https://doi.org/10.21289/KSIC.2022.25.3.467

액터-크리틱 모형기반 포트폴리오 연구

A Study on the Portfolio Performance Evaluation using Actor-Critic Reinforcement Learning Algorithms

이우식[1*]

Lee Woo Sik[1*]

〈Abstract〉

The Bank of Korea raised the benchmark interest rate by a quarter percentage point to 1.75 percent per year, and analysts predict that South Korea's policy rate will reach 2.00 percent by the end of calendar year 2022. Furthermore, because market volatility has been significantly increased by a variety of factors, including rising rates, inflation, and market volatility, many investors have struggled to meet their financial objectives or deliver returns. Banks and financial institutions are attempting to provide Robo-Advisors to manage client portfolios without human intervention in this situation. In this regard, determining the best hyper-parameter combination is becoming increasingly important. This study compares some activation functions of the Deep Deterministic Policy Gradient(DDPG) and Twin-delayed Deep Deterministic Policy Gradient (TD3) Algorithms to choose a sequence of actions that maximizes long-term reward. The DDPG and TD3 outperformed its benchmark index, according to the results. One reason for this is that we need to understand the action probabilities in order to choose an action and receive a reward, which we then compare to the state value to determine an advantage. As interest in machine learning has grown and research into deep reinforcement learning has become more active, finding an optimal hyper-parameter combination for DDPG and TD3 has become increasingly important.

Keywords : Quantitative Finance, Business Analytics, FinTech, Autonomous Portfolio, Optimization

1* 정회원, 제1저자, 경상국립대학교 경영대학, 교수
E-mail: woosiklee@gnu.ac.kr

1* College of Business Administration, Gyeongsang National University

1. 서 론

한국은행 금융통화위원회가 2022년 5월 기준금리 '연 1.75%'로 인상하였고, 주택담보대출 금리도 연 6%대까지 큰 폭으로 상승했지만, 예금 및 적금의 금리는 아직까지 2~3%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 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 금융시장 가격 변수의 등락에 따른 변동성으로 자산 가치에 손실이 발생하는 것을 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 배분으로 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 등은 인공지능 기반의 자산관리 서비스 제공을 위한 로보어드바이저 서비스를 제공하고 있다. 로보어드바이저를 통해 고객 맞춤형 포트폴리오를 제공한다지만 아직 충분히 검증된 상태가 아니고, 인공신경망을 포함한 강화학습과 금융이 연계된 연구들은 미흡한 실정이다. 기존 연구 중 김선웅[5]은 블랙리터만모델의 투자자 전망을 서포트벡터머신을 활용하여 자산 배분 모델을 제시하고, 이우식[1]은 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 모형을 이용하여 에이전트가 각 금융시장 상태 간의 전이 확률을 측정하였다. 본 연구는 활성화 함수에 따른 액터-크리틱 모형기반의 포트폴리오 성능을 비교·분석하고자 한다.

본 논문은 다음과 같이 구성되어 있다. 본 연구의 필요성을 밝힌 제1절의 서론에 이어 제2절에서는 주요 방법론인 Deep Deterministic Policy Gradient(DDPG), Twin-delayed Deep Deterministic Policy Gradient(TD3) 그리고 활성화 함수들에 대한 설명을 소개하였으며, 제3절에서는 실증분석 및 결과를 확인한다. 마지막으로 제4절에서는 결론과 시사점을 제시한다.

2. 이론적 배경

2.1 DDPG 알고리즘

최근 인공신경망을 강화학습에 적용한 심층 강화학습(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 결

정 과정으로 정의되는 환경으로부터 누적 보상 값의 기댓값을 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 주시하여 이를 기반으로 행동을 선택하고, 이때 환경의 상태가 변화하면서 정의된 에이전트는 행동에 따른 보상을 받는다. 학습 초기에 에이전트는 무작위 행동을 하지만, 학습이 점차 진행되면서 더 많은 보상을 얻을 수 있는 행동으로 학습하게 된다. 강화학습은 초기 상태로부터 정책에 기반을 두고 연속적인 행동(Continuous Action)을 취했을 때의 기대 누적 보상을 최대화하는 정책을 발견하는 것을 목적으로 하는데, 이것을 최적 정책(Optimal Policy)이라고 지칭한다[1].

심층 강화학습은 마르코프 결정 과정을 사용하여 순차적 의사결정 문제를 모형화하고 가치함수나 정책함수에 대한 근사자로 인공신경망을 강화학습에 활용하여 문제를 해결하는 방법론이다. 심층 강화학습 기법 중 가치함수와 정책함수에 대한 근사를 모두 활용한 DDPG는 비활성 정책 학습법(Off-policy) 기반으로 목표 신경망 $\Theta' = [\phi^{\mu'}, \theta^{\phi}]$을 갖고, 액터(Actor) 신경망 ϕ^{μ}이 행동을 계산하고 크리틱(Critic) 신경망 θ^{Q}이 행동 가치에 대한 계산을 통해 행동 개선을 도모하는 알고리즘이다. Q-함수를 위한 크리틱 신경망의 목적함수 $L(\theta^{Q})$는 다음과 같다[8].

$$L(\theta^{Q}) = \frac{1}{M}\sum_{t}\left[\left(Q(s_{t,j}, a_{t,j} \mid \theta^{Q}) - y_t\right)^2\right] \quad (1)$$

목표 $y_t = r(s_{t,j}, a_{t,j}) + \gamma Q^{\mu}(s_{t+1,j}, \mu(s_{t+1,j}))$는 보상과 목표 정책 μ'하에서의 Q-함수값의 합을 의미한다. 이때 이 행동 $a_{t+1,j}$을 선택할 때 결정론적인 $\arg max$방법을 사용하고 M은 배치크기를 의미한다[8]. 크리틱 신경망은 식 (1)의 목적함수가 최소화되는 방향으로 학습된다.

DDPG의 정책 최적화를 위한 액터 신경망의 목적함수는 다음과 같다[8].

$$L(\phi^{\mu}) = \frac{1}{M}\sum_{t} Q(s, \mu(s \mid \phi^{\mu}) \mid \theta^{Q}) \quad (2)$$

DDPG는 매번 목표 신경망을 다음과 같은 방식으로 업데이트한다.

$$\theta^{Q'} \leftarrow \tau\theta^{Q} + (1-\tau)\theta^{Q'} \quad (3)$$

$$\phi^{\mu'} \leftarrow \tau\phi^{\mu} + (1-\tau)\phi^{\mu'} \quad (4)$$

식 (3)의 θ^{Q}는 크리틱 신경망의 Q-함수를 근사하는 파라미터이며 $\theta^{Q'}$는 목표 Q-함수를 근사하는 파라미터이다. 식 (4)의 ϕ^{μ}는 액터 신경망의 정책 근사 파라미터이며 $\phi^{\mu'}$는 목표 정책을 근사하는 파라미터이다[8].

각각의 신경망의 업데이트 수식에 사용된 τ는 목표 신경망의 변화율을 조절하는 파라미터로 0에 가까울수록 적게 변화하고 1에 가까울수록 크게 변화한다[8]. 이런 방식을 소프트 갱신(Soft update)이라고 하며 DDPG는 이를 통해 목표 신경망이 천천히 변화할 수 있도록 강제한다[8]. 목적함수의 근사는 심층 신경망을 통해 이루어지며 DDPG의 심층 신경망 구조는 Fig. 1(왼쪽)을 통해 액터 신경망과 크리틱 신경망이 각각 1개인 것을 확인할 수 있다.

2.2 TD3 알고리즘

TD3의 신경망 $\Theta = [\phi^{\mu}, \theta^{Q_1}, \theta^{Q_2}]$과 목표 신경망

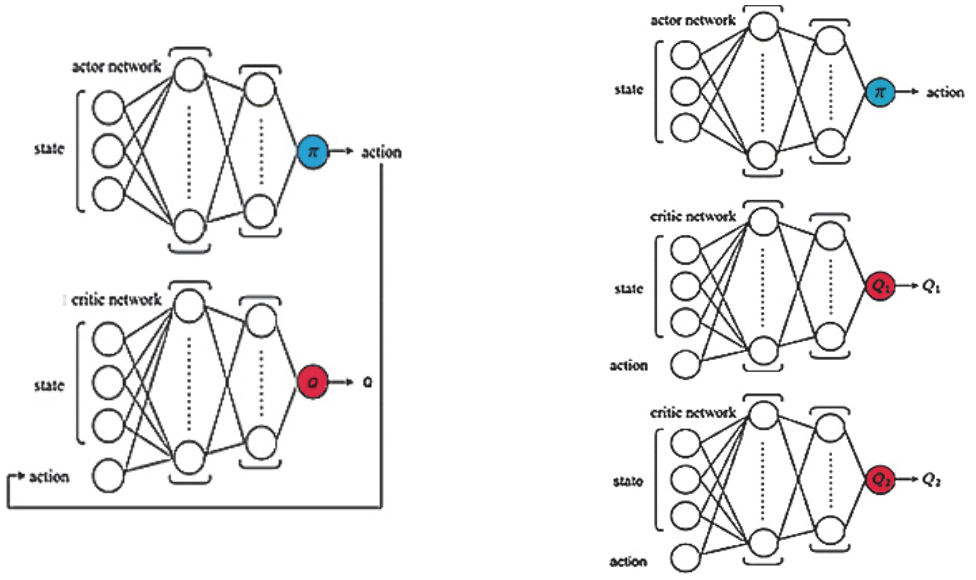

Fig. 1 Structure of DDPG(Left) and TD3(Right)

$\Theta' = [\phi^{\mu'}, \theta^{Q_1'}, \theta^{Q_2'}]$ 은 DDPG에서 크리틱 신경망이 한 개 더 추가된 구성이다[8]. 또한, TD3에서 사용되는 목적함수는 DDPG와 동일한 구성을 가지며 크리틱 신경망에서 사용되는 식 (1)의 목표 함수 y_t를 구하는 방법만 식 (5)로 변경된다[8].

$$\begin{aligned} y_t &= r(s_{t,j}, a_{t,j}) \\ &+ \gamma \min_{i=1,2} Q_{\theta^{Q_i'}}(s_{t+1,j}, \mu(s_{t+1,j}) + \omega) \\ \omega &\sim clip(N(0,\sigma), -\epsilon, +\epsilon) \end{aligned} \quad (5)$$

이는 DDPG의 과대 추정 편향을 해결하기 위해 개선된 방법이다[8]. $\min_{i=1,2} Q_{\theta'}$을 통해 두 신경망에서 근사한 Q-함수 중 더욱 작은 값을 사용한다. 또한 목표 정책에 평활화(smoothing) 기법을 적용하여 행동 선택 과정에서 클리핑 무작위 잡음(Clipped Random Noise) w를 가한다[8].

TD3와 DDPG는 신경망 업데이트 방식에서도 차이를 보인다. DDPG의 경우, 정해진 시간 스텝(Time Step)마다 모든 신경망이 차례로 학습을 하지만 TD3는 지연 업데이트(Delayed Update) 방식을 사용한다. 이는 크리틱 신경망보다 액터 신경망과 목표 신경망의 업데이트 주기를 늦추는 방식이다. 이를 통해 Q-함수가 안정되어 다른 신경망에서 발생하는 과대 추정 및 오류의 축적을 방지할 수 있다. 결과적으로 분산이 낮은 값을 추정할 수 있도록 하며 정책의 품질을 보장한다[8].

TD3의 심층 신경망 구조 역시 전반적으로 DDPG와 유사하다. 그러나 TD3의 경우, Twin Q-learning을 적용하였기 때문에 액터 신경망이 1개 그리고 크리틱 신경망이 2개(Q_1과 Q_2)인 것을 Fig. 1(오른쪽)을 통해 확인할 수 있다[8].

$$L(\theta^Q) = \frac{1}{M}\sum_t \left[\left(Q(s_{t,j}, a_{t,j} \mid \theta^Q) - y_t \right)^2 \right] \quad (6)$$

2.3 활성화 함수

본 연구에서는 액터-크리틱 모형을 이용한 금

융자산 배분의 성능평가를 위해 여러 가지 초매개변수 중에서 ReLU, PReLU, GELU, SiLU 그리고 Mish 활성화 함수로 DDPG와 TD3를 학습한다. 일반적으로 심층 강화학습에서 사용되는 인공신경망의 활성화 함수는 식(7)과 같이 정의된다.

$$f(x) = activation(\sum_{i=1}^{n} x_i w_i + b) \qquad (7)$$

Table 1은 본 논문에서 사용된 활성화 함수를 나타낸다. 시그모이드(Sigmoid)는 모든 인공신경망의 입력값에 대해 0에서 1 사이로 변환시키며, 미분이 가능하여 역전파를 사용할 수 있었다. 그러나 값이 매우 커지거나 매우 작아지는 경우, 기울기 소멸 문제가 발생하게 된다. 이를 해결한 것이 선형과 매우 유사한 성질을 가지고 있는 비선형 함수인 ReLU이다. 그러나 ReLU도 인공신경망의 입력값이 0에 가까워지거나 음수가 되면 함수의 미분 값이 0이 되어 역전파를 수행할 수 없어 학습을 더 이상할 수 없게 되는 한계에 부딪히게 된다. 이런 ReLU의 단점을 해결하기 위해 SiLU, Mish, PReLU, GELU가 제안되었다. SiLU는 큰 음수값이 입력되면 출력이 0이 되어 미분 값이 포화하지만, 작은 음수값이 입력되면 들어온 값을 어느 정도 보존하고 SiLU의 부드러운 오차 손실 경사는 학습률과 초깃값에 대한 민감성을 줄일 수 있는 장점이 있다[9]. Mish도 SiLU와 비슷하게 스스로 정규화되는 비단조함수로서 양수 영역에서 값이 부드럽게 증가하기 때문에 지속해서 미분할 수 있다[9]. PReLU는 기울기가 고정된 Leaky ReLU와 달리 유닛별로 기울기를 학습하여 성능을 개선하고 GELU는 입력값에 Gaussian 분포에 대한 누적분포함수를 곱한 값을 결과로 사용하여 학습의 성능을 개선한다[11].

Table 1. Activation functions

Name	Equation
ReLU	$f(x) = \begin{cases} 0, \text{for } x < 0 \\ x, \text{for } x \geq 0 \end{cases}$
PReLU	$f(x) = \begin{cases} ax, \text{for } x < 0 \\ x, \text{for } x \geq 0 \end{cases}$
GELU	$f(x) = x\Phi(x), where$ $\Phi(x) = x\frac{1}{2}\left[1 + erf(\frac{x}{\sqrt{2}})\right]$
SiLU	$f(x) = x\sigma(x), where$ $\sigma(x) = \frac{1}{1+e^{-x}}$
Mish	$f(x) = x\,Tanh(\ln(1+e^{x}))$

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 동안

Table 2. Descriptive statistics

	DJI	DJI Return(%)
Mean	18374.52	0.000432
Median	17576.96	0.000565
Max.	28645.26	0.049846
Min.	10655.30	-0.055464
S.D.	4840.076	0.008704
Skewness	0.417034	-0.447421
Kurtosis	-1.022344	4.208436

의 투자 기간자료를 확보하였다.

다우존스 산업 평균 인덱스 일별 종가에 대한 기술통계량(평균, 표준편차, 왜도와 첨도)은 Table 2에서 살펴볼 수 있다. 본 인덱스에서 나타나는 음의 첨도는 꼬리가 정규 분포보다 얇음을 나타내고, 인덱스 수익률의 왜도가 음의 값이라는 것은 부정적 극단 현상의 발생 가능성이 정규 분포에 비해 높다는 것을 뜻한다[1].

$$\text{주가지수 변화율} = \ln(\text{주가지수}(t)\ /\text{주가지수}(t-1)) \tag{8}$$

3.2 모형의 추정 및 분석

DDPG와 TD3 기반으로 한 포트폴리오의 기대 보상값 비교와 검증을 통하여 최종적으로 최적의 포트폴리오를 확인하였다. 즉 ReLU, PReLU, GELU, SiLU 그리고 Mish 활성화 함수로 이루어진 각각의 DDPG와 TD3의 학습을 위해 미국 다우 인덱스를 구성하고 있는 종목의 일별 수익률과 이에 대한 상관행렬을 상태변수로 사용하고 금융자산 비중에 따른 샤프지수의 비교와 검증을 시행하였다. 이를 위해 다층 퍼셉트론을 정책 신경망으로 이용하였고, 256개의 유닛 수를 가진 액터와 크리틱 신경망 공통아키텍처는 2개로 설정하였다. 더불어 배치크기는 64와 128개, 버퍼 크기는 50,000과 500,000으로 각각 구성하였다. 공매도와 증권거래세 등의 거래비용은 고려하지 않았다.

인공신경망을 구성할 때, 은닉층의 개수, 은닉층에 존재하는 유닛의 개수, 활성화 함수, 가중치 초기화 기법 그리고 최적화 알고리즘 등이 적절히 조합되지 못하면 높은 성능을 이끌어 낼 수 없다[1]. 이러한 인공신경망의 초매개변수들의 모든 가능한 조합을 통한 최적 조합을 찾아내는 것은 상당히 어려운 문제이며 많은 계산량이 요구된다[1]. 여러 초

Table 3. Performance of portfolio using DDPG and TD3

DDPG(a)	Return	Volatility	Sharpe Ratio
Benchmark	0.228985	0.124887	1.720666
ReLU	0.284268	0.120950	2.129915
GELU	0.286857	0.115796	2.236857
PReLU	0.263154	0.115727	2.077317
SiLU	0.28841	0.116959	2.226098
Mish	0.288553	0.121089	2.155133

DDPG(b)	Return	Volatility	Sharpe Ratio
Benchmark	0.228985	0.124887	1.720666
ReLU	0.259065	0.120513	1.972577
GELU	0.276156	0.123264	2.040762
PReLU	0.285618	0.115596	2.232181
SiLU	0.295032	0.121652	2.186994
Mish	0.284784	0.122893	2.101438

DDPG(c)	Return	Volatility	Sharpe Ratio
Benchmark	0.228985	0.124887	1.720666
ReLU	0.265013	0.124396	1.952746
GELU	0.263903	0.115221	2.091082
PReLU	0.281038	0.121465	2.100654
SiLU	0.279402	0.123173	2.062821
Mish	0.276505	0.120327	2.089866

DDPG(d)	Return	Volatility	Sharpe Ratio
Benchmark	0.228985	0.124887	1.720666
ReLU	0.275177	0.117061	2.135955
GELU	0.299347	0.124757	2.162317
PReLU	0.243471	0.117882	1.908140
SiLU	0.283027	0.12044	2.130354
Mish	0.275258	0.117227	2.133653

TD3(a)	Return	Volatility	Sharpe Ratio
Benchmark	0.228985	0.124887	1.720666
ReLU	0.261593	0.122194	1.963530
GELU	0.2708	0.117225	2.103805
PReLU	0.273204	0.116189	2.137783
SiLU	0.276218	0.121943	2.061923
Mish	0.284734	0.123743	2.087537

TD3(b)	Return	Volatility	Sharpe Ratio
Benchmark	0.228985	0.124887	1.720666
ReLU	0.290192	0.117197	2.233597
GELU	0.258218	0.120606	1.965561
PReLU	0.262992	0.118065	2.037404
SiLU	0.289158	0.121834	2.146543
Mish	0.280684	0.125809	2.030165

Table 3. (Continued)

TD3(c)	Return	Volatility	Sharpe Ratio
Benchmark	0.228985	0.124887	1.720666
ReLU	0.280173	0.120618	2.108946
GELU	0.287890	0.124094	2.101767
PReLU	0.260694	0.117432	2.032250
SiLU	0.268757	0.120947	2.029419
Mish	0.305001	0.123508	2.218129

TD3(d)	Return	Volatility	Sharpe Ratio
Benchmark	0.228985	0.124887	1.720666
ReLU	0.268592	0.121179	2.024669
GELU	0.266398	0.120752	2.017040
PReLU	0.273032	0.121067	2.055310
SiLU	0.287019	0.120384	2.157156
Mish	0.277593	0.117757	2.140135

매개변수 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. 동일한 구조와 동일한 초매개변수를 갖는 인공신경망이라도 초매개변수의 조합에 따라서 그 결과에 차이가 있을 수 있기 때문에 최적의 초매개변수를 찾는 것은 매우 중요하다[1]. 이에 본 논문에서는 ReLU, PReLU, GELU, SiLU 그리고 Mish 활성화 함수가 DDPG와 TD3 기반의 포트폴리오에 미치는 영향을 측정하는 연구를 수행했다. 그 결과 Table 3과 같이 모든 실험에서 정(+)의 샤프지수를 보여 이는 위험 대비 투자수익이 발생했음을 의미한다.

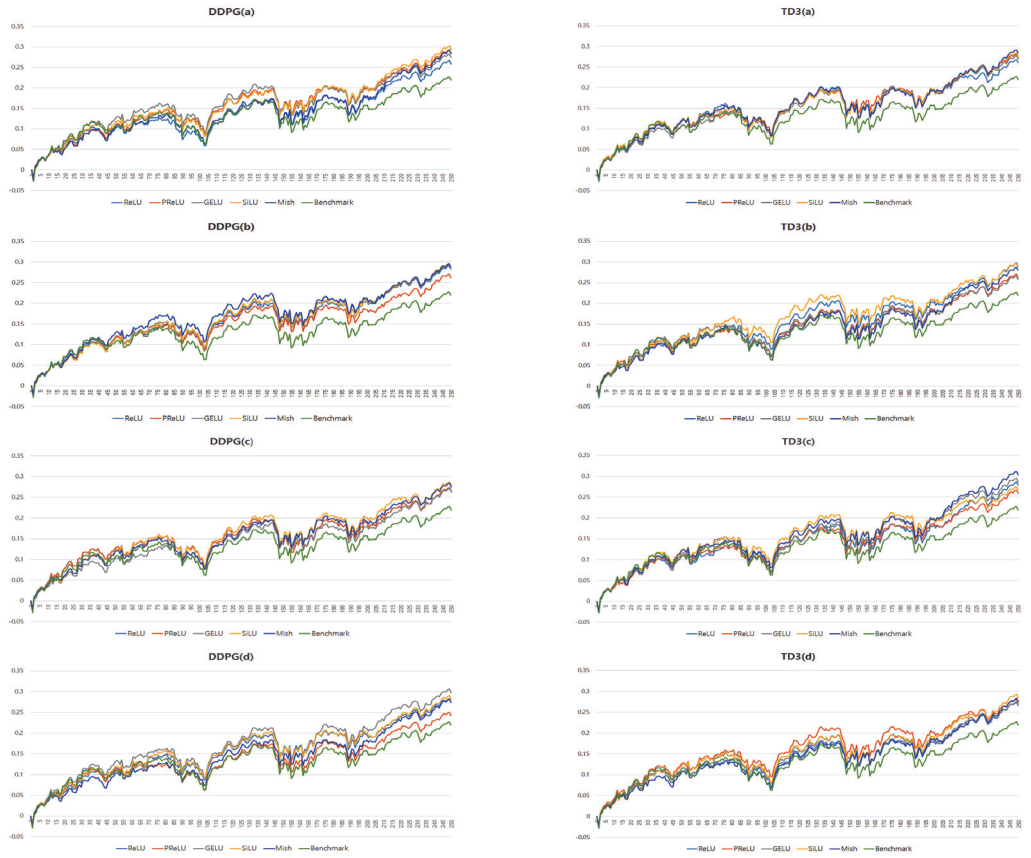

Fig. 2 Cumulative Returns of DDPG(Left) and TD3(Right)

이는 DDPG와 TD3가 액터-크리틱모형 기반으로 학습하기 때문에 동작 확률 분포에 따라 동작 선택 후 보상을 받고, 이것을 상태 가치와의 비교를 통해서 이익을 계산하기 때문에 최적의 정책을 학습시킬 확률이 높아졌다고 판단된다. 샤프지수가 가장 높을 때는 GELU 활성화 함수, 64개의 배치크기 그리고 500,000의 버퍼 크기를 이용한 DDPG을 사용할 때(2.236857)로 나타났고, 모든 활성화 함수에서도 대체로 위험 대비 높은 투자 성과를 보여주었다. 기존의 연구[8-10]와 같이 DDPG의 에이전트가 TD3의 에이전트에 비해 과대 평가되는 방향의 학습 경향을 보이지 않았다. 이는 이미지 자료로 이루어진 환경에서 TD3의 쌍둥이 Q 신경망이 특징 계층을 공유하는 것이 유용할 수 있지만 본 연구에서의 환경은 이미지 자료가 아니기 때문에 DDPG와 비슷한 학습 성과를 보여준 것으로 고려된다. 더불어 TD3의 경우, 정책의 갱신을 지연시키며 두 개의 Q-함수 중 최솟값을 선택적으로 사용하여 학습하기 때문에 급변하는 금융시장 상황에서의 대응이 어려울 수 있지만, DDPG는 다양한 상황별 대처가 가능하다[10].

Fig. 2에서와같이 먼저 모든 활성화 함수에서 벤치마크보다 높은 누적 수익률을 나타냈다. 즉 DDPG와 TD3의 학습성능을 살펴보면 ReLU, PReLU, GELU, SiLU 그리고 Mish가 포함된 모든 활성화 함수에서 높은 학습 성과를 보여주었다. 마지막으로 64개와 128개의 배치크기와 50,000와 500,000의 버퍼 크기를 이용한 DDPG와 TD3 성과 분석에서도 비슷한 성능을 보여주었다.

4. 결 론

미국 내 웰스프론트, 베터먼트 등 약 200여 개의 로보어드바이저업체가 이미 존재하고 우리나라 금융업체에서도 인공지능을 포함한 기계학습과 투자자문 전문가를 합성한 용어인 로보어드바이저(Robo-Advisor)에 관심을 보이면서 인공지능을 포함한 기계학습 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].

본 연구에서는 활성화 함수에 따른 액터-크리틱 모형기반의 포트폴리오 성능을 비교·분석하였다. 본 연구의 주요 분석 결과는 다음과 같다. 첫째, DDPG와 TD3 기반 포트폴리오의 샤프지수가 벤치마크보다 더 높은 수치를 기록했다. 이는 DDPG와 TD3가 액터-크리틱 모형기반으로 학습하기 때문에 동작 확률 분포로 동작 선택 후 보상을 받고, 이것을 상태 가치와 비교를 하여 이익을 계산하기 때문에 최적의 정책을 학습시킬 확률이 높아졌다고 판단된다. 둘째, DDPG와 TD3의 학습성능을 살펴보면 ReLU, PReLU, GELU, SiLU 그리고 Mish가 포함된 모든 활성화 함수에서 높은 학습 성과를 보여주었다. 셋째, 64개와 128개의 배치크기와 50,000와 500,000의 버퍼 크기로 이루어진 DDPG와 TD3의 성과 분석을 비교한 결과, 비슷한 성능을 보여주었다. 마지막으로 이미지 자료로 이루어진 환경을 가진 기존의 연구[8-10]에서는 DDPG가 TD3와 비교해 과대 평가되는 방향의 학습 경향을 보인다고 하지만 본 연구에서는 이런 결과를 보이지 않음을 확인할 수 있다. 이는 이미지 자료로 이루어진 환경에서 TD3의 쌍둥이 Q

신경망이 특징 계층을 공유하는 것이 유용할 수 있지만 본 연구에서의 환경은 이미지 자료가 아니기 때문에 DDPG와 비슷한 학습 성과를 보여준 것으로 고려된다. 더불어 TD3의 경우, 정책의 갱신을 지연시키며 두 개의 Q- 함수 중 최솟값을 선택적으로 사용하여 학습하기 때문에 급변하는 금융시장 상황에서의 대응이 어려울 수 있지만, DDPG는 다양한 상황별 대처가 가능하다[10].

인공신경망을 구성할 때, 은닉층의 수, 은닉층에 존재하는 유닛 개수, 활성화 함수, 신경망의 초기화 기법 그리고 최적화 등이 적절히 최적화하지 못하면 높은 결과를 이끌어 낼 수 없다. 이러한 인공신경망의 초매개변수들의 모든 가능한 조합을 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. 동일한 구조와 동일한 초매개변수를 갖는 인공신경망이라도 활성화 알고리즘에 따라 성능이 다르게 나타날 수 있으므로, 적합한 활성화 함수를 찾는 것은 매우 중요하다. 이에 본 논문에서는 ReLU, PReLU, GELU, SiLU 그리고 Mish 활성화 함수에 따른 DDPG와 TD3의 포트폴리오 학습성능에 미치는 영향을 검증하였다. 포트폴리오 학습성능을 확인한 결과, 벤치마크의 샤프지수가 1.72인 반면, DDPG와 TD3 알고리즘 기반의 샤프지수는 1.90~2.23으로 이들 알고리즘이 위험에 대한 초과수익률을 위한 포트폴리오 배분을 수행할 수 있음을 보여주었다. 하지만 ReLU, PReLU, GELU, SiLU 그리고 Mish 활성화 함수를 사용하여 DDPG와 TD3를 비교한 결과, 대부분의 활성화 함수에서 비슷한 성능을 보였다.

하지만 본 논문에도 향후 몇 가지 보완할 점이 필요하다. 기존의 많은 연구에서 DDPG가 TD3와 비교해 과대평가 되는 방향의 학습 경향을 보인다고 하지만 본 연구에서는 이런 결과를 보이지 않음을 확인할 수 있다. 이에 정형과 비정형 자료를 비교한 연구가 필요하다. 더불어 다양한 정책 신경망의 비교를 통해 더 높은 성과의 기대가 가능한 방안에 관해 후속 연구가 이루어질 필요가 있다.

참고문헌

[1] W. Lee, “Performance Evaluation of Portfolio using a Deep Q-Networks,” Journal of Next-generation Convergence Information Services Technology, vol.10, no. 4, pp. 459-470, (2021).
[2] H. Markowitz, “Portfolio Selection,” Journal of Finance,, pp. 77-91, (1952).
[3] G. P. Brinson, P. L. Randolph-Hood, and G. L. Beebower,, “Determinants of Portfolio Performance,” Financial Analysts Journal,, vol. 51, no. 1, pp. 133-138, (1955).
[4] I. Bajeux-Besnainou, . V. Jordan, and R. Portait, “Dynamic Asset Allocation for Stocks, Bonds, and Cash,” The Journal of Business, vol.76, no. 2, pp. 263-288, (2003).
[5] S. Kim, “Robo-Advisor Algorithm with Intelligent View Model”, Journal of intelligence and information systems, pp. 39-55, (2019).
[6] M. García-Galiciaab, A. A. Carsteanuab, and J. B. Clempnerab, “Continuous-time reinforcement learning approach for portfolio management with time penalization,” Expert Systems with Applications, vol. 7, no. 1, pp. 27-36, (2019).
[7] J. Lee, K. Kim, and J. Lee, “Singularity Avoidance Path Planning on Cooperative Task of Dual Manipulator Using DDPG Algorithm,” Journal of Korea Robotics Society, vol. 16, no. 2, pp. 137-146, (2021).
[8] D. Lee, and M. Kwon, “Combating Stop-and-Go Wave Problem at a Ring Road Using

Deep Reinforcement Learning Based Autonomous Vehicles," The Journal of Korean Institute of Communications and Information Sciences, vol. 46, no. 10, pp. 1667-1682, (2021).

[9] Y. Yoo, D. Kim, and J. Lee, "A Performance Comparison of Super Resolution Model with Different Activation Functions," KIPS Trans. Softw. and Data Eng, vol. 9, no. 10, pp. 303-308, (2020).

[10] Y. Kim, S. M. Hong, and J. Oh, "Design of Control Algorithm for Micro Electric Vehicle Suspension System Using Reinforcement Learning Algorithm," Transactions of the Korean Society for Noise and Vibration Engineering, vol. 32, no. 2, pp. 124-132, (2022).

[11] 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, pp. 117-122, (2021).

(접수: 2022.05.17. 수정: 2022.06.03. 게재확정: 2022.06.08.)