**Ensemble strategy**

Tác giả sử dụng 3 actor-critic based algorithms thực hiện cái trading agent. 3 thuật toán đó là : A2C, DDPG và PPO.

* actor-critic là : kỹ thuật RL giúp cho việc học policy function và value function 1 cách độc lập. Policy function sẽ đưa ra quyết định hành động còn value function giúp cải thiện training process.

1. **Advantage Actor Critic (A2C):**

* A2C để cải thiện các policy gradient updates. A2C sử dụng advantage function để giảm sự phân tán của các policy gradient. Thay vì tính value function, các mạng critic sẽ tính advantage function. Do đó việc đánh giá một hành động không chỉ phụ thuộc vào mức độ tốt của hành động đó mà còn phụ thuộc vào hành động đó sẽ mang lại kết quả tốt như thế nào. Điều này làm giảm độ phân tán của các policy network.
* A2C sử dụng bản sao của các agent giống nhau để update gradients với các mẫu dữ liệu khác nhau. Mỗi agent sẽ hoạt động độc lập trong cùng 1 môi trường. Kết thúc mỗi vòng lặp các agent sẽ tính toán gradient của nó. A2C sẽ sử dụng 1 bộ điều phối (coordinator) để tính gradient trung bình thông qua tất cả các agent trong global network. Vì vậy global network sẽ update actor and the critic network. Sự hiện diện của global network làm tăng tính đa dạng của dữ liệu đào tạo. Từ đó gradient update được đồng bộ hóa tiết kiệm chi phí hơn, nhanh hơn và hoạt động tốt hơn với batch size lớn.
* A2C là mô hình tuyệt vời cho giao dích chứng khoán vì sự ổn định của nó
* The objective function for A2C is:
* là policy network, là Advantage function

Or

1. **Deep Deterministic Policy Gradient (DDPG):**

* DDPG dùng để khuyến khích đầu tư tối đa. DDPG là sự kết hợp của Q-learning và policy gradient và sử dụng neural networks làm function approximators. Trái với DQN là học gián thông qua Q-values tables và gặp vấn đề về kích thước. DDPG học trực tiếp từ các observations thông qua policy gradient. DDPG lập bản đồ xác định trạng thái cho các hành động để phù hợp hơn trong không gian các hành động liên tục (continuous action space environment).
* Mỗi 1 bước, DDPG agent thực hiện 1 action tại trạng thái , nhận được reward và tiến đến trạng thái . Transitions được luu trữ trong replay buffer ***R.*** Một dây chuyền của **N** transitions được rút ra từ buffer ***R*** và Q-value được thực hiện như sau:
* Sau đó critic network sẽ được updated bằng việc minimizing hàm loss function , difference kỳ vọng giữa outputs của target critic network và critic network :
* DDPG hoạt động hiệu quả trong việc xử lí continuous action space và nó rất thích hợp cho việc kinh doanh chứng khoán.

1. **Proximal Policy Optimization (PPO):**

* PPO dùng để kiểm soát policy gradient update và đảm bảo là policy mới sẽ không khác biệt quá nhiều so với policy trước đó. PPO cố gắng đơn giản hóa objective của Trust Region Policy Optimization (TRPO) bằng việc nhét clipping term vào objective function.
* Tỉ lệ giữa old và new policy như sau:
* Hàm clipped thay thế cho objective function của PPO là:
* là policy gradient objective bình thường, là advantage function dự đoán. Công thức sẽ clip tỉ lệ trong khoảng []. Objective function sẽ lấy giá trị tối thiểu giữa clipped và normal objective. PPO không khuyến khích việc thay đổi large policy bên ngoài thời gian clipped. Do đó PPO cải thiện tính ổn định của mạng policy network bằng cách hạn chế policy update ở mỗi training step. PPO được sử dụng trong giao dích chứng khoán bởi vì nó ổn định, nhanh chóng, dễ triển khai và điều chỉnh.

1. **Ensemble Strategy:**

* Ensemble Strategy sẽ tự động chọn agent hoạt động tốt nhất trong 3 agent PPO, A2C và DDPG để giao dịch dựa trên Sharpe ratio. Quá trình tổng hợp như sau:
* **Bước 1:** Sử dụng 1 khoảng thời gian *n* months để retrain cả 3 agent cùng lúc. Trong bài báo này là 3 months
* **Bước 2:** Sau khi train xong thì sẽ đánh giá cả 3 agent bằng 1 một tập test 3-month validation để chọn ra agent cho kết quả Sharpe ratio tốt nhất. Công thức tính Sharpe ratio:

là tỉ suất lợi nhuận của danh mục đầu tư

là tỉ suất lợi nhuận phi rủi ro

là độ lệch chuẩn của tỉ suất lợi nhuận vượt quá của danh mục

Để phù hợp với việc risk-aversion (ngại rủi ro) thì trong bước validation có những chỉ số hỗn loạn.

* **Bước 3:** Sau khi chọn được agent tốt nhất thì sẽ dùng nó để dự đoán cho việc giao dịch cổ phiếu ở quý tiếp theo. Lí do cần phải lựa chọn 1 trong 3 agents mỗi tháng là vì mỗi loại agent sẽ nhạy cảm với 1 loại xu thế khác nhau. Có agent hoạt động tốt trong xu hướng tăng nhưng hoạt động xấu trong xu hướng giảm. Một agent khác thì điều tiết tốt trong 1 thị trường biến động. Sharpe ratio của agent càng cao thì lợi nhuận càng cao so với mức rủi ro đầu tư mà nó đã thực hiện. Do đó việc chọn ra này sẽ có thể tối đa hóa được lợi nhuận đầu tư điều chỉnh theo rủi ro.