RLTA: 강화학습을 이용한 AI 트레이딩 구현

강민지 [1], 최윤정 [2], 이지성 [3], 이규영 [4]
[1] 성균관대학교 시스템경영공학과
[2] 서울대학교 연합전공 인공지능
[3] 한서대학교 항공교통물류학부
[4] 한국과학기술원 전산학부 정보보호대학원

minji0801@g.skku.edu, racheal0@snu.ac.kr, ba5614@naver.com, leeahn1223@kaist.ac.kr

RLTA: Implementation of AI Stock Trading using Reinforcement Learning

Min-Ji Kang[1], Yun-Jeong Choi[2], JiSung Lee[3], Gyuyoung Lee[4]
[1]Dept. of Systems Management Engineering, SungKyunkwan University
[2]Interdisciplinary Major in Artificial Intelligence, Seoul National University
[3]Dept. of Air Transportation and Logistics, Hanseo University
[4]Graduate School of Information Security, KAIST

요 약

인류는 주가를 과학적으로 예측하기 위해 수많은 학문적 노력을 기울여왔지만, 아직까지도 풀지 못한 난제로 남아 있다. 이에 본 연구에서는 깊은 수학적 원리에 기반하고 알파고 등에서 인간을 능가하는 성능을 보여준 강화학습 기술을 주식 트레이딩에 적용한 RLTA 모델을 제안하고, 실험을 통해 그 유용성을 입증하였다.

1. 서론

과거부터 현재까지 주식시장에 대한 주가 변동 예측은 풀리지 않는 난제이다. 주가를 과학적으로 예측하기 위해 다양한 시도와 연구들이 있어왔지만, 아직까지 정확한 미래를 예측하는 것은 불가능하다. 하지만, 주가 예측은 경제, 수학, 물리 그리고 전산학 등 여러 관련 분야에서 오랜 관심의 대상이 되어왔다. [1]

그동안 RNN 및 LSTM AI 모델을 이용하여 시계열 회귀예측을 수행하는 트레이딩 시스템들이 등장하였으나, 과거의 추세를 그대로 추종하는 결과를 보이는 등 큰 효과를 거두지 못하였다. 반면 강화학습은 알파고, 게임 에이전트 등을 탄생시킨 강력한 AI 기술로서, 이를 주식 트레이딩에 응용하면 매수, 매도 등의 투자행위 자체를 결정해 주기 때문에, 사용에 편리하고 높은 수익을 창출할 수 있을 것으로 기대된다.

본 연구에서는 강화학습과 인공신경망 기술을 결합한 DQN(Deep Q-Network) 알고리즘을 사용하여 주가 트레이딩을 수행하는 RLTA(Reinforcement Learning Trading Agent) 모델을 제안한다. RLTA 는 과거 주가데이터를 학습한 후 트레이딩 시점마다 높은 누적가치를 지향하는 최적의 액션을 예측해 낸다.

본 논문에서는 RLTA 강화학습 모델의 구현과정을 설명하고 실험을 통해 그 유용성을 입증하고자 한다.

2. 제안 모델

본 연구에서 강화학습 환경(Environment)은 주식시장 종가의 일봉 데이터이고, 에이전트가 매수, 매도, 관망 액션 중에서 매 시점 마다 1 개를 결정하게 된다.

<표 1> RLTA 강화학습 트레이딩 수행단계

단계	단계명	핵심 내용
1	DQN 구축	입력단, 은닉층, 출력층 구성
2	에피소드 수집	탐험 & 탐사(벨만최적방정식)
3	정답값 산출	Temporal Difference Target
4	DQN 학습	Capture & Replay Separate networks
5	트레이딩 실시	매수, 매도, 관망 액션을 예측

<표 1> 내용을 세부적으로 설명하면 다음과 같다.

1 단계에서는 각 상태에서 가능한 액션(행동) 중 향후 기대 보상이 가장 큰 액션을 선택하는 DQN 가치

기반 강화학습 모델을 구축한다. 입력단의 피처수는 50 개이며, 종가, 거래량, 주식매입율 등 다양한 입력 자료로 구성하였다. 은닉층은 총 3 개이고, 뉴런개수는 각각 256 개, 64 개, 16 개로 설정하였다. 출력층은 매수와 매도 액션으로만 이루어져 있으며, 따라서 뉴런은 2 개로 구성하였다.

2 단계에서는 탐험(exploration)과 탐사(exploitation)를 통해 에피소드를 수집하며, 일정한 비율로 탐험과 탐사를 섞어서 진행한다. 탐험은 액션을 랜덤하게 선택하는 것으로서, 학습이 진행될수록 그 수행비율은 줄어든다.

탐사에서는 특정 상태(state)에서 특정 액션(action)을 취했을 때 얻을 수 있는 기대값인 액션가치함수의 리턴값이 가장 큰 액션을 선택한다. 벨만최적방정식 V(s) = max E[Rt+1 + γV (St+1) | St = s, At = a] 을 사용하며, 이는 액션을 선택할 때 확률적으로 선택하는 것이 아니라 최대값 연산자를 통해 가장 좋은 액션을 선택한다는 것을 뜻한다.[2]

3 단계에서는 에피소드 내 취득한 샘플 별로 정답을 계산하는데, 다음 상태 데이터의 예측값을 정답값으로 삼는 강화학습 TD(Temporal Difference) 방법론을 사용한다. 강화학습 Q-Network 는 구조적으로 지도학습 모델을 취하기 때문에 정답이 필요한데, 이를 Q-Network 에서 계산한 예측값으로 조달한다.

에피소드의 끝이 존재하는 경우에는, 에피소드 마다 그 길이가 다르고, 에피소드의 보상값이 있는 종료지점에서 역순으로 정답값을 산출해 나가는 MC(Monte Carlo) 리턴 방식을 사용한다. MC 는 업데이트를 하기 위해서는 리턴이 필요한데 리턴은 에피소드가 끝나기 전까지는 알 수 없으므로, 반드시 종료하는 MDP(Episodic MDP)에서만 사용할 수 있다. (예: 바둑, 게임 등)

반면, 주가데이터는 에피소드의 종료상태가 존재하지 않는 Non-Episodic MDP 이기 때문에, 일정 주기마다 샘플링이 일어나서 모든 에피소드의 길이가 동일한 특징이 있고, 따라서 종료상태 도착을 기다릴 필요 없이 각각의 상태전이 시점마다 다음 상태의 최적 액션가치를 대입하여 업데이트하는 TD 학습기법을 사용한다. TD 기법의 수식은 아래와 같고, 좌항 $Q(S_t, a_t)$ 은 다음 상태의 예측값으로서 해당 샘플의 정답값으로 사용된다.

$$Q(S_t, a_t) = Q(S_t, a_t) + \propto [r_{t+1} + \gamma max_a Q(s_{t+1}, a) - Q(S_t, a_t)]$$

4 단계에서는 시계열 인접 데이터 간 연관성(Correlation)을 완화하기 위한 “Capture & Replay” 기법과 정답값이 흔들리는 “Non-stationary targets” 문제를 해결하기 위한 “Separate networks” 기법을 적용하여 DQN 신경망을 학습한다.

5 단계에서는 학습을 완료한 DQN 신경망을 사용하여, 입력한 주가데이터에 대한 최적의 매수, 매도, 관망 액션을 예측하고 이를 실전 트레이딩에 투입한다.

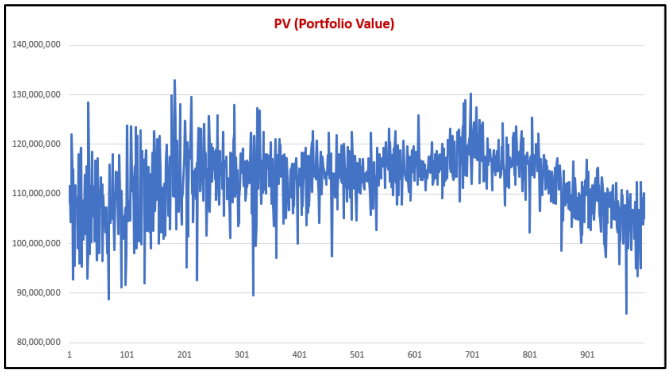
3. 실험

3.1. 데이터셋

2018.01.01 부터 2019.12.31 까지 2 년간의 삼성전자 주가데이터를 사용하여, DQN 모델을 학습하였다.

3.2. 실험결과

PV(Portfolio value)는 현재주가로 계산한 주식잔고의 평가액과 현금잔고를 합한 것으로서, 그림 1 은 투자 자본금 1 억원으로 시작하여 1,000 epoch 를 학습하였을 때 PV 의 변동을 그래프로 표시한 것이다. RLTA 에이전트는 미래를 알 수 없는 상황에서 해당 날짜에 가장 적합하다고 판단한 투자액션(매수, 매도, 관망)을 제시하였고, 이를 적용한 결과 그림 1 과 같이 전체 epoch 에 걸쳐 자본금인 1 억원을 상회하는 PV 분포(Max PV:133,156,143 원)를 보였으며, 이를 통해 강화학습 기반 RLTA 모델의 수익발생 가능성이 높다는 사실을 지속적으로 확인할 수 있었다.



(그림 1) 강화학습 트레이딩 모델의 PV 변화

4. 결론

본 연구에서는 강화학습 기술을 주식 트레이딩에 적용한 RLTA 모델을 제안하고, 실험을 통해 그 유용성을 입증하였다.

주식 트레이딩은 불연속한 액션을 선택하는 Q-Learning 기반의 DQN 이 적합하지만, 향후에는 로봇제어와 같은 연속적인 액션공간에서도 효과적인 정책기반 강화학습 기법을 함께 결합한 A2C(Advantage Actor-Critic), A3C(Asynchronous Advantage Actor-Critic) 등 액터-크리틱 강화학습 기법을 적용한 연구를 수행할 예정이다.

※ 본 논문은 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 ICT 멘토링 프로젝트 결과물입니다.

참고문헌

[1] 송유정, 이재원, 이종우, "텐서플로우를 이용한 주가 예측에서 가격-기반 입력 피쳐의 예측 성능 평가", 정보과학회 컴퓨팅의 실제 논문지 제 23 권 제 11 호, 2017.11

[2] 노승은, "바닥부터 배우는 강화학습", 영진닷컴, 2021.