech, Optimization, Quantitative Finance

1 College of Business Administration, Gyeongsang National University, Jinju, Korea [Professor]
e-mail: woosiklee@gnu.ac.kr

Received(May 3, 2022), Review Result(1st: May 20, 2022), Accepted(June 10, 2022), Published(June 30, 2022)

© 2022 The Authors. Published by NCISS.
This is an open access article licensed under the Creative Commons Attribution-NonCommercial 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc/4.0/.

1. 서론

한국은행 금융통화위원회가 2022년 4월 기준금리 '연 1.5%'로 인상하였고, 주택담보대출 금리도 연 6%대까지 큰 폭으로 상승했지만, 예금 및 적금의 금리는 지금까지 2%대로 낮은 수준에 머물러 있다. 우크라이나 사태, 미국과 중국 간의 무역전쟁 등으로 급변하는 국제 정세 속에 글로벌 공급망의 경제 회복에 대한 전망은 갈리는 것으로 분석이 나오고 있다. 이러한 불투명한 국내·외 경제 환경에 노후 대비 자금 마련, 목돈 마련 등의 자산 증식을 위해서는 금융 투자가 필수가 된 상황이다 [1].

하지만 에너지 가격 물가 상승, 인플레이션 등 불확실하고 예측이 어려운 금융환경에서 수익률 관리에 매우 어려움이 존재한다. 이러한 금융시장 위험에 대응하기 위해 변동성을 최소화하는 전략으로 안정적인 수익률 달성을 추구하는 금융자산 배분에 관심을 둘 필요가 있다 [1]. 금융자산 배분은 학문적 측면에서뿐만 아니라 실무적 측면에서도 오랫동안 널리 활용되고 있는 금융 전략이다. 해리 마코위츠(Markowitz) [2]는 위험과 수익 관계의 정량적 분석을 통한 투자자산 배분을 제시하였고, 브린슨(Brinson) [3]은 금융자산 배분에 관한 전략 성과가 투자 시기(Market Timing), 종목 선택(Asset Selection) 그리고 금융자산 배분(Asset allocation) 중 약 91.5%가 자산의 효율적 배분에 의해 영향을 받는다고 보고한 바 있다.

금융자산 배분은 투자자의 투자 목적과 투자 성향 등을 고려하여 금융자산 배분을 구성하고, 감내 가능한 투자 위험 수준과 투자 목적 달성을 위한 금융자산 배분 최적화 등 다수의 과정을 포괄하는 과학적 투자전략 [4]으로 최근 주요 자산운용사, 은행, 증권회사, 스타트업(Startup)기업 등은 인공지능 기반의 자산관리 서비스 제공을 위한 로보어드바이저(Robo-Advisor) 서비스를 제공하고 있다. 로보어드바이저를 통해 고객 맞춤형 포트폴리오를 제공한다지만 아직 충분히 검증된 상태가 아니고, 인공신경망(Artificial Neural Network)을 포함한 강화학습(Reinforcement Learning)과 금융(Finance)이 연계된 연구들은 미흡한 실정이다. 기존 연구 중 김선웅 [5]은 블랙리터만 모델(Black-Litterman Model)의 투자자 전망을 서포트벡터머신을 활용하여 자산 배분 모델을 제시하였고, 이우식 [1]은 Deep Q-Networks(DQN)기반의 포트폴리오와 목표시장지수의 샤프지수와 비교 분석하여 인공신경망이 포함된 강화학습의 적용 가능성을 제시하였다. García-Galicia 외 [6]는 자산 배분 관리를 위해 어드밴티지 액터 크리틱(Advantage Actor-Critic)모형을 이용하여 에이전트가 각 금융시장 상태 간의 전이 확률(Transition Rate Matrices)을 측정하였다. 본 연구는 활성화 함수(Activation Function)를 기반으로 심층 결정론적 정책 경사모형에 따라 포트폴리오 성능을 비교 · 분석하고자 한다.

본 논문은 다음과 같이 구성되어 있다. 본 연구의 필요성을 밝힌 제1절의 서론에 이어 제2절에

서는 주요 방법론인 심층 결정론적 정책과 활성화 함수들에 대한 설명을 소개하였으며, 제3절에서는 실증분석 및 결과를 확인한다. 마지막으로 제4절에서는 결론과 시사점을 제시한다.

2. 이론적 배경

2.1 심층 결정론적 정책 강하 알고리즘

최근 인공 신경망(Artificial Neural Network)을 강화학습(Reinforcement Learning)에 적용한 심층강화학습(Deep Reinforcement Learning)을 게임, 로보틱스, 자율주행, 데이터 냉방 솔루션 그리고 프로그램 작성 등 여러 분야에 활용하고 있다. 강화학습은 { S, A, P, R, γ }로 구성된 마르코프 결정과정(MDP: Markov Decision Process)으로 정의되는 환경으로부터 누적 보상값의 기대값(Expected Cumulative Reward)을 최대화하는 최적화 기법이다 [1]. 여기서 S는 유한한 크기를 갖는 상태(State) 집합이고, A는 유한한 크기를 갖는 행동(Action) 집합이며, $P(s'|s,a)$는 상태 전이(State Transition) 확률로 현재 상태 $s \in S$ 에서 행동 $a \in A$를 취했을 때 다음 상태가 $s' \in S$이 되는 확률 분포(Probability Distribution)를 의미한다. 또한 R은 보상 함수, $\gamma \in (0,1)$은 할인 계수를 나타낸다 [1]. 강화학습의 학습 순서는 각 시간 단계(Time Step)에서 정의된 에이전트(Agent)가 주어진 환경에서 현재의 상태를 관찰하여 이를 기반으로 행동을 선택하고, 이때 환경의 상태가 변화하면서 정의된 에이전트는 행동에 따른 보상을 받는다 [1]. 학습 초기에 에이전트는 무작위 행동을 하지만, 학습이 점차 진행되면서 더 많은 보상을 얻을 수 있는 행동으로 학습하게 된다 [1]. 강화학습은 초기 상태로부터 정책에 기반을 두고 연속적인 행동(Continuous Action)을 취했을 때의 기대 누적 보상(Expected Cumulative Reward)을 최대화하는 정책을 발견하는 것을 목적으로 하는데, 이것을 최적 정책(Optimal Policy)이라고 지칭한다 [1].

심층강화학습은 마르코프 결정 과정을 사용하여 순차적 의사결정 문제를 모형화하고 가치 함수(Value Function)나 정책 함수(Policy Function)에 대한 근사자(Approximator)로 인공 신경망을 강화학습에 활용하여 문제를 해결하는 방법론이다. 심층강화학습 기법 중 가치 함수와 정책 함수에 대한 근사를 모두 활용한 심층 결정론적 정책 강하법(Deep Deterministic Policy Gradient Algorithm)은 액터(Actor) 신경망이 행동을 계산하고 크리틱(Critic) 신경망이 행동 가치에 대한 계산을 통해 행동 개선을 도모하는 알고리즘으로 [표 1]와 같이 수행한다. [그림 1]에서 특정 상태 s_t하에서 식 (1)에서 보듯이 ϵ-탐욕 정책으로 행동 a_t를 결정한다. |A(s)|는 상태 s 하에서 작동 가능한 행동의 개수이고, A* 는 Q 값을 최대로 설정한 행동 a 인 것이며, $0 \le \epsilon \le 1$ 이다. 다시 말해, ϵ=1 의 조건에서는 행동을 랜덤하게 100% 선택하고, ϵ=0 의 조건에서는 큐러닝(Q-learning)을 기반으로 한 탐욕 정

책에 따라 행동을 결정하게 된다. 탐욕 정책은 식 (2)에서 보듯이 Q값이 최대에 도달하는 행동을 선택하는 정책이다.

$$\pi(s,a)=\begin{cases}1-\epsilon+\frac{\epsilon}{|A(s)|} & (a=A^*)\\ \frac{\epsilon}{|A(s)|} & (a\neq A^*)\end{cases} \tag{1}$$

$$\mu(s)=\arg max_a \quad Q(s,a) \tag{2}$$

Q 값은 행동 가치함수로써 다음과 같이 계산한다.

$$Q(s_t,a_t)=\int_{s_{t+1},a_{t+1}} \{r_t+\gamma Q(s_{t+1})\} \\ p(s_{t+1}\mid s_t,a_t)p(a_{t+1}\mid s_{t+1})d(s_{t+1},a_{t+1}) \tag{3}$$

행동 a_t의 실행 후 결정되는 다음 상태 s_{t+1}와 보상에 대한 피드백 이후 리플레이 버퍼에 저장하는데, 데이터가 일정 수준 이상 쌓이게 되면 무작위 추출을 통해 신경망을 미니배치로 갱신한다. 신경망의 업데이트에 필요한 데이터들은 상호 독립적이라는 가정을 하고 있지만, 탐욕 정책으로 산출된 데이터는 독립적이지 않으므로 독립성의 부여를 위해 리플레이 버퍼를 사용하게 된다. 행동 가치함수 Q 값을 생성하는 크리틱 신경망은 식 (4)에서 보듯이 손실함수로 업데이트가 이루어지는데, 이러한 과정에서 TD (Time Difference) 타깃 $(r_i+\gamma Q_\phi(s_{t+1},a_{t+1})$ 이 영향을 받게 되기에 크리틱 신경망을 복사한 타깃 크리틱 신경망을 만들고 여기에서 TD 타깃을 계산한다 [7].

행동 a_t를 생성하는 액터 신경망은 식 (5)에서 보듯이 정책 강하법으로 업데이트가 이루어진다. 식 (4)과 (5)에서 N 은 리플레이 버퍼에서 추출한 샘플 수로서 미니배치의 크기라 할 수 있다. [7].

$$L=\frac{1}{N}\sum_{i=1}^{N}\left(r_i+\gamma Q_\phi(s_{t+1},a_{t+1})-Q_\phi(s_t,a_t)\right)^2 \tag{4}$$

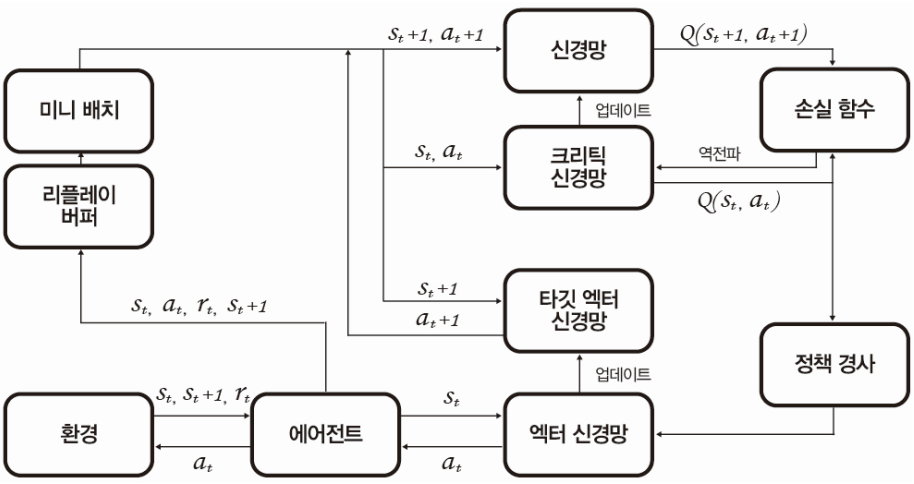
$$\nabla_{\theta^\mu}J\approx\frac{1}{N}\sum_i \quad \nabla_a Q(s,a\mid\theta^Q)_{s=s_i,a=\mu(s_i)}\nabla_{\theta^\mu}\mu(s\mid\theta^\mu)\Big)\mid s_i \tag{5}$$

타깃 액터 신경망은 다음 상태 s_{t+1}의 입력을 받아서 다음 상태 행동 a_{t+1}을 식 (5)로 표현된 탐욕 정책으로 생성된다. 이것은 TD 타깃의 계산을 위해 필요한 자료이다 [7]. 즉, 심층 결정론적 정책 강하법은 앞서 제시한 과정을 통해 액터, 크리틱 신경망의 가중치를 업데이트 함으로써 학습한다. 학습 종료 후 상태 s_t를 입력받게 됐을 때 신경망의 결과인 행동 a_t를 도출해 내는 알고리즘이다 [7].

[표 1] 심층 결정론적 정책 강하 알고리즘

[Table 1] DDPG Algorithm

1. 크리틱 신경망의 모수 θ와 Actor 액터 신경망의 모수 ϕ에 초기치를 부여
2. $\theta' = \theta$ 그리고 $\phi' = \phi$
3. 리플레이 버퍼 D를 초기화
4. $t = 0$부터 $T-1$까지 다음 (a) ~ (e)를 반복
   (a) $\mu_\phi(s_t) + N$으로부터 a_t를 추출
   (b) Transition(s_t, a_t, r_t, s_{t+1})를 D에 저장
   D가 일정 수준에 도달하지 않으면 : 계속
   (c) D로부터 k개의 Transition들을 임의로 추출
   (d) $y_i = r_i + \gamma Q_{\theta'}(s_i^{'}, \mu_{\phi'}(s_i^{'}))$를 계산
   (e) $L(\theta) = \frac{1}{k}\sum_{i=1}^{k}(y_i - Q_\theta(s_i, a_i))^2$으로부터 $\theta = \theta - \alpha \nabla_\theta L(\theta)$으로 θ를 업데이트
   (f) $J(\phi) = \frac{1}{k}\sum_{i=1}^{k} Q_\theta(s_i, \mu_\phi(s_i))$으로부터 $\phi = \phi + \alpha \nabla_\phi J(\phi)$으로 ϕ를 업데이트
   (g) $\theta' = \tau\theta + (1-\tau)\theta', \phi' = \tau\phi + (1-\tau)\phi'$으로 θ'과 ϕ'를 업데이트



[그림 1] 심층 결정론적 정책 강하의 구조 [7]

[Fig. 1] Structure of DDPG

2.2 연구모형

본 연구에서는 심층 결정론적 정책 강하 모형을 이용한 금융자산 배분의 성능평가를 위해 여러 가지 초매개변수(Hyperparameter) 중에서 정류 선형 단위(Rectified Linear Unit, ReLU), 누출 정류 선형 단위(Leaky Rectified Linear Unit, LeakyReLU), 그리고 지수 선형 단위(Exponential Linear Unit,

ELU) 활성화 함수로 심층 결정론적 정책 강하 모형을 학습한다. 일반적으로 심층강화학습에서 사용되는 인공신경망의 활성화 함수는 다음과 같이 정의된다.

$$f(x) = activation(\sum_{i=1}^{n} x_i w_i + b) \tag{6}$$

[표 2]는 본 논문에서 사용된 활성화 함수를 나타낸다. 시그모이드(Sigmoid)는 모든 인공 신경망의 입력값에 대해 0에서 1사이로 변환시키며, 미분이 가능하여 역전파(Backpropagation)를 사용할 수 있었다. 그러나 x값이 매우 커지거나 매우 작아지는 경우, 기울기 소멸 문제 (Vanishing Gradient Problem)가 발생하게 된다. 이를 해결한 것이 선형과 매우 유사한 성질을 가지고 있는 비선형 함수인 정류 선형 단위이다. 그러나 정류 선형 단위도 인공신경망의 입력값이 0에 가까워지거나 음수가 되면 함수의 미분값이 0이 되어 역전파를 수행할 수 없어 학습을 더 이상할 수 없게 되는 한계에 부딪히게 된다. 이러한 문제를 해결하기 위해 0보다 큰 값이 들어올 때는 정류 선형 단위와 같이 동작하지만, 0보다 작은 값이 들어올 때 인공신경망의 입력값에 비례하여 출력으로 보내는 특성을 가진 누출 정류 선형 단위를 사용한다 [8]. 누출 정류 선형 단위는 음의 영역에서도 작은 양의 기울기를 가지므로 음의 입력값에도 역전파가 가능하다 [9]. 누출 정류 선형 단위와 더불어 지수 선형 단위도 양수 입력값에 대한 출력값은 정류 선형 단위와 같지만, 음수 입력값에 대해 출력값은 0에 가까울수록 기울기 변화가 크고 0에서 멀어질수록 기울기 변화가 작은 특성을 통해 죽은 정류 선형 단위(Dying ReLU) 문제를 해결할 수 있을 뿐 아니라 누출 정류 선형 단위와 비교해 출력값의 분산이 작은 장점이 있다 [8].

[표 2] 활성화 함수
[Table 2] Activation Functions

활성화 함수명	활성화 함수식
정류 선형 단위	$f(x) = \begin{cases} 0, & \text{for } x < 0 \\ x, & \text{for } x \geq 0 \end{cases}$
누출 정류 선형 단위	$f(x,a) = \begin{cases} ax, & \text{for } x < 0 \\ x, & \text{for } x \geq 0 \end{cases}$
지수 선형 단위	$f(x) = \begin{cases} x, & \text{for } x \geq 0 \\ a(e^x - 1), & \text{for } x < 0 \end{cases}$

3. 실증분석

3.1 자료의 구성

본 연구에서 활용할 표본은 미국 다우존스 산업 평균 인덱스로 미국 시장 전체를 대표할 수는

없지만, 인덱스를 구성하는 주식 수가 우량기업 30개로 상대적으로 쉽게 추종 인덱스 움직임에 따른 포트폴리오를 만들 수 있고, 이를 통해 투자 포트폴리오의 다변화를 꾀하는 국내 투자자가 미국 금융시장에 투자할 수 있는 간단한 투자수단에 부합한다고 볼 수 있다. 실험을 위해 2011년 1월 3일부터 2018년 12월 31일까지 일별 종가자료를 활용하고, 모형의 성과 측정을 위해 2019년 동안의 투자 기간자료를 확보하였다.

다우존스 산업 평균 인덱스 일별 종가에 대한 기술통계량(평균, 표준편차, 왜도와 첨도)은 [표 3]에서 살펴볼 수 있다. 본 인덱스에서 나타나는 음의 첨도는 꼬리가 정규 분포보다 얇음을 나타내고, 인덱스 수익률의 왜도가 음의 값이라는 것은 부정적 극단 현상의 발생 가능성이 정규분포에 비해 높다는 것을 뜻한다 [1].

$$\text{주가지수변화율} = \ln(\text{주가지수}(t) \ / \text{주가지수}(t-1)) \tag{7}$$

[표 3] 기술 통계
[Table 3] Descriptive Statistics

	다우존스 산업평균지수	다우존스 산업평균지수 변화율(%)
평균	18374.52	0.000432
중앙값	17576.96	0.000565
최대값	28645.26	0.049846
최소값	10655.30	-0.055464
표준편차	4840.076	0.008704
왜도	0.417034	-0.447421
첨도	-1.022344	4.208436

3.2 모형의 추정 및 분석

심층 결정론적 정책 강하 모형 기반으로 한 포트폴리오의 기대 보상 값 비교와 검증을 통하여 최종적으로 최적의 지능형 포트폴리오를 확인하였다. 즉 정류 선형 단위, 누출 정류 선형 단위 그리고 지수 선형 단위 활성화 함수로 이루어진 각각의 심층 결정론적 정책 경사모형의 학습을 위해 미국 다우 인덱스를 구성하고 있는 종목의 일별 수익률과 이에 대한 상관행렬을 상태변수로 사용하고 금융자산 비중에 따른 샤프지수의 비교와 검증을 하였다. 이때 다층 퍼셉트론(Multi-Layer Perceptron, MLP)을 정책 신경망으로 이용하였고, 심층 결정론적 정책 강하 모형의 액터와 크리틱 신경망 공통아키텍처는 유닛 수를 256개로 설정하고 각각 2개와 4개인 다층 퍼셉트론으로 구성하였다. 그러나 공매도와 증권거래세 등의 거래비용은 고려하지 않았다.

인공신경망을 구성할 때, 은닉층(Hidden Layer)의 개수, 은닉층에 존재하는 유닛(Unit)의 개수, 활성화 함수, 가중치 초기화 (Weight Initialization)기법 그리고 최적화 알고리즘 등이 적절히 조합되지

못하면 높은 성능을 이끌어 낼 수 없다. [1]. 이러한 인공신경망의 초매개변수들의 모든 가능한 조합(Combination)을 통한 최적 조합을 찾아내는 것은 상당히 어려운 문제이며 많은 계산량이 요구된다 [1]. 여러 초매개변수 중 학습과 직접적으로 관련된 활성화 함수, 특히 비선형 활성화 함수는 신경망의 표현 능력을 강화시키고 모형이 학습할 수 있는 입력-출력 관계의 범위를 증가시키는 중요한 역할을 한다. 같은 구조와 같은 초매개변수를 갖는 인공신경망이라도 초매개변수의 조합에 따라서 그 결과에 차이가 있을 수 있으므로 최적의 초매개변수를 찾는 것은 매우 중요하다 [1]. 이에 본 논문에서는 정류 선형 단위, 누출 정류 선형 단위 그리고 지수 선형 단위 활성화 함수가 심층 결정론적 정책 경사모형 기반의 포트폴리오에 미치는 영향을 측정하는 연구를 수행했다. 그 결과 [표 4] 및 [표 5]와 같이 모든 실험에서 정(+)의 샤프지수를 보여 이는 위험 대비 투자수익이 발생했음을 의미한다. 이는 심층 결정론적 정책 경사모형이 액터-크리틱(Actor-Critic) 모형 기반으로 학습하기 때문에 동작확률분포에 따라 동작 선택 후 보상을 받고, 이것을 상태 가치와의 비교를 통해서 이익을 계산하기 때문에 최적의 정책을 학습시킬 확률이 높아졌다고 판단된다. 샤프지수가 가장 높을 때는 엘루 활성화 함수와 4개의 다층 퍼셉트론으로 이루어진 정책 신경망을 이용한 심층 결정론적 정책 경사모형을 사용할 때(2.22556)로 나타났고, 모든 활성화 함수에서도 대체로 위험 대비 높은 투자 성과를 보여주었다. 하지만 딥러닝 모형에서와같이 학습 진행이 증가함에 따라 심층 결정론적 정책 경사모형의 에이전트는 주어진 상태에서 점점 더 많은 보상 혹은 샤프지수를 추구하는 쪽으로 행동을 취하는 경향성을 보이지 않음을 확인할 수 있다.

[표 4] 다층 퍼셉트론 정책 신경망을 이용한 포트폴리오의 성능(2개 계층)
[Table 4] Performance of Portfolio using MLP Policy Networks(Two layers)

	수익률	표준편차	샤프지수
Benchmark	0.228985	0.124887	1.720666
학습진행 50			
ReLU	0.292582	0.123538	2.140159
LeakyReLU	0.272190	0.121708	2.039688
ELU	0.264810	0.118496	2.042583
학습진행 500			
ReLU	0.283637	0.118496	2.167413
LeakyReLU	0.250920	0.116939	1.973679
ELU	0.272281	0.121388	2.045320
학습진행 5000			
ReLU	0.265013	0.124396	1.952746
LeakyReLU	0.271568	0.116518	2.121059
ELU	0.273411	0.121057	2.057943

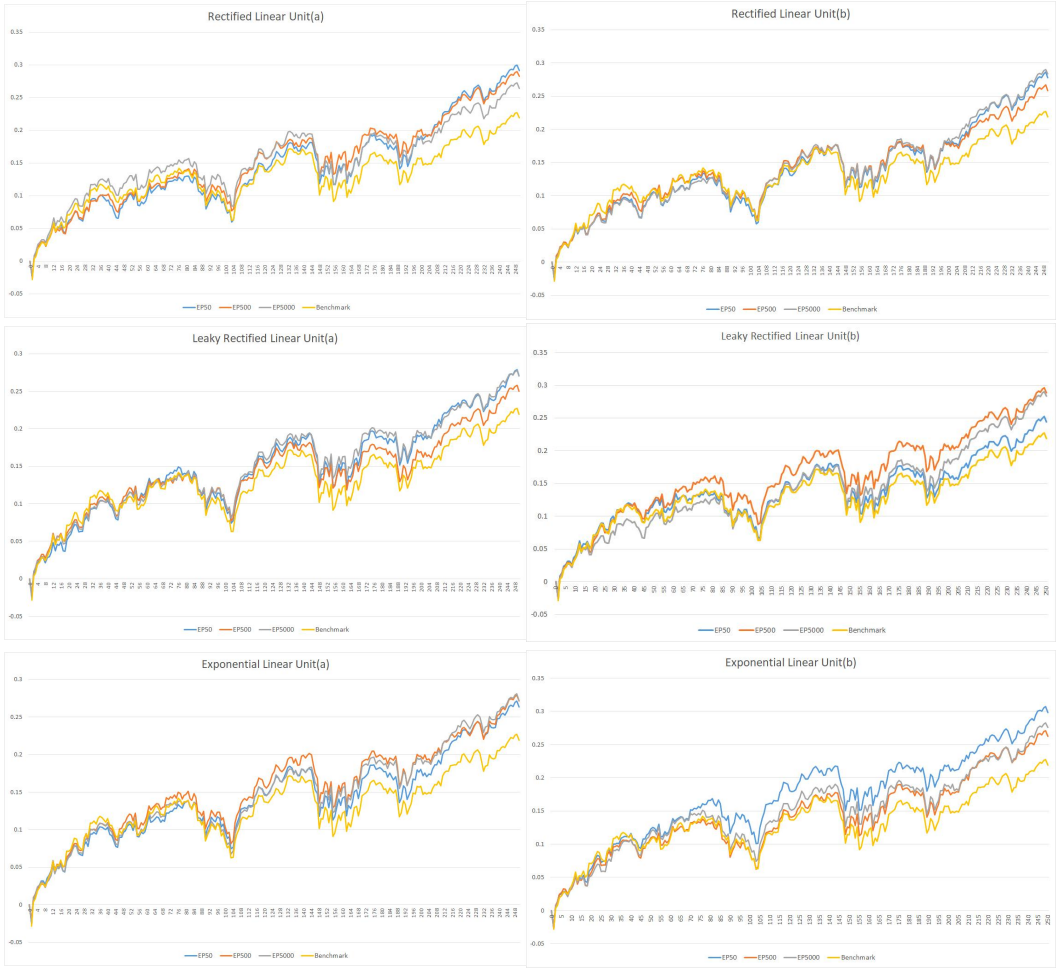
[표 5] 다층 퍼셉트론 정책 신경망을 이용한 포트폴리오의 성능(4개 계층)
[Table 5] Performance of Portfolio using MLP Policy Networks(Four layers)

	수익률	표준편차	샤프지수
Benchmark	0.228985	0.124887	1.720666
학습진행 50			
ReLU	0.27921	0.12237	2.074329
LeakyReLU	0.245378	0.12516	1.816481
ELU	0.299393	0.121021	2.22556
학습진행 500			
ReLU	0.259898	0.120121	1.984126
LeakyReLU	0.289729	0.122711	2.135707
ELU	0.263236	0.122007	1.97703
학습진행 5000			
ReLU	0.25426	0.125102	1.874128
LeakyReLU	0.284536	0.118361	2.175679
ELU	0.276832	0.121902	2.066558

[그림 2]에서와 같이 먼저 모든 에피소드에서 벤치마크보다 높은 누적 수익률을 나타냈다. 더불어 심층 결정론적 정책 경사모형의 학습성능을 살펴보면 정류 선형 단위, 누출 정류 선형 단위 그리고 지수 선형 단위 모두 대부분 짧은 학습 진행에도 불구하고 높은 학습 성과를 보여주었다. 마지막으로 2개(a)와 4개(b)의 다층 퍼셉트론으로 이루어진 정책 신경망을 이용한 심층 결정론적 정책 경사모형 성과분석을 비교한 결과, 신경망을 깊게 하여도 비슷한 성능을 보여주었다.

4. 결론

미국 내 웰스프론트(Wealthfront), 베터먼트(Betterment) 등 약 200여 개의 로보어드바이저 업체가 이미 존재하고 우리나라 금융업체에서도 인공지능을 포함한 머신러닝과 투자자문 전문가를 합성한 용어인 로보어드바이저(Robo-Advisor)에 관심을 보이면서 인공지능을 포함한 머신러닝 기반의 자산관리 서비스 제공을 위한 신생 기업들이 생겨나고, 주요 금융업체들은 인공지능과 빅데이터 기반의 금융 서비스 고도화를 위한 업무 협약이 활발하게 진행되고 있다. 이와 더불어 대한민국 금융위원회에서도 디지털금융의 혁신과 함께 안정적 · 균형적 발전을 도모하기 위해 핀테크(FinTech)를 더욱 활성화하는 내용의 전자금융거래법 개정을 추진하고 있다. 이처럼 디지털 자산관리(Digital Asset Management)에 대한 중요성이 커지면서 로보어드바이저를 포함한 디지털금융의 성장성은 매우 높을 것으로 예상된다 [1].



[그림 2] 누적 수익률

[Fig. 2] Fig. Cumulative Returns

본 연구에서는 심층강화학습을 이용한 금융자산 배분의 성능평가를 위해 정류 선형 단위, 누출 정류 선형 단위 그리고 지수 선형 단위 활성화 함수 기반의 심층 결정론적 정책 경사모형 성능을 분석하였다. 본 연구의 주요 분석 결과는 다음과 같다. 첫째, 심층 결정론적 정책 경사모형 기반 포트폴리오의 샤프지수가 벤치마크보다 더 높은 수치를 기록했다. 이는 심층 결정론적 정책 경사모형이 액터-크리틱(Actor-Critic) 모형 기반으로 학습하기 때문에 동작확률분포(Action Probabilities)로 동작 선택 후 보상을 받고, 이것을 상태 가치와 비교를 하여 이익을 계산하기 때문에 최적의 정책을 학습시킬 확률이 높아졌다고 판단된다. 둘째, 일반적으로 학습 진행이 증가함에 따라 성능이 뛰어나다고 알려진 딥러닝 분석 [10]과 다르게 본 연구에서는 심층 결정론적 정책 경사모형의 에이전트는 모든 활성화 함수에서 점점 더 많은 보상 혹은 샤프지수를 추구하는 쪽으로 행동을 취

하는 경향성을 보임을 확인할 수 있다. 셋째, 2개와 4개의 다층 퍼셉트론으로 이루어진 정책 신경망을 이용한 심층 결정론적 정책 경사모형 성과분석을 비교한 결과, 신경망을 깊게 하여도 비슷한 성능을 보여주었다. 마지막으로, 심층 결정론적 정책 경사모형 학습성능을 살펴보면 정류 선형 단위, 누출 정류 선형 단위 그리고 지수 선형 단위 모두 대부분 짧은 학습 진행에도 불구하고 높은 학습 성과를 보여주었다.

인공신경망을 구성할 때, 은닉층의 수, 은닉층에 존재하는 유닛 개수, 활성화 함수(Activation function), 신경망의 초기화 기법 그리고 최적화 등이 적절히 최적화하지 못하면 높은 결과를 이끌어 낼 수 없다. 이러한 인공신경망의 초매개변수들의 모든 가능한 조합(Combination)을 통한 최적 조합을 찾아내는 것은 상당히 어려운 문제이며 많은 계산량이 요구된다. 여러 초매개변수 중 인공신경망 학습과 가장 직접적으로 관련된 활성화 알고리즘은 기울기 소실(Vanishing Gradient) 및 기울기 폭주(Exploding Gradient) 문제를 효율적인 활성화 알고리즘 선택을 통해 어느 정도 해결할 수 있다. 같은 구조와 같은 초매개변수를 갖는 인공신경망이라도 활성화 알고리즘에 따라 성능이 다르게 나타날 수 있으므로, 적합한 활성화 알고리즘를 찾는 것은 매우 중요하다. 이에 본 논문에서는 정류 선형 단위, 누출 정류 선형 단위 그리고 지수 선형 단위 활성화 함수에 따른 심층 결정론적 정책 경사모형의 학습성능에 미치는 영향을 검증하였다.

하지만 본 논문에도 향후 몇 가지 보완할 점이 필요하다. 심층 결정론적 정책 경사모형은 확정적 정책이므로, 환경을 구석구석 지속적인 탐색을 하기 위한 에이전트의 무작위적인 행동이 필요하다. 더불어 본 연구에서 제시하였던 심층강화학습 모형에 다양한 경제지표 등을 추가하여 더 높은 성과의 기대가 가능한 방안에 관해 후속 연구가 이루어질 필요가 있다.

References

[1] W. Lee, “Performance Evaluation of Portfolio using a Deep Q-Networks”, Journal of Next-generation Convergence Information Services Technology, vol. 10, no. 4, June 2021, pp. 459-470, doi: 10.29056/jncist.2021.08.10.

[2] H. Markowitz, “Portfolio Selection”, Journal of Finance, vol. 7, no. 1, March 1952, pp. 77-91, doi: 10.2307/2975974.

[3] G. P. Brinson, P. L. Randolph-Hood, G. L. Beebower, “Determinants of Portfolio Performance”, Financial Analysts Journal, vol. 51, no. 1, December 1955, pp. 133-138, doi: 10.2469/faj.v42.n4.39.

[4] I. Bajeux-Besnainou, J. V. Jordan, R. Portait, “Dynamic Asset Allocation for Stocks, Bonds, and Cash”, The Journal of Business, vol. 76, no. 2, December 2003, pp. 263-288, doi: 10.1086/367750.

[5] S. Kim, “Robo-Advisor Algorithm with Intelligent View Model”, Journal of intelligence and information systems, vol. 25, no. 2, December 2019, pp. 39-55, doi: 10.13088/jiis.2019.25.2.039.

[6] M. García-Galiciaab, A. A. Carsteanuab, J. B. Clempnerab, “Continuous-time reinforcement learning approach for portfolio management with time penalization”, Expert Systems with Applications, vol. 7, no. 1, September 2019, pp. 27-36, doi: 10.1016/j.eswa.2019.03.055.

[7] J. Lee, K. Kim, J. Lee, “Singularity Avoidance Path Planning on Cooperative Task of Dual Manipulator Using DDPG Algorithm”, Journal of Korea Robotics Society, vol. 16, no. 2, June 2021, pp. 137-146, doi: 10.7746/jkros.2021.16.2.137.

[8] D. Lee, “Comparison of Activation Functions using Deep Reinforcement Learning for Autonomous Driving on Intersection”, The Journal of The Institute of Internet, Broadcasting and Communication, vol. 21, no. 6, December 2021, pp. 117-122, doi: 10.7236/JIIBC.2021.21.6.117.

[9] T. Jiang, J. Cheng, “Target Recognition Based on CNN with LeakyReLU and PReLU Activation Functions”, International Conference on Sensing, Diagnostics, Prognostics, and Control, August 2019, pp. 718-722, doi: 10.1109/SDPC.2019.00136.

[10] W. S. Lee, H. J. Chun, “A deep learning analysis of the Chinese Yuan’s volatility in the onshore and offshore markets”, Journal of the Korean Data & Information Science Society, vol. 27, no. 2, March 2016, pp. 327-335, doi: 10.7465/jkdi.2016.27.2.327.