Actor-Critic 모델을 이용한 포트폴리오 자산 배분에 관한 연구

칼리나 바야르체첵[*1], 이주홍[*2], 송재원[**3]
[*]인하대학교 전기컴퓨터공학과
[**](주)밸류파인더스
[1]kb0422.bk@gmail.com, [2]juhong@inha.ac.kr, [3]jwsong@valuefinders.co.kr

A Study on Portfolio Asset Allocation Using Actor-Critic Model

Bayartsetseg Kalina[*], Ju-Hong Lee[*], Jae-Won Song[**]
[*]Dept. of Computer Engineering, Inha University
[**]ValueFinders Co., Ltd

요 약

기존의 균등배분, 마코위츠, Recurrent Reinforcement Learning 방법들은 수익률을 최대화하거나 위험을 최소화하고, Risk Budgeting 방법은 각 자산에 목표 리스크를 배분하여 최적의 포트폴리오를 찾는다. 그러나 이 방법들은 미래의 최적화된 포트폴리오를 잘 찾아주지 못하는 문제점들이 있다. 본 논문은 자산 배분을 위한 Deterministic Policy Gradient 기반의 Actor Critic 모델을 개발하였고, 기존의 방법들보다 성능이 우수함을 검증한다.

1. 서론

포트폴리오 자산 배분이란 개인의 목표를 달성하기 위한 위험과 수익률의 적절한 균형을 맞추어서 자산을 배분하는 투자전략을 말한다. 자산의 예로는 주식, 채권, 상품 및 현금 등이 있다. 포트폴리오 이론은 마코위츠(Markowitz)에 의해서 체계화되었다. 마코위츠 모델[3]은 공분산 행렬형태의 위험을 최소화하면서 포트폴리오의 기대 수익률을 최대화하는 평균-분산 최적화 방법이다. Risk budgeting[6]은 자본이 아닌 포트폴리오의 위험에 따라 자산을 할당하는 방법이다. 이 방법은 포트폴리오 매니저가 일련의 위험 예산(Risk Budget)을 정의한 다음 포트폴리오의 가중치를 계산한다. 균등배분(Equally weighted)은 포트폴리오 각 자산에 동일한 가중치를 부여하는 가중치 방법이다. 일반적으로 포트폴리오를 할당한다는 것은 투자 매니저의 의사결정 프로세스를 의미한다. 강화학습은 순차적인 의사결정 작업을 학습하는 일반적인 프레임워크다. Temporal difference 방법[8]으로 매개변수를 조정하여 시스템의 action에 따른 기대보상(expected reward)을 최대화한다. 자산 배분에서 포트폴리오 가중치는 강화학습으로 정의된다. RRL(Recurrent Reinforcement Learning) 알고리즘[4]은 샤프지수(Sharpe ratio)를 최소화하여 네트워크를 학습한다. RRL[4]이 포함된 자산 배분 시스템은 행동(action)의 value를 학습한 후, 측정된 행동들의 value를 기반으로 행동을 선택한다. 이러한 방법은 기대 수익률이나 샤프지수 같은 value 함수에 따라 달라진다. 이 문제점을 해결하기 위하여 DPG(Deterministic Policy Gradient)[7]를 사용하는 actor critic 모델을 제안한다.

2. 제안 방법

2.1 Problem Statement

먼저 포트폴리오 자산 배분을 위한 금융 시계열을 정의한다. 자산 가격 행렬 는 m개의 자산의 t 날 자산의 가격을 열 단위로 나타내었다.

$$P_{1:t} = [p^1 p^2 \dots p^m] = \begin{bmatrix} p_1^1 & p_1^2 \cdots & p_1^m \\ p_2^1 & p_2^2 \cdots & p_2^m \\ \cdots & \cdots \cdots & \cdots \\ p_t^1 & p_t^2 \cdots & p_t^m \end{bmatrix} \quad (1)$$

다른 방법과 마찬가지로 강화학습 시스템의 상태(state) 입력으로 과거 d일 동안의 자산 수익률 데이터를 사용한다.

$$s_t = \{z_{t-d:t}^1, \dots, z_{t-d:t}^m\} = Z_{t-d:t}$$

시간 t에서 i번째 자산의 수익률은 $z_t^i = \frac{p_t^i}{p_{t-1}^i} - 1$로 정의된다. 그래서 자산 수익률 행렬을 $Z_{2:t}$로 정의할 수 있다.

$$Z_{2:t} = \left[z_{2:t}^1 z_{2:t}^2 \cdots z_{2:t}^m\right] = \begin{bmatrix} z_2^1 & z_2^2 \cdots & z_2^m \\ z_3^1 & z_3^2 \cdots & z_3^m \\ \cdots & \cdots \cdots & \cdots \\ z_t^1 & z_t^2 \cdots & z_t^m \end{bmatrix} \quad (2)$$

강화학습 에이전트의 행동은 포트폴리오 가중치 $w \in \{w^1, w^2, ..., w^m\}$에 의해 정의되고, trader는 매수($w^i \geq 0$)만 사용한다. 본 논문에서는 샤프지수를 즉각적인 보상(immediate reward)으로 사용한다.

2.2 포트폴리오 자산 배분의 Actor Critic 모델

Actor 네트워크는 상태를 특정 행동에 결정적으로 대응하는 CNN[2]을 가진 LSTM[1]을 사용하고, 훈련 알고리즘은 DPG 방법을 사용한다.

$$\theta \leftarrow \theta + \alpha_\theta \frac{\partial u_\theta(s)}{\partial \theta} \frac{\partial Q_w(s,w)}{\partial w}\Big|_{w=u_\theta(s)} \quad (3)$$

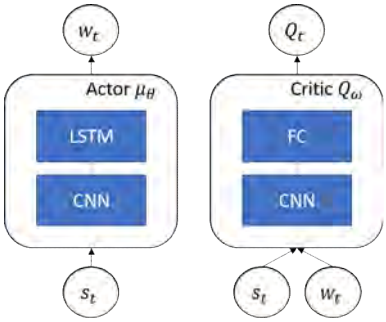
$Q_w(s,w)$는 critic 네트워크의 출력이고 α_θ는 학습률이다. Actor 네트워크가 $\sum_{i=1}^{m} w^i = 1$과 $w^i \geq 0$ 조건을 만족하기 위해 softmax 함수 식(4)를 사용한다.

$$w^i = \exp(y^i) / \sum_{j=1}^{m} \exp(y^j) \quad (4)$$

$y \in \{y^1, \ldots, y^m\}$ LSTM의 출력이고, CNN은 과거 수익률을 입력으로 사용하고, 중요한 특징(feature)을 찾아내서 출력한다. 출력값은 LSTM의 입력으로 사용된다. $Q_{w'}$을 타겟 critic 네트워크로 사용하고, 매개변수 w'와 experience replay로 critic 네트워크를 학습한다. Critic 네트워크의 에이전트는 상태(s_t)와 행동(w_t)를 입력으로 사용하여, $\hat{q} = Q_w(s_t, w_t)$로 정의된 q-value를 추정한다. 타겟 q-value는 시간 t에서 다음과 같이 정의된다.

$$q_t = r(s_t, w_t) + \gamma Q_{w'}(s_{t+1}, w_{t+1}) \quad (5)$$

여기서 $r(s_t, w_t)$는 즉각적인 보상(샤프지수), $Q_{w'}(s_{t+1}, w_{t+1})$는 타겟 critic 네트워크의 출력이다. 평균 제곱 오차(Mean Square Error)를 최소화하여 critic 네트워크를 학습한다.



<그림 1> Actor Critic 모델

$$L(w) = E_{s_t \sim p^u, w_t \sim u}[(q - Q_w(s_t, w_t))^2] \quad (6)$$

Gradient를 사용하여 critic 네트워크의 매개변수를 업데이트할 수 있다.

$$w \leftarrow w + \alpha_w \frac{\partial L(w)}{\partial \theta} \quad (7)$$

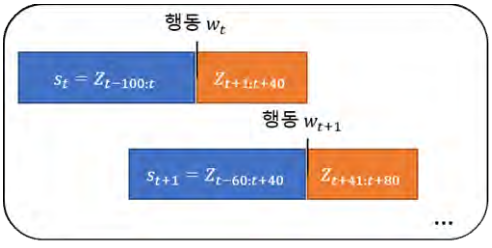
α_w는 학습률이다. Q 네트워크의 구조는 FC(fully connected) layer와 CNN을 사용하였다. Actor Critic 모델의 일반적인 프레임워크는 그림1과 같다.

	포트폴리오-A	포트폴리오-B	포트폴리오-C
Markowitz	0.0630	0.2448	0.8613
EquallyWeighted	-0.4643	-0.3111	0.4386
Risk Budgeting	0.2051	0.2084	1.0246
RRL	1.6197	1.4177	1.7444
Actor Critic	**1.7803**	**1.4736**	**1.7925**

<표 1> 실험결과 (샤프지수)

3. 실험

Actor Critic 포트폴리오 자산 배분 시스템은 일 단위 데이터로 학습하며, 제안된 모델의 특성을 평가하기 위해 3개의 포트폴리오(포트폴리오-A, 포트폴리오-B, 포트폴리오-C)를 사용한다. 각 포트폴리오는 총 10개의 자산으로 구성되어 있다. 2012년 1월부터 2019년 7월까지 총 8년의 일일 데이터(1970 trading days)를 사용한다. 2012년 1월부터 2019년 1월 사이의 데이터(1850 trading days)를 훈련 세트로 사용하고 다른 데이터(120 trading days)는 테스트 세트로 사용한다.



<그림 2> 자산배분 모델의 입력 및 포트폴리오 재조정

Benchmark 모델과 Actor Critic 모델의 상태 입력 값으로 과거 5개월(100 trading days) 동안의 자산 수익률 데이터를 사용한다.

$$s_t = \left\{ z^1_{t-100:t}, \cdots, z^{10}_{t-100:t} \right\}$$

두 달(40 trading days)에 한 번 모든 포트폴리오를 재조정한다. 그림2는 자산배분 모델들의 입력(과거 수익률 데이터)과 포트폴리오를 재조정한 기간을 보여준다. 본 논문에서는 모델의 학습을 위하여 0.001의 학습률을 가진 Adam Optimizer[5]을 사용한다. 포트폴리오-A는 니케이 225, 나스닥 종합, 코스피 200, 독일 DAX, S&P500 5개의 주가지수와 S-Oil, SK텔레콤, POSCO, 삼성전자, 한국전력 5개의 주식으로 구성된다. 포트폴리오-B는 니케이 225, 영국 FTSE 100, 코스피 200, 독일 DAX, S&P500 5개의 주가지수와 SK텔레콤, POSCO, 삼성전자, 한국전력, 현대차 5개의 주식으로 구성된다. 포트폴리오-C는 니케이 225, 나스닥 종합, 영국 FTSE 100, 독일 DAX, S&P500 5개의 주가지수와 SK텔레콤, POSCO, 삼성전자, S-Oil, 현대차 5개의 주식으로 구성되어 있다. 각 포트폴리오의 실험결과로 나온 샤프지수는 표1과 같다. 실험결과를 통해 Actor Critic 강화학습 모델이 가장 우수한 성능을 나타냄을 알 수 있다.

4. 결론

자산 배분을 위한 Actor Critic 강화학습 모델을 제시하였다. 강화학습을 적용한 모델들의 결과가 기존 방법들에 비해 매우 우수한 성능을 보인 점을 통해 강화학습이 자산 배분 문제에 적합함을 알 수 있었다. 시계열 데이터의 temporal dependency를 반영하기 위해 강화학습 모델에 LSTM을 적용함으로써 모델의 성능을 개선할 수 있었다. 그러나 샤프지수를 최대화하는 RRL 모델은 샤프지수의 영향을 많이 받는다는 단점이 발생하였다. 이를 해결하기 위하여 q-value 함수를 근사화하는 Actor Critic 모델을 제안하였다. Actor Critic 모델은 타겟 네트워크와 experience replay를 통하여서 훈련 모델을 효과적으로 학습한다.

감사의 글

이 논문은 2019년도 정부(과학기술정보통신부) 의 재원으로 한국연구재단의 기초연구사업(과제번호: 2019R1F1A1062094)과 정보통신기획평가원의 지원(과제번호: 2019-0-01124)을 받아 수행된 연구임

참고문헌

[1] Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation*, 9(8): 1735-1780, 1997.

[2] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *NIPS*, pp. 1106-1114, 2012.

[3] Markowitz, H. Portfolio selection. *The Journal of Finance*, 7(1): 77-91, 1952.

[4] Moody, J., Wu, L., Liao, Y., and Saffel, M. Performance functions and reinforcement learning for trading systems and portfolios. *Journal of Forecasting*, 17: 441-470, 1998.

[5] P.Kingma, D. and Ba, J. Adam: A method for stochastic optimization, 2014.

[6] Roncalli, T. Introduction to risk parity and budgeting, 2014.

[7] Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. Deterministic policy gradient algorithms. In *Proceedings of the 31st International Conference on Machine Learning* (ICML 2014), pp. 387-395

[8] Sutton, R. S. Learning to predict by the methods of temporal differences. *Machine Learning*, 3: 9-44, 1988.