Tạp chí Khoa học Trường Đại học Cần Thơ
Phần D: Kinh tế và Pháp luật
website: ctujsvn.ctu.edu.vn

DOI:10.22144/ctu.jvn.2023.151

THUẬT TOÁN HỌC TĂNG CƯỜNG CẢI TIẾN DỰA TRÊN XU HƯỚNG DỮ LIỆU ĐỂ RA QUYẾT ĐỊNH MUA BÁN TRÊN THỊ TRƯỜNG TIỀN ĐIỆN TỬ

Trần Kim Toại[1*], Võ Minh Huân[1], Lê Ngọc Thanh[2] và Võ Thị Xuân Hạnh[1*]

[1]*Trường Đại học Sư phạm Kỹ thuật Thành phố Hồ Chí Minh*

[2]*Đại học Kiến trúc Đà Nẵng*

**Người chịu trách nhiệm về bài viết: Trần Kim Toại (email: toaitk@hcmute.edu.vn);*
Võ Thị Xuân Hạnh (email: hanhvtx@hcmute.edu.vn)

Thông tin chung:
Ngày nhận bài: 17/12/2022
Ngày nhận bài sửa: 03/03/2023
Ngày duyệt đăng: 03/03/2023

Title:
Enhanced reinforcement learning based on trend line for trading decisions on cryptocurrency

Từ khóa:
Chiến lược mua bán, học tăng cường, tiền điện tử, xu hướng giá

Keywords:
Cryptocurrency, reinforcement learning, trading strategy, trend line

ABSTRACT

The use of a machine learning algorithm in combination with the data on cryptocurrency market price trends to make buying and selling decisions is studied in this paper. We suggest that agents combined with data trends should be used to make financial decisions instead of just using reinforcement learning. The research question is to answer the problem that can reinforcement learning (RL) win the financial market? RL that makes market-based buying and selling decisions can win in terms of profitability and limit risks? Results suggest that agents combined with data trends should be used to make financial decisions instead of just using reinforcement learning. Various financial indicators are applied such as Maximum Drawdown, Annual Profit, and Accuracy. Analysis results are conducted on two datasets of Bitcoin and Dogecoin, finding that the RL based on the trendline has more advantages than the RL without the trendline based on different performance indicators.

TÓM TẮT

Việc dùng thuật toán máy học với sự kết hợp dữ liệu đường xu hướng giá của thị trường tiền điện tử để ra quyết định mua bán được nghiên cứu trong bài viết. Thay vì chỉ sử dụng mô hình học tăng cường để thực thi hành động trong môi trường tài chính, học tăng cường kết hợp với xu hướng dữ liệu để ra quyết định hành động. Nghiên cứu trả lời cho câu hỏi dùng học tăng cường có thể chiến thắng được thị trường tài chính hay không? Học tăng cường tự ra các quyết định mua bán dựa trên thị trường có mang lại lợi nhuận cho nhà đầu tư, giúp giảm rủi ro đầu tư hay không? Kết quả nghiên cứu cho thấy các tác nhân được kết hợp với xu hướng dữ liệu nên được dùng để ra quyết định tài chính thay vì chỉ sử dụng học tăng cường. Các thước đo tài chính về mức sụt giảm tối đa, lợi nhuận hằng năm, độ chính xác được dùng để đánh giá. Kết quả phân tích được thực hiện trên hai tập dữ liệu là Dogecoin và Bitcoin chỉ ra thuật toán học tăng cường dựa trên đường xu hướng có ưu điểm hơn so với học tăng cường không theo đường xu hướng trong các khía cạnh sử dụng chỉ số đánh giá hiệu năng khác nhau.

1. GIỚI THIỆU

Trong lĩnh vực trí tuệ nhân tạo (Artificial Intelligence - AI) nói chung và lĩnh vực học máy nói riêng (Machine Learning - ML), học tăng cường (Reinforcement learning - RL) là nhiệm vụ học từ dữ liệu không được gắn nhãn và mục tiêu của nó là phân cụm (Li, 2017). RL là một kỹ thuật trong AI, trong đó các tác nhân (agent) tương tác với môi trường (environment) thông qua các hành động (action). Một trạng thái (state) được cung cấp cho môi trường và tác nhân chọn một hành động dựa trên trạng thái đó để tối đa hóa phần thưởng. Tác nhân tìm hiểu thông qua các trạng thái và hành động để tối đa hóa phần thưởng của nó. RL tập trung vào việc làm thế nào để một tác nhân trong một môi trường có thể hành động sao cho lấy được phần thưởng nhiều nhất có thể. Học tăng cường không có cặp dữ liệu có gán nhãn trước làm đầu vào và cũng không có đánh giá các hành động là đúng hay sai (Fischer, 2018).

Công nghệ tài chính còn gọi là công nghệ Fintech đang phát triển nhanh. Mục tiêu của Fintech dựa trên điểm mạnh của công nghệ để cải tiến các hành động trong lĩnh vực tài chính (Mhlanga, 2020; Yuan & Jing, 2018**)**. Thời gian gần đây, công nghệ Fintech được mong đợi sẽ làm thay đổi cách ra quyết định liên quan đến các lĩnh vực tài chính như hoạt động mua bán, hoạt động đầu tư, quản lý rủi ro, quản lý danh mục đầu tư, tư vấn tài chính (Chopra & Sharma, 2021**)**. Vấn đề ra quyết định thì rất phức tạp để tìm ra hành động bởi vì tính chất ngẫu nhiên và thay đổi đột ngột của dữ liệu. Vì vậy, xây dựng thuật toán mua bán là quan trọng và cũng là thách thức trong nền công nghiệp Fintech. Mục tiêu chính của thuật toán mua bán là để trả lời câu hỏi cách nào thiết kế một thuật toán dựa dựa trên thuật toán AI để có thể chiến thắng trong lĩnh vực tài chính (Cao, 2020, 2021).

Thị trường tài chính bao gồm các loại giao dịch trên các loại tài sản như cổ phiếu, vàng, tiền điện tử, tài sản cố định… Mục đích của thị trường là thực hiện giao dịch mua và bán nhằm tìm kiếm lợi nhuận (Jay et al., 2020). Dự đoán giá tiền điện tử là một thách thức của lĩnh vực tài chính bởi vì sự thay đổi giá của nó phụ thuộc vào những sự ảnh hưởng của tình hình kinh tế, chính trị trong nước và ngoài nước. Điều này làm cho việc dự báo giá trở nên phức tạp. Dự báo giá vẫn còn là một thách thức lớn đối với những nhà đầu tư. Vì vậy, nó thu hút nghiên cứu từ nhiều lĩnh vực nghiên cứu như kỹ thuật tài chính, thống kê và cả mô hình học máy. Những năm gần đây, mô hình máy học được nghiên cứu và phát triển trên nhiều ứng dụng, với các mô hình thuật toán có thể giải quyết được mức độ phức tạp của dữ liệu ngày càng cao (Culkin, 2017). Bắt đầu từ các phương pháp máy học cổ điển như phương pháp hồi quy tuyến tính, hồi quy Ridge (Toai et al., 2022), cây quyết định, rừng cây ngẫu nhiên (Braham et al., 2022), ARIMA (Toai et al., 2022). Dần dần, các phương pháp học sâu đưa ra các kết quả hiệu quả và dự đoán có độ chính xác cao hơn so với các phương pháp máy học cổ điển (OECD, 2021). Những kỹ thuật dự báo tài chính thường được chia làm hai nhánh, phân tích kỹ thuật và phân tích cơ bản. Phương pháp phân tích cơ bản dựa trên các yếu tố của nền kinh tế như các báo cáo tài chính, lãi suất ngân hàng, mô hình kinh doanh của công ty, kinh tế vĩ mô ảnh hưởng tới xu hướng của thị trường tài chính. Phương pháp phân tích kỹ thuật dùng các chỉ số để tính toán đặc trưng dữ liệu ở quá khứ để dự báo xu hướng tương lai của thị trường (Shah et.al, 2019; Singh & Khushi, 2021). Phương pháp phân tích cơ bản ngày nay cũng thường sử dụng máy học để dự đoán hành vi của thị trường, giúp nhà đầu tư biết được thị trường đang tăng hay giảm để đưa ra chiến lược mua bán. Yuxuan et al. (2021) đã sử dụng mạng noron, random forest, mạng fuzzy để phân tích báo cáo tài chính hàng quý để xác định các yếu tố ảnh hưởng tích cực hoặc tiêu cực tới xu hướng. Nhiều nghiên cứu sử dụng các mô hình máy học khác nhau để dự đoán xu hướng thị trường dựa trên dữ liệu theo thời gian quá khứ để biết được hành vi thị trường trong tương lai để ra quyết định nhằm thu lợi nhuận cao nhất trong một khoảng thời gian xác định (Shahi et al., 2020; Tsung-Jung et al., 2021; Al-Sulaiman, 2022). Một ví dụ như chiến lược mua bán sử dụng mạng học sâu DNN kết hợp phương pháp biến đổi wavelet với mạng noron hồi quy (Tsung-Jung et al., 2021), sử dụng mạng LSTM (Shahi et al., 2020) và nhiều nghiên cứu khác.

Tuy nhiên, những nghiên cứu dựa trên dự báo xu hướng dữ liệu ở trên gặp khó khăn. Các phương pháp máy học này chỉ xem xét dữ liệu quá khứ với hành vi dữ liệu phi tuyến tính, không hỗn độn của thị trường tài chính, vì vậy chịu đựng vấn đề quá khớp (overfitting) khi đưa vào dự báo giá trị thực tế (Al-Sulaiman, 2022). Nhiều nghiên cứu sử dụng các phương pháp kỹ thuật học sâu khác nhau để tối thiểu hóa các vấn đề dự đoán giá tài chính này (Olorunnimbe & Viktor, 2022). Hơn nữa, việc dự báo giá cũng khó để đưa chi phí giao dịch vào để đưa ra quyết định. Ngoài ra, phương pháp phân tích kỹ thuật cũng không sử dụng các quyết định được thực thi trước đó vào để dự báo cho các hành động kế tiếp (Kabbani & Ekrem 2022).

Kỹ thuật học tăng cường được dùng để vượt qua các khó khăn từ mô hình dự báo giá này. Ở học tăng cường, mục tiêu của tác nhân không dựa trên các mẫu dữ liệu qua việc gắn nhãn mà cố gắng thực hiện các hành động để đạt hàm phần thưởng lớn nhất qua các vòng lặp huấn luyện. Nhiều nhà nghiên cứu cũng đề xuất thuật toán dùng kỹ thuật học tăng cường để giải quyết vấn đề mua bán này. Deng et al. (2017) đã đưa ra kiến trúc mô hình mạng học mờ (fuzzy network) để lấy đặc trưng kỹ thuật dữ liệu nhằm giảm sự không chắc chắn của dữ liệu theo thời gian kết hợp với học tăng cường để dự báo giá tài chính. Học tăng cường với đa tác nhân cùng với phương pháp học tập hợp để chọn ra hành động tối ưu từ đa tác nhân này để ra quyết định đầu tư mà không cần sử dụng mô hình dự báo cũng đã được nghiên cứu (Carta et al., 2020; Yang et al., 2020). Hai nhà nghiên cứu này sử dụng thuật toán học tăng cường sâu (Deep RL) để dự báo thị trường tài chính (Carta et al., 2020; Yang et al., 2020). Cả Q-learning và deep Q learning là những thuật toán của học tăng cường được sử dụng để giải quyết vấn đề mà một tác nhân tương tác với môi trường để học một nhiệm vụ nào đó. Deep Q-learning bao gồm Q-learning cùng với mạng neuron học sâu để xấp xỉ hàm Q. Với các vấn đề lớn, bao gồm không gian trạng thái phức tạp và nhiều trạng thái, mô hình deep Q-learning có khả năng xử lý hiệu quả ứng dụng này. Q-learning dùng một bảng để chứa các giá trị Q cho mọi hành động-trạng thái tương ứng. Q-learning thường giải quyết tốt các vấn đề với không gian trạng thái nhỏ. Phương pháp mua bán dựa theo xu hướng với sự phân tích chỉ số kỹ thuật cùng kết hợp với Q-learning để ra quyết định mua bán (Jagdish & Manish, 2019).

Nghiên cứu này đề xuất một phương pháp học tăng cường cải tiến để dự đoán hành động trong giao dịch thị trường tiền ảo nhằm tăng lợi nhuận cho nhà đầu tư. Quy trình ra quyết định của Markov được áp dụng bằng cách kết hợp với mô hình dự đoán xu hướng để hỗ trợ mô hình học tăng cường nhằm ra quyết định hiệu quả hơn, giúp giao dịch thành công và đạt lợi nhuận cao hơn. Tại một thời điểm, hệ thống sẽ đưa ra quyết định chọn hành động để đạt được giá trị phần thưởng lớn nhất. Mô hình học tăng cường kết hợp với xu hướng giá trong dữ liệu chuỗi thời gian được đề xuất để đưa ra các hành động mua hoặc bán hoặc giữ để đạt được lợi nhuận cao nhất cho nhà đầu tư.

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Markov Decision Process – MDP

Quy trình quyết định Markov là một quá trình kiểm soát ngẫu nhiên theo thời gian rời rạc. Nó cung cấp một khung toán học để mô hình hóa việc đưa ra quyết định trong các tình huống mà kết quả là một phần ngẫu nhiên và một phần nằm dưới sự kiểm soát của người đưa ra quyết định.

Mỗi trạng thái trong một môi trường là hệ quả của trạng thái trước đó. Tuy nhiên, việc lưu trữ tất cả các thông tin này, ngay cả đối với môi trường có các tập ngắn, việc này trở nên không dễ dàng và khả thi. Để giải quyết vấn đề này, giả định được đưa ra rằng mỗi trạng thái tuân theo một thuộc tính Markov, tức là mỗi trạng thái chỉ phụ thuộc vào trạng thái trước và sự chuyển đổi từ trạng thái đó sang trạng thái hiện tại (Li, 2017).

MDP bao gồm các biến: **S, A, P, R, γ**.

- **S** là một tập hợp các trạng thái, thường là hữu hạn.

- **A** là một tập hợp các hành động, thường là hữu hạn.

- $P(s, s', a) = P(s_{t+1} = s' | s_t = s, a_t = a)$ là hàm chuyển trạng thái, mô tả xác suất để chuyển từ trạng thái hiện tại sang trạng thái mới với hành động $\mathbf{a_t}$.

- $R(s, s', a) \in R$ là phần thưởng nhận được ngay lập tức sau khi chuyển từ trạng thái **s** đến trạng thái **s',** tiếp với hành động **a**.

- $\gamma \in [0,1]$ được gọi là hệ số chiết khấu và xác định việc tập trung vào phần thưởng.

Chính sách $\boldsymbol{\pi}$ là một hàm có tính xác suất $\pi: S \rightarrow A$.

Môi trường sử dụng là môi trường ngẫu nhiên. Hệ số chiết khấu cho phép đánh giá hàm phần thưởng. Tác nhân sẽ hoạt động tốt nếu chọn hành động tối đa hóa được hàm phần thưởng trong tương lai ở các bước tiếp theo. Phần thưởng trong tương lai được tính bằng công thức (1) như sau. Ở đây r_t là phần thưởng tại thời điểm t, r_n là phần thưởng tại thời điểm n.

$$R_t = r_t + \gamma r_{t+1} + \cdots + \gamma^{n-t} r_n = r_t + \gamma R_{t+1} \quad (1)$$

Hàm **V** hay còn được gọi là hàm giá trị trạng thái, thể hiện mức độ tốt nhất của trạng thái khi chuyển sang một trạng thái khác. Hàm **V** phụ thuộc vào trạng thái **s** hiện tại và tuân theo chính sách $\boldsymbol{\pi}$ mà tác nhân tuân theo, được đưa ra bởi công thức (2).

$$V_\pi(s) = E\left(\sum_{t\geq 0}\gamma^t r_t\right), \forall s \in S \quad (2)$$

Chúng ta có thể suy ra hàm giá trị trạng thái tối ưu có giá trị cao nhất cho tất cả các trạng thái (3).

$$V_*(s) = \max_\pi V_\pi(s)\ \forall s \in S \quad (3)$$

Tác nhân không thể trực tiếp kiểm soát được trạng thái khi nó kết thúc, tác nhân có thể tác động bằng cách chọn một số hành động **a**. Hàm **Q** là hàm giá trị hành động, thể hiện chất lượng của một hành động nhất định cho một trạng thái, sẽ trả về tổng phần thưởng mong đợi bắt đầu từ trạng thái **s,** lấy hành động **a** và tuân theo chính sách $\boldsymbol{\pi}$, được ký hiệu là $\boldsymbol{Q_\pi(s,a)}$**.** Tương tự hàm **V**, chúng ta có thể xác định được hàm **Q** tối ưu $\boldsymbol{Q_*(s,a)}$ mang lại tổng phần thưởng dự kiến tối ưu cho tác nhân bắt đầu từ trạng thái **s** cùng với hành động **a**. Mối liên hệ giữa hàm $\boldsymbol{V_*}$ tối ưu và hàm $\boldsymbol{Q_*}$ tối ưu được thể hiện trong công thức (4) như sau :

$$V_*(s) = \max_a Q_*(s,a)\ \forall s \in S \quad (4)$$

Tức là, tổng phần thưởng dự kiến tối đa có thể đạt được khi bắt đầu ở trạng thái **s** là giá trị lớn nhất của $\boldsymbol{Q_*(s,a)}$ trên toàn bộ các hành động có thể xảy ra. Sử dụng hàm $\boldsymbol{Q_*}$ tối ưu, chúng ta có thể suy ra chính sách tối ưu $\boldsymbol{\pi_*}$ bằng cách chọn hành động **a** mang lại phần thưởng tối đa $\boldsymbol{Q_*(s,a)}$ ở trạng thái **s** trong công thức (5) (Li, 2017).

$$\pi_*(s) = \arg\max_a Q_*(s,a)\ \forall s \in S \quad (5)$$

Chúng ta đã xác định được mối liên lệ giữa tất cả các trạng thái và cặp trạng thái và hành động. Bây giờ, nếu ta xác định mối liên hệ giữa hàm $\boldsymbol{V_*}$ và $\boldsymbol{Q_*}$, chúng ta có thể xây dựng một tác nhân tương tác với môi trường.

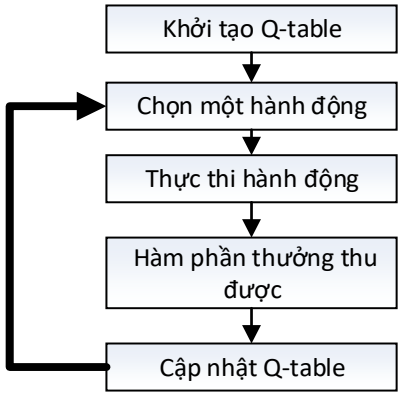
2.2. Thuật toán học với Q- LEARNING

Quá trình ra quyết định Markov mô hình hóa vấn đề ra chuỗi quyết định. Đặc trưng quan trọng nhất của MDP là sự chuyển trạng thái và hàm phần thưởng chỉ phụ thuộc vào trạng thái hiện tại và hành động được thực hiện. Mô hình MDP bao gồm hàm phần thưởng **R** và hàm chuyển trạng thái **P**. Trong trường hợp hai hàm này được xác định cho trạng thái kế tiếp với hành động tương ứng, MDP được thực thi như một phương pháp lập trình động. Tuy nhiên, trong hầu hết trường hợp hàm **P** và **R** không thể dự đoán một cách chính xác. Khi đó Q-learning là một thuật toán được dùng để giải quyết vấn đề của MDP với hàm phần thưởng và hàm chuyển trạng thái. Ý tưởng chính của Q-learning là để khám phá ra các cặp hành động-trạng thái và ước lượng hàm phần thưởng nhận được bởi áp dụng một hành động trong một trạng thái.

Q-learning là một thuật toán học tăng cường không chính sách. Q-learning học từ các hành động được thực hiện ngẫu nhiên dựa trên bảng Q, Q-table. Bảng Q-table là một bảng tra cứu đơn giản, nơi phần thưởng dự kiến tối đa được tính toán trong tương lai cho một hành động ở mỗi trạng thái. Về cơ bản, bảng này sẽ hướng dẫn chúng ta quyết định hành động tốt nhất ở mỗi trạng thái. Mỗi điểm Q-table sẽ tạo ra phần thưởng dự kiến tối đa trong tương lai sẽ nhận được khi thực hiện hành động ở trạng thái đó. Đây là một quá trình lặp đi lặp lại, vì ta cần cải thiện bảng Q-table ở mỗi lần lặp lại.

Để tìm hiểu từng giá trị của Q-table, ta sử dụng thuật toán Q-learning**.** Mô hình Q-learning thay vì dựa trên giá trị các trạng thái của hàm **V(s)** để đưa ra quyết định về hành động, Q-learning tập trung vào việc đánh giá chất lượng của hành động được phát triển từ hàm giá trị hành động.

Có một quá trình lặp đi lặp lại để cập nhật các giá trị. Khi bắt đầu**,** hàm Q cung cấp cho chúng ta các giá trị gần đúng hơn và tốt hơn bằng cách liên tục cập nhật các giá trị Q trong bảng.



Hình 1. Quy trình giải thuật của Q-learning. Sau nhiều lần lặp lại, chúng ta thu được Q-table tối ưu (Li, 2017)

Quy trình của Q-learning thể hiện trong ***Hình 1*** bao gồm các bước: Bước 1 là khởi tạo Q-table. Đầu

tiên, bảng Q-table được xây dựng, bao gồm n cột, trong đó n là số hành động. Trong nghiên cứu này n=3, với ba hành động a_0=mua, a_1=bán và a_3=giữ. m là số hàng, trong đó m là số trạng thái. Chúng ta sẽ khởi tạo các giá trị trong bảng Q-table ban đầu bằng 0 được minh họa qua **Bảng 1**.

Bảng 1. Giá trị khởi tạo của Q-table (Li, 2017)

State \ Action	a_0	a_1	a_2
s_0	$Q(s_0, a_0)$	$Q(s_0, a_1)$	$Q(s_0, a_2)$
s_1	$Q(s_1, a_0)$	$Q(s_1, a_1)$	$Q(s_1, a_2)$
s_2	$Q(s_2, a_0)$	$Q(s_2, a_1)$	$Q(s_2, a_2)$
...	...	...	...

Bước 2 và 3 dùng để chọn và thực hiện một hành động. Mô hình sẽ chọn một hành động tại một trạng thái dựa trên Q-table và thực hiện nó. Tuy nhiên, tại thời điểm ban đầu, mọi giá trị Q đều bằng 0. Chúng ta sẽ sử dụng kỹ thuật **epsilon greedy** để tìm các giá trị Q này. Trong thời gian đầu, **epsilon** sẽ cao hơn hàm ngẫu nhiên **random**, hàm này có chức năng gieo ngẫu nhiên các giá trị trong khoảng từ 0 đến 1. Mô hình sẽ lựa chọn ngẫu nhiên một trong các hành động để khám phá môi trường. Khi thực hiện nhiều lần, **epsilon** giảm dần, đến mức giá trị nhất định, khi đó hàm **random** sẽ cho ra giá trị cao hơn. Khi đó, mô hình sẽ chọn hành động có giá trị Q cao nhất theo từng trạng thái, đồng thời đây là lúc mà bảng Q-table dần ổn định và chính xác hơn.

Bước 4 và 5 dùng để đánh giá kết quả. Bây giờ mô hình đã thực hiện xong một hành động, bước tiếp theo sẽ là quan sát kết quả và phần thưởng. Chúng ta có thể biểu diễn **Q (s, a)** một cách đệ quy theo giá trị của trạng thái **s'** tiếp theo theo công thức (6).

$$Q(s,a) = r(s,a) + \gamma \max_{a'} Q(s',a') \qquad (6)$$

Phương trình (6) được gọi là phương trình Bellman, thể hiện phần thưởng tối đa trong tương lai mà tác nhân nhận được khi ở trạng thái hiện tại cộng với phần thưởng tối đa trong tương lai cho trạng thái **s'** tiếp theo.

Mục tiêu chính của việc học Q-learning là chúng ta lặp đi lặp lại để tiệm cận tới hàm $\boldsymbol{Q}_*$ dùng phương trình Bellman trên. Phương trình của hàm Q-learning được cho bởi công thức (7):

$$New\ Q(s,a) = \qquad (7)$$
$$Q(s,a) + \alpha\,[r(s',a) + \gamma\, maxQ'(s',a') - Q(s,a)]$$

Trong đó: $New\ Q(s,a)$: Giá trị Q mới cho trạng thái và hành động vừa thực hiện.

α : là tốc độ học (Learning rate).

$r(s',a)$: Phần thưởng đạt được khi thực hiện hành động a tại trạng thái s'.

$Q(s,a)$: Giá trị Q hiện tại.

$maxQ'(s',a')$ là phần thưởng mong đợi lớn nhất với tất cả các hành động khả thi trong trạng thái **s'** mới.

Mô hình sẽ lặp lại điều này nhiều lần cho đến khi quá trình học được dừng lại. Bằng cách này, Q-table sẽ được tối ưu. Sau khi tác nhân khám phá càng nhiều về môi trường, giá trị Q xấp xỉ hội tụ về $\boldsymbol{Q}_*$.

2.3. Giao dịch dựa theo xu hướng dữ liệu

Xu hướng dữ liệu là một chiến lược giao dịch, trong đó các quyết định giao dịch như mua và bán được thực hiện theo xu hướng giá trong thị trường tài chính. Xu hướng dữ liệu không thực hiện việc dự đoán giá, xu hướng dữ liệu chỉ xác định xu hướng thị trường và động lượng hiện tại của thị trường. Khi mô hình RL giao dịch dựa trên xác định xu hướng, các giao dịch được thực hiện theo chiến lược xác định trước. Bởi theo dõi liên tục xu hướng dữ liệu, nếu có sự thay đổi về xu hướng dữ liệu, mô hình thực hiện hoán đổi vị trí giao dịch từ hành động mua sang hành động bán hoặc ngược lại từ hành động bán sang hành động mua. Phương pháp để xác định xu hướng của dữ liệu phổ biến và được sử dụng trong phân tích tài chính như các chỉ báo phân tích kỹ thuật ví dụ đường trung bình động giản đơn **(**Simple Moving Average), đường trung bình (moving average), phân tích chỉ số sức mạnh (Relative Strength Index), chỉ số kênh hàng hóa (Commodity Channel Index - CCI) (Jagdish & Manish, 2019). Các phân tích kỹ thuật này có lợi trong các giao dịch ngắn hạn, như giao dịch có thời gian ngắn nắm giữ khác nhau từ một phút, một giờ, một ngày hoặc một tuần, nhiều nhất là vài tuần. Phân tích kỹ thuật cung cấp các chỉ báo cho biết xu hướng giá hiện tại. Nếu xu hướng động lượng đi lên, chúng ta tiên lượng giá tiền điện tử sẽ tăng. Bởi vậy, chúng ta quyết định thực hiện hành động mua. Ngược lại, nếu xu hướng động lượng giảm, chúng ta sẽ thực hiện hành động bán. Ngược lại, nếu thấy xu hướng không tăng, không giảm, chúng ta sẽ giữ lại, không thực hiện hành động mua hoặc bán trong thời điểm này. Trong nghiên cứu này, phương pháp được đề xuất là xác định xu hướng dữ liệu dựa trên dự báo giá của thị trường. Nếu dự báo giá tương lai lớn hơn giá hiện tại, có nghĩa là xu hướng dữ liệu đang tăng,

ngược lại nếu giá tương lai nhỏ hơn giá hiện tại, xu hướng dữ liệu đang giảm.

3. THƯỚC ĐO ĐÁNH GIÁ MÔ HÌNH BỞI CÁC CHỈ SỐ ĐỊNH LƯỢNG

3.1. Mức sụt giảm tối đa (Maximum Drawdown) – MDD

MDD là khoản lỗ tối đa quan sát được từ đỉnh đến đáy của danh mục đầu tư. MDD là một chỉ báo về rủi ro giảm giá trong một khoảng thời gian xác định và chỉ được ghi nhận khi đáy vốn hình thành sau đỉnh vốn. MDD có thể được sử dụng như làm thước đo cho các chỉ số, thước đo hiệu suất của các quỹ đầu tư. MDD được định nghĩa theo công thức (8):

$$DD_t[\%] = \frac{P_{max} - P_t}{P_{max}} 100\%$$
$$MDD_t = \underbrace{max}_{0 \le q \le t} (\mathrm{DD}_q) \quad (8)$$

Ở đây, P_t là giá của mục đầu tư tại thời điểm t. P_{max} là giá cao nhất của mục đầu tư. Độ sụt giảm tối đa là giá trị lớn nhất của các mức sụt giảm tới thời điểm t.

MDD là một chỉ báo được sử dụng để đánh giá mức độ rủi ro tương đối của một chiến lược sàng lọc cổ phiếu hoặc tiền ảo so với các chiến lược khác. Nó tập trung vào việc bảo toàn vốn. Đây là mối quan tâm chính của hầu hết các nhà đầu tư. MDD càng thấp sẽ chứng minh rằng khoản lỗ từ đầu tư sẽ càng nhỏ.

3.2. Lợi nhuận hằng năm (Annualized return)

Annualized return (Lợi nhuận hằng năm) là số tiền trung bình mà một khoản đầu tư kiếm được mỗi năm trong một khoảng thời gian nhất định. Lợi nhuận hằng năm cung cấp một tổng quan về khả năng đầu tư, chỉ ra lợi nhuận mà nhà đầu tư kiếm được trong khoảng thời gian một năm. Tổng lợi nhuận hằng năm cung cấp hiệu suất của các khoản đầu tư và không cung cấp cho nhà đầu tư bất kỳ dấu hiệu nào về sự biến động giá của thị trường.

$$Lợi\ nhuận\ hằng\ năm = (1 + Tỷ\ suất\ \mathrm{sinh}\ lời)^{\frac{365}{Days\ Held}} \quad (9)$$

Ở đây, biến Days Held là tổng số ngày mà quỹ đang nắm giữ. Biến tỷ suất sinh lời (cumulative return) là số tiền mà việc đầu tư thu được hoặc mất mát được tính theo phần trăm theo công thức

$$Tỷ\ suất\ \mathrm{sinh}\ lời = \frac{Giá\ hiện\ tại - Giá\ ban\ đầu}{Giá\ ban\ đầu} * 100\% \quad (10)$$

3.3. Độ chính xác

Độ chính xác của một mô hình máy học là thước đo được dùng để xác định mô hình nào là tốt nhất. Mô hình nào đặc trưng được dữ liệu không nhìn thấy tốt hơn là mô hình có độ chính xác cao hơn. Giới hạn trên của độ chính xác là 1.

$$\text{Độ } chính\ xác = \frac{Lợi\ nhuận\ lớn\ nhất}{Lợi\ nhuận\ lý\ tưởng} * 100\% \quad (11)$$

Công thức (11) được dùng tính độ chính xác của mô hình. Trong đó, biến lợi nhuận lớn nhất đề cập tới quá trình thực hiện việc mua bán sau khi trừ chi phí và giá mua vào, phần lợi nhuận còn lại là lớn nhất. Lợi nhuận lý tưởng là lợi nhuận được tính dựa vào dữ liệu thực tế so với dữ liệu trước đó. Nếu dữ liệu dự báo lớn hơn dữ liệu thực tế trước đó, hành động mua được thực hiện và lợi nhuận được tính từ hiệu của giá trị đóng cửa thực và giá trị đã mua. Trong mô hình mô phỏng, hành động mua được thực hiện trong số vốn đầu tư ban đầu. Ở đây, khi hành động bán được thực hiện thành công, hành động mua kế tiếp mới thực hiện. Tương tự, hành động bán chỉ thực hiện khi đã có hành động mua toàn số vốn đã thực thi thành công. Vì vậy, nhà đầu tư sẽ không mua hai lần liên tiếp hoặc bán hai lần liên tiếp cho dù giá mua đang thấp hoặc giá bán đang cao.

4. THIẾT KẾ HỆ THỐNG HỌC TĂNG CƯỜNG KẾT HỢP VỚI XU HƯỚNG DỮ LIỆU DÙNG GIẢI THUẬT Q-LEARNING

Một mô hình học tăng cường bắt đầu từ trạng thái S(t). Ở trạng thái này, tác nhân dựa trên chính sách đã được thiết kế sẵn để đưa ra một hành động A(t) trong một môi trường. Môi trường sau khi quan sát hành động sẽ chuyển sang trạng thái tiếp theo S(t+1) đối với tác nhân và đồng thời dựa theo hành động mà tác nhân đã thực hiện, môi trường sẽ đưa ra phần thưởng R(t) tương ứng.

Tác nhân sẽ lặp đi lặp lại quy trình cho đến khi tìm được chuỗi những hành động tối đa hóa được những phần thưởng được nhận. Để đảm bảo tác nhân thực hiện đúng những hành động tối ưu mà chúng ta mong muốn, chúng ta sẽ tạo ra một bộ nhớ cho các tác nhân này, sử dụng hàm V, là hàm giá trị trạng thái của phương trình Bellman. Từ đó, ta sẽ có giá trị tại các trạng thái như đề cập trong mục 2.

Một mô hình RL gồm có các thành phần chính: tác nhân (Agent), môi trường (Environment), trạng

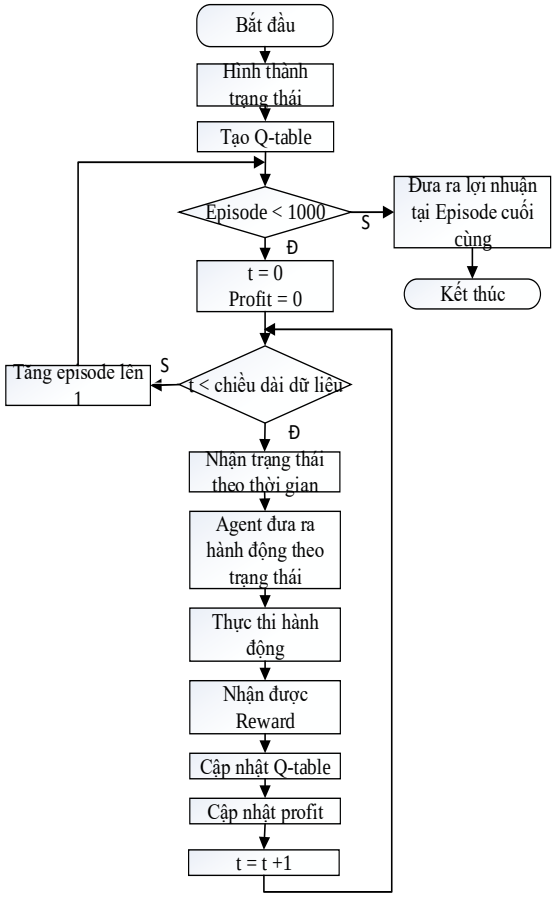
thái (State), hành động (Action), hàm phần thưởng (Reward), số lần lặp (giá trị Episode) và hàm chính sách (Policy). Ở đó, **tác nhân** được định nghĩa là máy dùng để quan sát môi trường và đưa ra các quyết định để thực thi các hành động tương ứng.

Môi trường là không gian tồn tại xung quanh tác nhân, nơi để tác nhân tồn tại và tương tác. Trong quyết định mua bán tài chính, môi trường tài chính là toàn bộ các yếu tố môi trường phức tạp của thị trường tài chính xung quanh có ảnh hưởng tới hoạt động tác nhân. Vấn đề chính của thuật toán mua bán là khả năng quan sát môi trường bị giới hạn. Một số thông tin của môi trường bị ẩn hoặc không nhận biết được của các tác nhân, ví dụ như độ tin cậy của các dữ liệu báo cáo tình hình kinh doanh của công ty hoặc là các chiến lược của người tham gia vào thị trường mua bán. Vì vậy, thông tin cung cấp tới tác nhân bị giới hạn so với mức độ phức tạp của thị trường. Quan trọng hơn, thông tin thu thập được từ môi trường được xem xét như là chuỗi thời gian thay vì các giá trị riêng rẽ độc lập. Ở đó, mỗi bước thời gian **t**, tác nhân quan sát thị trường tài chính có một trạng thái là $\mathbf{s_t}$. Thông tin thu thập trong môi trường mua bán này được ký hiệu $\mathbf{O_t}$. Một cách lý tưởng, môi trường quan sát **O** nên bao gồm tất cả các khả năng thông tin có ảnh hưởng tới giá của thị trường tài chính. Bởi vì thuật toán mua bán xem xét tính tuần tự của dữ liệu, môi trường phải được xem xét như chuỗi thời gian **t**. Trong nghiên cứu này, giả sử rằng thông tin quan sát chỉ được xem xét bao gồm dữ liệu giá hàng ngày của thị trường tài chính, giá đóng, giá mở cửa, giá cao nhất và số lượng *coin* giao dịch trong một ngày. Trạng thái của môi trường tài chính là các giá trị giá đóng của đồng coin được hình thành theo từng ngày. Tác nhân nhận các trạng thái này như ngõ vào để xử lý. Hành động là phương thức của tác nhân cho phép nó tương tác với môi trường. Dựa trên trạng thái S(t) hiện tại của môi trường mà tác nhân sẽ đưa ra hành động A(t). Hành động bao gồm mua hoặc bán hoặc giữ. Hành động mua nhằm gia tăng số lượng đồng coin được thực hiện bởi tác nhân. Giá trị đầu tư ban đầu là tổng số tiền đầu tư để thực hiện việc mua hoặc bán trong thị trường tài chính. Gọi N_S là số coin hoặc cổ phiếu. N_S được tính dựa vào giá mở cửa ban đầu như sau:

$$N_S = \frac{Tổng\ vốn\ đầu\ tư}{Giá\ mở\ cửa\ ban\ đầu} \quad (12)$$

Ở mỗi hành động, môi trường gửi đến cho tác nhân (agent) một phần thưởng xác định. Mục tiêu của tác nhân là tối đa hóa tổng phần thưởng mà nó nhận được trong một thời gian dài. Tín hiệu phần thưởng giúp xác định đâu là sự kiện tốt và xấu đối với tác nhân, đồng thời, nó cũng là cơ sở chính để thay đổi chính sách. Nếu một hành động được lựa chọn bởi chính sách mang đến phần thưởng thấp, thì chính sách đó có thể bị thay đổi. Tác nhân sẽ lựa chọn các hành động khác trong các tình huống tương tự ở tương lai. Mục tiêu của hàm phần thưởng đưa ra chiến lược lợi nhuận. Một loạt các tương tác giữa tác nhân và môi trường từ thời điểm bắt đầu đến khi quyết định hành động được gọi là một Episode.

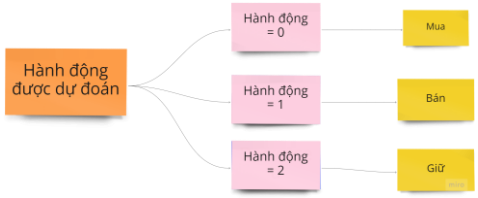
Chính sách là yếu tố xác định cách thức hoạt động của tác nhân tại một thời điểm nhất định. Nói cách khác, chính sách là một ánh xạ từ các trạng thái của môi trường đến các hành động sẽ được thực hiện khi ở trong các trạng thái đó. Chính sách là cốt lõi của tác nhân trong việc xác định hành vi. Trong một số trường hợp, chính sách có thể là một hàm hoặc bảng tra cứu đơn giản. Trong một số trường hợp khác, chính sách có thể liên quan đến tính toán mở rộng, ví dụ như quá trình tìm kiếm.



Hình 2. Lưu đồ giải thuật tổng quát của mô hình

Hình 2 trình bày lưu đồ giải thuật tổng quát của mô hình. Bước đầu khi đọc file csv, mô hình hình

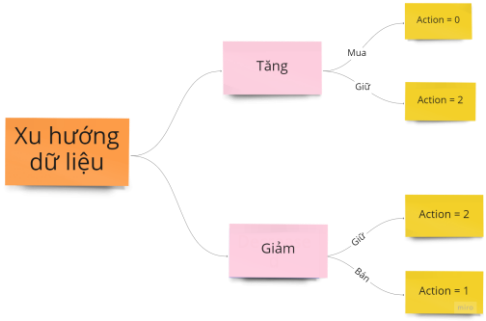
thành và xác định số trạng thái (state) mà chương trình cần theo chiều dài dữ liệu. Vì môi trường khảo sát là tài chính nên trong nghiên cứu này, môi trường sẽ có 3 hành động (action) là mua, bán và giữ. Từ số hành động và số trạng thái, Q-table được hình thành với số hàng là số trạng thái và số cột là số hành động. Tiếp đến số vòng thực hiện chương trình (episode) được xác định là 1.000. Sau đó khởi tạo biến t là 0 đại diện cho ngày đầu tiên và lợi nhuận (profit) ban đầu sẽ là 0. Sau đó chương trình sẽ nhận trạng thái (State) theo t và gửi trạng thái này về cho tác nhân để tác nhân ra quyết định chọn hành động nào. Sau đó chương trình sẽ thực hiện hành động mà tác nhân gửi về. Qua đó, phần thưởng (reward) được xác định sau khi thực hiện hành động. Phần thưởng nếu hành động thực hiện đúng chiều xu hướng sẽ được cho là luôn dương. Ngược lại, nếu không đúng theo xu hướng thì hàm phần thưởng sẽ âm. Nếu hành động sinh lợi nhuận, thì hàm phần thưởng sẽ có giá trị dương và khi sinh lỗ sẽ có giá trị âm. Sau đó, phần thưởng này được áp dụng vào công thức Bellman trong công thức (6) để cập nhật giá trị trong Q-table theo vị trí là trạng thái (state) và hành động (action) hiện tại. Cuối cùng là cập nhật lợi nhuận (profit) sau đó là tăng t lên một đơn vị và thực hiện như các bước trên. Khi khảo sát hết chiều dài dữ liệu (điều kiện t < chiều dài dữ liệu không thỏa) thuật toán sẽ ra được lợi nhuận tích lũy cho một episode đồng thời tăng episode lên một đơn vị. Quá trình lặp đi lặp lại cho đến khi kết thúc episode 1.000 chương trình sẽ đưa ra lợi nhuận tích lũy thu được. Lợi nhuận tích lũy được tính toán trên từng episode và lấy kết quả lớn nhất trong 1.000 episode này. Đây cũng là lợi nhuận được tối ưu hóa sau khi thực hiện hết chương trình theo kỹ thuật chính là Q-leaning.



Hình 3. Sơ đồ khối dự báo hành động mua/bán/giữ bởi mô hình tối ưu ngẫu nhiên theo Q-learning với điều kiện không xem xét xu hướng dữ liệu

Hình 3 trình bày sơ đồ khối dự báo các hành động mua/bán/giữ bởi mô hình tối ưu ngẫu nhiên bởi thuật toán Q-learning với không xem xét đường xu hướng dữ liệu. Các hành động tối ưu này được thực hiện bởi thuật toán Q-learning. Các hành động mua bán được thực hiện theo hàm phần thưởng để đạt giá trị lớn nhất mà không xem xét các đường xu hướng hoặc các chỉ số kỹ thuật để xác định xu hướng. Mỗi hành động được thực hiện bởi bảng Q-table được tối ưu trả về. Nếu hành động = 0, mô hình sẽ thực thi hành động mua và ngược lại nếu hành động = 1, mô hình sẽ thực hiện chỉ thị bán. Nếu hành động hành động = 2, mô hình sẽ giữ và không thực hiện việc mua và bán.

Hình 4 trình bày sơ đồ thực hiện mua/bán/giữ dựa trên đường xu hướng kết hợp thuật toán Q-learning. Không như mô hình ngẫu nhiên ở trên Hình 3, việc mua/bán được thực hiện dựa trên xu hướng dữ liệu kết hợp thuật toán Q-learning. Nếu đường xu hướng tăng và hành động action=0, hành động mua sẽ được thực hiện. Nếu đường xu hướng giảm và hành động action=1 thì hành động bán được thực hiện. Nếu đường xu hướng tăng mà hành động action=2 được trả về từ Q-learning thì hành động giữ được thực hiện. Nếu đường xu hướng giảm mà hành động action=0 thì hành động giữ cũng được thực hiện. Khi action=2 được trả về từ Q-learning, thì cả đường xu hướng tăng hay giảm, khi đó hành động giữ luôn được thực hiện.



Hình 4. Sơ đồ thực hiện mua/bán/giữ dựa trên xu hướng dữ liệu kết hợp thuật toán Q-learning

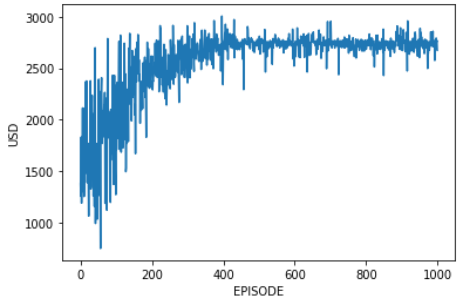
5. KẾT QUẢ ĐÁNH GIÁ

Kết quả đánh giá của mô hình được thực hiện dựa trên hai tập dữ liệu bao gồm Dogecoin và Bitcoin từ Yahoo! Finance. Tập Dogecoin được lấy từ 7/2017 tới tháng 3/2021 cho quá trình huấn luyện và từ 4/2021 đến 4/2022 cho quá trình kiểm tra. Một cách tương tự, tập dữ liệu BITCOIN được lấy từ Yahoo! Finance từ ngày 01/2015 đến 8/2021 cho quá trình huấn luyện và từ 9/2021 tới 9/2022 cho quá trình kiểm tra.

5.1. Kết quả đánh giá hành động mua bán dựa trên mô hình Q-learning không dựa theo xu hướng

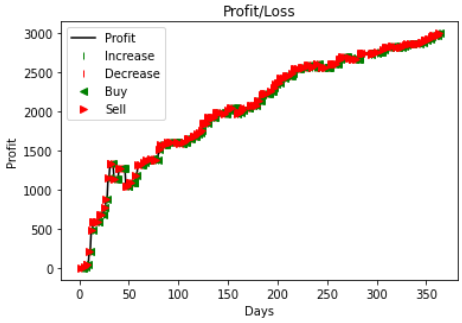
Hình 5 trình bày lợi nhuận thu được khi thực thi mô hình Q-learning không có sự kết hợp với đường xu hướng trên tập dữ liệu Dogecoin. Mô hình Q-

learning thông thường thực hiện việc mua bán không dựa trên đường xu hướng. Việc thực hiện hành động mua bán dựa vào việc tối ưu hành động được thực hiện từ môi trường trả về.



Hình 5. Sự hội tụ của lợi nhuận khi thực hiện 1.000 episode bởi thuật toán Q-learning không dựa theo đường xu hướng

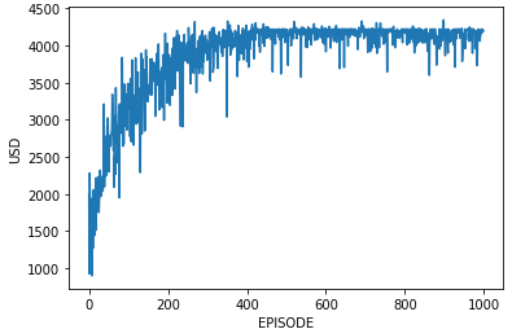
Hình 5 phân tích sự hội tụ của thuật toán Q-learning khi không xem xét kết hợp với đường xu hướng. Thuật toán Q-learning hội tụ khi thực hiện ở khoảng episode thứ 400. Các episode từ 401 tới 1.000 thể hiện sự hội tụ của lợi nhuận. Trục y thể hiện số đơn vị lợi nhuận lớn nhất trên mỗi episode. Biên độ lợi nhuận thay đổi theo từng episode. Các episode đầu, thuật toán đang trong quá trình khám phá ra bảng Q-table để tìm kiếm hành động có hàm phần thưởng tối ưu nhất. Khi quá trình khám phá kết thúc, lợi nhuận sẽ được hội tụ và hầu như không thay đổi. Các hành động khác nhau sẽ đưa ra hàm phần thưởng khác nhau. Giá trị lợi nhuận này đã được chuẩn hóa trên giá mở bán của đồng coin ban đầu. Lợi nhuận đạt được sẽ bão hòa khi tăng số episode.



Hình 6. Lợi nhuận thu được khi thực thi mô hình Q-learning không có sự kết hợp đường xu hướng. Mô hình Q-learning thực thi hành động mua/bán không dựa trên đường xu hướng

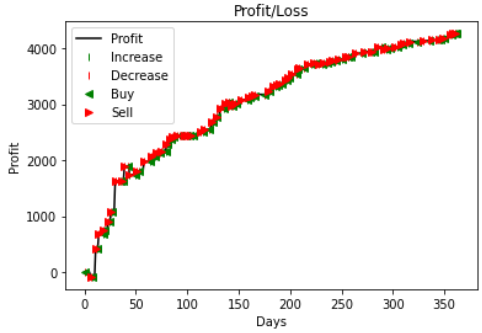
Hình 6 thể hiện lợi nhuận thu được sau 365 ngày khi thực hiện mua, bán hoặc giữ. Lợi nhuận được tính bằng giá bán trừ đi giá mua. Việc thực hiện mua (Buy) và bán (Sell) này không theo xu hướng tăng hoặc giảm của giá thị trường. Quá trình mua bán được thực thi dưới kết quả được trả về để hàm phần thưởng đạt được giá trị lớn nhất trong quá trình thực thi mô hình học tăng cường. Hành động mua bán được thực hiện liên tục và biểu diễn bằng dấu mũi tên theo từng ngày trong vòng 365 ngày trong tập dữ liệu kiểm tra. Đường lợi nhuận là đường ký hiệu liên tục, màu đen. Đường lợi nhuận có thể giảm hoặc tăng tùy theo hành động đưa ra mang lại lợi nhuận hay thua lỗ.

5.2. Thực hiện mua bán dựa trên đường xu hướng dữ liệu +



Hình 7. Sự hội tụ của lợi nhuận khi thực hiện 1.000 episode bởi thuật toán Q-learning theo đường xu hướng

Hình 7 trình bày sự hội tụ của lợi nhuận khi thực hiện 1.000 episode bởi thuật toán Q-learning kết hợp xu hướng giá của thị trường. Kỹ thuật tìm ra xu hướng của dữ liệu giúp cho tác nhân ra các quyết định tối ưu hơn về lợi nhuận so với kỹ thuật mua bán không theo xu hướng như được mô tả trong Hình 5. Giá trị lợi nhuận của mô hình đề xuất khoảng 4.200 đơn vị cao hơn mô hình mua bán không theo xu hướng. Lợi nhuận đạt được của mô hình mua bán không theo xu hướng khoảng 3.000 đơn vị.



Hình 8. Lợi nhuận thu được khi thực thi mô hình Q-learning dựa theo đường xu hướng. Mô hình Q-learning thực thi hành động mua/bán dựa trên đường xu hướng

Hình 8 phân tích các hành động mua và bán trong 365 ngày. Ở đây, hành động giữ không mang lại lợi nhuận nên không được hiển thị trên đồ thị lợi nhuận. Các dấu hiệu giá tăng hoặc giảm được đưa vào mô hình học tăng cường để cung cấp ngõ vào cho các tác nhân, giúp hệ thống ra các quyết định mua hoặc bán hoặc giữ hiệu quả hơn tới lợi nhuận.

Bảng 2 trình bày kết quả của lợi nhận sau khi thực hiện 1.000 episode và 4.000 episode. Đầu tư ban đầu với 1.714 đơn vị coin, được định nghĩa là 100% vốn ban đầu. Sau 365 ngày, lợi nhuận tích lũy tăng thêm 249,465% với hệ thống mua bán theo xu hướng so với 172,62% của hệ thống mua bán không dựa theo xu hướng, hệ thống chỉ thực hiện hành động tối ưu theo mô hình học tăng cường.

Nhằm phân tích ảnh hưởng của số episode tới việc huấn luyện một agent, nghiên cứu thực hiện so sánh kết quả giữa việc học được thực hiện 4.000 và 1.000 episode. Dựa trên kết quả Bảng 2, nhận thấy rằng khi tăng số lượng episode các agent, hiệu quả học của agent qua môi trường tốt hơn. Cụ thể, khi thực hiện với 4.000 episode, mô hình học tăng cường không theo xu hướng lợi nhuận trở nên tốt hơn do quá trình học tăng cường đã tạo ra các hành động tối ưu hơn so với thực hiện vòng lặp 1.000 episode. Điều này là bởi vì càng nhiều episode cung cấp cho agent nhiều cơ hội để khám phá ra môi trường và học từ những kinh nghiệm trước của nó.

Khi đó, lợi nhuận mua bán không theo xu hướng tăng thêm 190,7%. Lợi nhuận của mô hình mua bán không theo xu hướng thấp nhất chỉ đạt 172,622% so với 249,465% khi mua bán theo xu hướng. Mô hình không theo xu hướng lợi nhuận thấp hơn mô hình theo xu hướng. Trong hai lần thực nghiệm với 1.000 và 4.000 episode, lợi nhuận của mô hình mua bán không theo xu hướng chỉ đạt 172,6222% và 190,721% tương ứng. Trong khi đó, mô hình thực hiện theo xu hướng lợi nhuận đạt 249,465% và 195,47% tương ứng.

Tương tự, Bảng 3 chỉ ra lợi nhuận thu được khi kiểm tra với tập Bitcoin. Lợi nhuận thu được khi hệ thống thực hiện học tăng cường có xem xét yếu tố xu hướng thị trường sau 1.000 episode là 182,6% so với học tăng cường không xem xét xu hướng thị trường với chỉ 131,98%. Khi thực hiện 4.000 episode, mô hình học tăng cường không theo xu hướng tăng lợi nhuận hơn nguyên nhân chủ yếu mô hình đã hội tụ và đưa ra các hành động tốt hơn so với chỉ thực hiện 1.000 episode.

Bảng 2 và Bảng 3 cũng đánh giá thêm hiệu năng của hai mô hình trên hai tập dữ liệu Dogecoin và Bitcoin khác nhau. Nghiên cứu đánh giá hiệu quả của mô hình qua các thước đo hiệu năng bao gồm độ chính xác, mức sụt giảm tối đa và lợi nhuận tích lũy. Kết quả chỉ ra rằng độ chính xác của mô hình mua bán dựa theo xu hướng có kết quả vượt trội so với mô hình khác trong các thước đo hiệu năng trong các khía cạnh về độ chính xác, độ sụt giảm và lợi nhuận tích lũy được thực hiện trên cả hai tập dữ liệu khác nhau với số lượng vòng lặp khác nhau.

Bảng 2. Kết quả đánh giá độ chính xác và các thước đo hiệu năng giữa các mô hình với tập dữ liệu Dogdecoin

	Episode = 1000			Episode = 4000		
	Độ chính xác	**Mức sụt giảm tối đa (MDD)**	**Lợi nhuận tích lũy**	**Độ chính xác**	**Mức sụt giảm tối đa (MDD)**	**Lợi nhuận tích lũy**
Theo xu hướng	88,76%	79,31%	249,46%	69,55%	91,72%	195,48%
Không theo xu hướng	61,42%	75,1%	172,62%	67,86%	96,59%	190,72%

Bảng 3. Kết quả đánh giá độ chính xác và các thước đo hiệu năng giữa các mô hình với tập dữ liệu Bitcoin

	Episode = 1000			Episode = 4000		
	Độ chính xác	**Mức sụt giảm tối đa (MDD)**	**Lợi nhuận tích lũy**	**Độ chính xác**	**Mức sụt giảm tối đa (MDD)**	**Lợi nhuận tích lũy**
Mua bán theo xu hướng	55,29%	99,33%	175,92%	57,81%	99,92%	183,97%
Mua bán không theo xu hướng	41,48%	98,97%	131,99%	48,82%	99,99%	155,36%

Phân tích một cách cụ thể, độ chính xác càng cao nghĩa là mô hình có lợi nhuận gần với lợi nhuận lý tưởng nhất. Độ chính xác thực hiện các hành động mua và bán của mô hình theo xu hướng là 88,76% so với 61,42% của mô hình không theo xu hướng với tập Dogecoin tại 1.000 vòng lặp như được mô tả trong Bảng 2.

Lợi nhuận lý tưởng được tính toán dựa trên giá trị bán thực tế trừ cho chi phí đã mua trước đó. Độ sụt giảm càng thấp thể hiện sự ổn định việc mua bán

càng cao. Các kết quả độ sụt giảm của hai mô hình trên hai tập dữ liệu Bitcoin và Dodgecoin thể hiện mô hình mua bán theo xu hướng có độ sụt giảm lớn nhất gần như bằng mức độ sụt giảm tối đa của mô hình mua bán không theo xu hướng. Vì vậy, hai mô hình thể hiện sự ổn định của lợi nhuận gần như nhau.

6. KẾT LUẬN

Sau khi phân tích, đánh giá và so sánh giữa các mô hình, những kết quả chỉ ra sự hiệu quả của học tăng cường kết hợp với xu hướng dữ liệu sẽ có nhiều ưu điểm trong việc dự báo giá thị trường tiền điện tử theo chuỗi dữ liệu thời gian. Mô hình đưa ra sự lựa chọn đúng cho người dùng để đạt được lợi nhuận không những tốt mà còn giảm nguy cơ thua lỗ bởi chỉ số MDD ổn định, đặc biệt quan trọng khi thị trường tiền điện tử có sự biến động lớn. Những thước đo đánh giá hiệu năng các mô hình trên các thước đo độ chính xác, lợi nhuận tích lũy, độ sụt giảm lớn nhất được sử dụng để so sánh đánh giá. Kết quả chỉ ra rằng, mô hình mua bán dựa trên học tăng cường với sự kết hợp của xu hướng dữ liệu mang lại nhiều ưu điểm hơn so với mô hình học tăng cường không xem xét xu hướng dữ liệu. Những kết quả được thực hiện trên cả hai tập dữ liệu Dogecoin và Bitcoin.

LỜI CẢM TẠ

Nghiên cứu này thuộc đề tài T2022-12 được hỗ trợ kinh phí bởi Trường Đại học Sư phạm Kỹ thuật Thành phố Hồ Chí Minh.

TÀI LIỆU THAM KHẢO

Braham, R., Samad, M.E., Bakhach, A.M., El-Chaarani, H., Sardouk, A., Nemar, S.E. & Jaber, D. (2022). Forecasting a Stock Trend Using Genetic Algorithm and Random Forest. *J. Risk Financial Manag. 15*, 188. https://doi.org/10.3390/jrfm15050188.

Cao, L. (2021). AI in Finance: Challenges, Techniques, and Opportunities. *ACM Computing Surveys (CSUR)* 55: 1 - 38. https://doi.org/10.48550/arXiv.2107.09051

Cao, L. (July 10, 2020). AI in Finance: A Review Available at *SSRN* https://ssrn.com/abstract=3647625 or http://dx.doi.org/10.2139/ssrn.3647625

Carta, S., Ferreira, A., Podda, A.S., Recupero, D.R., & Sanna, A. (2020). Multi-DQN: an Ensemble of Deep Q-Learning Agents for Stock Market Forecasting, *Expert Systems with Applications.* doi: https://doi.org/10.1016/j.eswa.2020.113820

Chopra, R., & Sharma, G.D. (2021). Application of Artificial Intelligence in Stock Market Forecasting: A Critique, Review, and Research Agenda. *J. Risk Financial Manag.*, *14*, 526. https://doi.org/10.3390/jrfm14110526

Culkin, R. (2017). Machine Learning in Finance: The Case of Deep Learning for Option Pricing.

Fischer, T. G. (2018). Reinforcement learning in financial markets - a survey. *Economics*.

Jagdish, C., & Manish, K.. (2019). Trend following deep Q-Learning strategy for stock trading. *Expert Systems*. https://doi.org/10.1111/exsy.12514

Kabbani, T., & Ekrem D. (2022). Deep Reinforcement Learning Approach for Trading Automation in the Stock Market. *IEEE Access.* 10: 93564-93574. https://doi.org/10.48550/arXiv.2208.07165

Li, Y. (2017). Deep Reinforcement Learning: An Overview. *ArXiv* abs/1701.07274.

Mhlanga, D. (2020). Industry 4.0 in Finance: The Impact of Artificial Intelligence (AI) on Digital Financial Inclusion. *Int. J. Financial Stud.*, *8*, 45. https://doi.org/10.3390/ijfs8030045

Jay, P., Kalariya, V., Parmar, P., Tanwar, S., Kumar, N. & Alazab, M. (2020). Stochastic Neural Networks for Cryptocurrency Price Prediction. *IEEE Access. vol. 8*, pp. 82804-82818, doi: 10.1109/ACCESS.2020.2990659.

OECD (2021). Artificial Intelligence, Machine Learning and Big Data in Finance: Opportunities, Challenges, and Implications for Policy Makers, https://www.oecd.org/finance/artificial-intelligence-machine-learningbig-data-in-finance.htm.

Olorunnimbe, K., & Viktor, H. (2022). Deep learning in the stock market—a systematic survey of practice, backtesting, and applications. *Artif Intell Rev*. https://doi.org/10.1007/s10462-022-10226-0.

Shah, D., Isah, H. & Zulkernine, F. (2019). Stock Market Analysis: A Review and Taxonomy of Prediction Techniques. *Int. J. Financial Stud. 7*, 26. https://doi.org/10.3390/ijfs7020026

Shahi, T.B., Shrestha, A., Neupane, A. & Guo, W. (2020). Stock Price Forecasting with Deep Learning: A Comparative Study. *Mathematics*, 8, 1441. https://doi.org/10.3390/math8091441.

Singh, J. & Khushi, M. (2021). Feature Learning for Stock Price Prediction Shows a Significant Role of Analyst Rating. *Appl. Syst. Innov.*, *4*, 17. https://doi.org/10.3390/asi4010017.

Al-Sulaiman, T. (2022). Predicting reactions to anomalies in stock movements using a feed-forward deep learning network. *International Journal of Information Management Data Insights, Volume 2*, Issue 1, https://doi.org/10.1016/j.jjimei.2022.100071.

TOAI, T.K., HANH, V. T. X., HUAN, & V. M. (2022). Applying ridge regression and ANN to predict ICO price after six months. *Journal of Science*.

TOAI, T. K.; SENKERIK, R.; ZELINKA, I.; ULRICH, A.; HANH, V.T. X.; & HUAN, V. M. (2022). ARIMA for Short-Term and LSTM for Long-Term in Daily Bitcoin Price Prediction. ICAISC2022 https://doi.org/10.1007/978-3-031-23492-7_12

Tsung-Jung, H., Hsiao-Fen, H., & Wei-Chang, Y. (2011). Forecasting stock markets using wavelet transforms and recurrent neural networks: An integrated system based on artificial bee colony algorithm. *Applied Soft Computing. Volume 11*, Issue 2, Pages 2510-2525. https://doi.org/10.1016/j.asoc.2010.09.007.

Deng, Y., Bao, F., Kong, Y., Ren, Z. & Dai, Q. (2017). Deep Direct Reinforcement Learning for Financial Signal Representation and Trading. *IEEE Transactions on Neural Networks and Learning Systems, vol. 28*, no. 3, pp. 653-664. doi: 10.1109/TNNLS.2016.2522401.

Yang, H., Liu, X. Y., Zhong, S. & Walid, A. (2020). Deep Reinforcement Learning for Automated Stock Trading: An Ensemble Strategy. *Available at SSRN*: https://ssrn.com/abstract=3690996 or http://dx.doi.org/10.2139/ssrn.3690996

Yuan, Q., & Jing, X. (2018). Fintech: AI powers financial services to improve people's lives. Commun. *ACM 61*, 11, 65–69. https://doi.org/10.1145/3239550.

Yuxuan, H., Luiz Fernando, C. & Danny, Ho. (2021). Machine Learning for Stock Prediction Based on Fundamental Analysis. 2021 IEEE Symposium Series on Computational Intelligence.