SỬ DỤNG TRÍ TUỆ NHÂN TẠO GIẢI BÀI TOÁN THỜI ĐIỂM DỪNG TỐI ƯU TRONG ĐẦU TƯ TÀI CHÍNH

Phạm Văn Khánh[*], Nguyễn Thành Trung[**]

Nhận bài: 18/06/2021; Nhận kết quả bình duyệt: 15/07/2021; Chấp nhận đăng: 30/07/2021

Tóm tắt

Trong bài báo này, chúng tôi trình bày một công cụ cao cấp của Trí tuệ nhân tạo, học tăng cường để thử nghiệm trong đầu tư cổ phiếu. Trí tuệ nhân tạo về cơ bản gồm có học máy, học sâu và học tăng cường. Học tăng cường sử dụng các lý thuyết toán học như quy hoạch động, quá trình quyết định Markov để cải tiến hành động trở nên tối ưu hơn. Học tăng cường có rất nhiều thuật toán khác nhau. Trong bài báo này, chúng tôi sử dụng thuật toán Zap Q-Learning để áp dụng trong việc đầu tư 30 mã cổ phiếu của thị trường chứng khoán Việt Nam. Chúng tôi thu được kết quả khá khiêm tốn: sau khi chiết khấu phần lãi suất ngân hàng, thì lợi nhuận còn khoảng 3%.

Từ khóa: *Trí tuệ nhân tạo; Học tăng cường; Thời điểm dừng tối ưu; Đầu tư tài chính; Xích Markov*

1. Giới thiệu

Trí tuệ nhân tạo hay trí thông minh nhân tạo (Artificial Intelligence) là một nhánh của khoa học liên quan đến việc làm cho máy tính có những khả năng của trí tuệ con người, tiêu biểu như các khả năng "suy nghĩ", biết "học tập", biết "lập luận" để giải quyết vấn đề, biết "học" và "tự thích nghi",... được ra đời tại hội nghị ở Dartmouth College mùa hè năm 1956, do Minsky và McCarthy tổ chức. Trí tuệ nhân tạo về cơ bản được hiểu là trí tuệ do con người lập trình tạo nên với mục tiêu giúp máy tính có thể tự động hóa các hành vi thông minh của con người.

Trí tuệ nhân tạo đang dần đi vào mọi lĩnh vực của mỗi quốc gia, của cuộc sống mỗi con người và đã cho thấy những ưu điểm nổi trội khi có thể xử lí dữ liệu nhanh hơn, khoa học hơn, thông minh hơn, hệ thống hơn với quy mô rộng hơn so với con người.

Trong toán học, lý thuyết thời điểm dừng tối ưu liên quan tới vấn đề chọn thời điểm để thực hiện một hành động cụ thể, nhằm tối đa hóa phần thưởng kì vọng hoặc giảm thiểu chi phí kỳ vọng. Đây là một trong những lý thuyết mang ý nghĩa rất quan trọng trong lĩnh vực xác suất, thống kê, kinh tế, đặc biệt là trong lĩnh vực toán tài chính.

[*] Viện Toán học và Khoa học ứng dụng (TIMAS), Trường Đại học Thăng Long
[**] Học viên cao học Phân tích dữ liệu QH-2018.T.CH, Khoa Toán – Cơ – Tin, Đại học Khoa học Tự nhiên

Thị trường chứng khoán vẫn luôn được coi là phong vũ biểu của nền kinh tế, là chỉ báo tương lai của sự chuyển động nền kinh tế. Có rất nhiều chủ thể tham gia thị trường chứng khoán như: các tổ chức phát hành, các nhà đầu tư cá nhân và nhà đầu tư tổ chức trong và ngoài nước, các nhà tạo lập thị trường,... bao gồm rất nhiều định chế tài chính quan trọng của nền kinh tế như: ngân hàng, công ty bảo hiểm, quỹ đầu tư, quỹ hưu trí, công ty chứng khoán,... và số lượng chứng khoán mà các chủ thể này hiện đang nắm giữ lên tới 6.679.640 tỷ đồng (UBCK Nhà Nước T12/2020). Câu hỏi mà tất cả các chủ thể trên thị trường chứng khoán đều quan tâm là các chiến lược nắm giữ tài sản như thế nào là hiệu quả. Các chủ thể trên thị trường luôn quan tâm tới những thay đổi bất lợi về giá trị của các trạng thái hoặc các danh mục tài sản của mình trong đó có tài sản là chứng khoán.

Những thành tựu của AI kết hợp với những lý thuyết toán học quan trọng đang là một hướng đi nhiều tiềm năng để có thể giúp cho các chủ thể trong nền kinh tế nói chung và của thị trường chứng khoán nói riêng có thể đưa ra các quyết định nắm giữ tài sản chính xác và kịp thời.

Bài báo nghiên cứu về mặt lý thuyết quy hoạch động, lý thuyết quá trình Markow, lý thuyết thời điểm dừng tối ưu, thuật toán Q-Learning, thuật toán Zap Q-Learning. Và từ đó, ứng dụng các lý thuyết và thuật toán này vào giải quyết bài toán thời điểm dừng tối ưu trong đầu tư tài chính với những bộ dữ liệu thực tế.

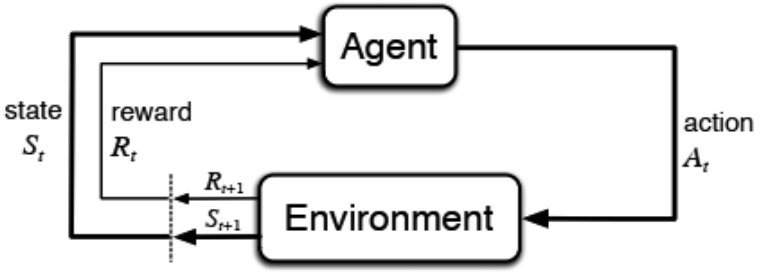
Các thuật toán Q-learning được biết là có các vấn đề hội tụ trong các cài đặt xấp xỉ hàm và điều này là do thực tế là toán tử quy hoạch động có thể không phải là một toán tử co. Nhiều thuật toán đã được đề xuất để cải thiện tốc độ hội tụ [3].

Học tăng cường hay còn được gọi là học củng cố (Reinforcement Learning) là lĩnh vực liên quan đến việc dạy cho máy (agent) thực hiện tốt một nhiệm vụ (task) bằng cách tương tác với môi trường (environment) thông qua hành động (action) và nhận được phần thưởng (reward). Học tăng cường đôi khi còn được gọi là học thưởng-phạt (reward-penalty learning), thuật toán học máy này có thể không yêu cầu dữ liệu huấn luyện, mà mô hình sẽ học cách ra quyết định bằng cách giao tiếp trực tiếp với môi trường xung quanh. Các thuật toán thuộc nhóm này liên tục ra quyết định và nhận phản hồi từ môi trường để củng cố hành vi.

Ví dụ như AlphaGo chơi cờ vây thắng con người trong bối cảnh cờ vây là một trò chơi có độ phức tạp cao với tổng số thế cờ xấp xỉ 10^{761} . Hay Google DeepMind không cần học dữ liệu từ các ván cờ của con người, hệ thống này tự chơi với chính mình để tìm ra các chiến thuật tối ưu và thắng tất cả con người và hệ thống khác bao gồm cả AlphaGo.

Một số thuật ngữ trong học tăng cường:

- *Environment* (môi trường): Là không gian mà máy tương tác.
- *Agent* (máy): Máy quan sát môi trường và sinh ra hành động tương ứng.
- *Policy* (chiến thuật): Máy sẽ theo chiến thuật như thế nào để đạt được mục đích.
- *Reward* (phần thưởng): Phần thưởng tương ứng từ môi trường mà máy nhận được khi thực hiện một hành động.
- *State* (trạng thái): Trạng thái của môi trường mà máy nhận được.



Hình 1. Sơ đồ học tăng cường

- *Episode*: Một chuỗi các trạng thái và hành động cho đến trạng thái kết thúc $s_1, a_1, s_2, a_2, \ldots s_T, a_T$
- *Accumulative Reward* (phần thưởng tích lũy): Tổng phần thưởng tích lũy từ state 1 đến state cuối cùng. Như vậy, tại state s_t, agent tương tác với environment với hành động a, dẫn đến state mới s_{t+1} và nhận được reward tương ứng r_{t+1}. Vòng lặp như thế cho đến trạng thái cuối cùng s_T.

Bài báo được chia làm 5 phần như sau: Phần 1 dành cho giới thiệu, phần 2 trình bày về quá trình Markov và quá trình quyết định Markov hữu hạn, phần 3 trình bày các thuật toán Q-Learning và Zap Q-Learning và phần 4 trình bày kết quả thực nghiệm và kết luận.

2. Quá trình Markow và quá trình quyết định Markow hữu hạn

2.1. Xích Markow (xem [5],[6])

Trong lý thuyết xác suất và các lĩnh vực liên quan, quá trình Markov (đặt theo tên của nhà toán học người Nga Andrey Markov) là một quá trình ngẫu nhiên thỏa mãn một tính chất đặc biệt, gọi là tính chất Markov (còn gọi là tính mất trí nhớ). Tính chất này giúp dự báo được tương lai chỉ dựa vào trạng thái hiện tại. Xích Markov là quá trình Markov đặc biệt mà trong đó hoặc có trạng thái rời rạc hoặc thời gian rời rạc. Quá trình Markov được nhà toán học Markov bắt đầu nghiên cứu từ khoảng đầu thế kỷ 20 và được ứng dụng nhiều trong các lĩnh vực công nghiệp, tin học, viễn thông, kinh tế, ...

Ta xét một hệ nào đó được quan sát tại các thời điểm rời rạc 0, 1, 2,... Giả sử các quan sát đó là $X_0, X_1, \ldots, X_n, \ldots$ Khi đó ta có một dãy các đại lượng ngẫu nhiên (ĐLNN) (X_n), trong đó X_n là trạng thái của hệ tại thời điểm n. Ký hiệu E là tập giá trị của các (X_n). Khi đó E là một tập hữu hạn hay đếm được, các phần tử của nó được ký hiệu là i, j, k... Ta gọi E là không gian trạng thái của dãy.

Định nghĩa 1 (Tính Markow)

Ta nói rằng dãy các ĐLNN (X_n) là một xích Markov nếu với mọi $n_1 < \ldots < n_k < n_{k+1}$ và với mọi $i_1, i_2, \ldots i_{k+1} \in E$

$$P\{X_{n_{k+1}} = i_{k+1} \mid X_{n_1} = i_1, X_{n_2} = i_2 \ldots, X_{n_k} = i_k\}$$
$$= P\{X_{n_{k+1}} = i_{k+1} \mid X_{n_k} = i_k\} \qquad (1.1)$$

Định nghĩa 2

Một xích Markow được gọi là thuần nhất nếu và chỉ nếu $P\{X_{m+n} = j \mid X_m = i\}$ là xác suất để xích tại thời điểm m ở trạng thái i sau n bước tại thời điểm m + n chuyển sang trạng thái j không phụ thuộc vào m.

2.2. Quá trình quyết định Markow hữu hạn

Quá trình quyết định Markow (MDP) được sử dụng để mô tả một môi trường học tăng cường. Một quá trình quyết định Markov là một tập 5-dữ liệu: (S, A, P. (. , .), R. (. , .), γ), trong đó:

- S là một tập hữu hạn các trạng thái
- A là một tập hữu hạn các hành động (ngoài ra A_s là tập hữu hạn các hành động có sẵn từ trạng thái s)
- $P_a(s,s') = Pr(s_{t+1} = s' \mid s_t = s, a_t = a)$ là xác suất mà hành động a ở trạng thái s tại thời điểm t chuyển sang trạng thái s' tại thời điểm t+1
- $R_a(s,s')$ là phần thưởng nhận được khi chuyển trạng thái từ s sang s'
- $\gamma \in [0,1]$ là hệ số chiết khấu đại diện cho sự khác biệt quan trọng giữa các phần thưởng tương lai và các phần thưởng hiện tại.

Trong MDP hữu hạn, tập hợp các trạng thái, hành động và phần thưởng (S, A và R) đều có một số hữu hạn các phần tử. Trong trường hợp này, các biến ngẫu nhiên R_t và S_t có phân bố xác suất rời rạc được xác định rõ ràng và chỉ phụ thuộc vào trạng thái và hành động của thời điểm trước đó. Nghĩa là, đối với các giá trị cụ thể của các biến ngẫu nhiên này ta có với mọi $s', s \in S$; $r \in R$; $a \in A_{(s)}$:

$$p(s',\, r \mid s,\, a) \;=\; Pr\{S_t = s',\, R_t = r \mid S_{t-1} = s,\, A_{t-1} = a\} \tag{1.3}$$

p là phân phối xác suất của mỗi lựa chọn s và a, vì vậy:

$$\sum_{s' \in S} \sum_{r \in R} p(s',\, r \mid s,\, a) \;=\; 1,\, \forall s \in S,\, a \in A_{(s)}$$

Xác suất chuyển trạng thái của môi trường:

$$p(s' \mid s,\, a) \;=\; Pr\{S_t = s' \mid S_{t-1} = s,\, A_{t-1} = a\} = \sum_{r \in R} p(s',\, r \mid s,\, a) \tag{1.4}$$

Phần thưởng kì vọng của cặp trạng thái – hành động là hàm hai đối số $r: S \times A \to \mathbb{R}$

$$r(s,\, a) \;= E\,[R_t \mid S_{t-1} = s,\, A_{t-1} = a] \;= \sum_{r \in R} r \sum_{s' \in S} p(s',\, r \mid s,\, a) \tag{1.5}$$

Và phần thưởng kì vọng cho trạng thái-hành động-tiếp theo là hàm với ba đối số $r: S \times A \times S \to \mathbb{R}$:

$$r(s,a,s') \;= E\,[R_t \mid S_{t-1} = s,\, A_{t-1} = a,\, S_t = s'] \;= \sum_{r \in R} r \frac{p(s',\, r \mid s,\, a)}{p(s' \mid s,\, a)} \tag{1.6}$$

2.2.1. Mục tiêu và phần thưởng

Mục tiêu của agent là tối đa hóa phần thưởng tích lũy mà nó nhận được trong thời gian dài. Nếu chuỗi phần thưởng nhận được sau bước thời gian t được ký hiệu là R_{t+1}, R_{t+2}, R_{t+3}, . . ., vậy thì khía cạnh chính xác nào của chuỗi này mà agent muốn tối đa hóa? Nói chung, agent luôn tìm cách tối đa hóa lợi nhuận kì vọng. Trong đó, lợi nhuận được ký hiệu là G_t, được định nghĩa là một số hàm cụ thể của chuỗi phần thưởng. Trong trường hợp đơn giản nhất, lợi nhuận là tổng phần thưởng:

$$G_t = R_{t+1} + R_{t+2} + R_{t+3} + ... + R_T$$

Trong đó, T là bước cuối cùng.

Khái niệm bổ sung mà học viên cần đề cập tới là chiết khấu. Theo cách tiếp cận này, agent cố gắng chọn các hành động để tối đa hóa tổng phần thưởng chiết khấu mà đại lý nhận được trong tương lai. Cụ thể, nó chọn A_t để tối đa hóa lợi nhuận chiết khấu kì vọng:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

Với γ là một tham số được gọi là tỷ số chiết khấu $0 \le \gamma \le 1$

Lợi nhuận ở các bước thời gian liên tiếp có liên quan với nhau theo cách quan trọng đối với lý thuyết và thuật toán của việc học củng cố:

$$\begin{aligned} G_t &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + ... \\ &= R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + ...) \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned}$$

2.2.2. Chính sách và hàm giá trị

Hàm giá trị của trạng thái s theo chính sách π, được ký hiệu là $v_\pi(s)$, là lợi tức kì vọng khi bắt đầu từ s và theo sau π sau đó. Đối với MDP, chúng ta có thể xác định $v_\pi(s)$ chính thức bằng cách:

$$v_\pi(s) = E_\pi[G_t \mid S_t = s] = E_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s], \forall s \in S \quad (1.7)$$

trong đó, $E_\pi[\cdot]$ biểu thị giá trị kỳ vọng của một biến ngẫu nhiên cho rằng tác nhân tuân theo chính sách π và t là bất kỳ bước thời gian nào. Lưu ý rằng giá trị của trạng thái đầu cuối, nếu có, luôn bằng 0. Chúng tôi gọi hàm $v_\pi(s)$ là hàm giá trị trạng thái cho chính sách π.

Tương tự xác định giá trị của việc thực hiện hành động a ở trạng thái s theo chính sách π, được ký hiệu là $q_\pi(s,a)$, là lợi nhuận kỳ vọng bắt đầu từ s, thực hiện hành động a, và sau đó theo chính sách π

$$q_\pi(s,a) = E_\pi[G_t \mid S_t = s, A_t = \text{a}] = E_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = \text{a}] \quad (1.8)$$

Gọi $q_\pi(s,a)$ là hàm giá trị hành động cho chính sách .

Đặc tính cơ bản của các hàm giá trị được sử dụng trong suốt quá trình học củng cố và lập trình động là chúng thỏa mãn các mối quan hệ đệ quy tương tự như các mối quan hệ mà chúng ta đã thiết lập cho kết quả trả về (1.8). Đối với bất kỳ chính sách π và bất kỳ trạng thái nào, điều kiện nhất quán sau đây giữ giữa giá trị của s và giá trị của trạng thái kế thừa có thể có của nó:

$$\begin{aligned} v_\pi(s) &= E_\pi[G_t \mid S_t = s] \\ &= E_\pi[\mathrm{R}_{t+1} + \gamma G_{t+1} \mid S_t = s] \\ &= \sum_a \pi(a|s) \sum_{s'} \sum_r p(\mathrm{s'},\mathrm{r}|\mathrm{s},\mathrm{a})\,[r + \gamma E_\pi[G_{t+1} \mid S_{t+1} = s']] \\ &= \sum_a \pi(a|s) \sum_{s',r} p(\mathrm{s'},\mathrm{r}|\mathrm{s},\mathrm{a})\,[r + \gamma v_\pi(s')] \end{aligned} \tag{1.10}$$

2.2.3. Các chính sách tối ưu và hàm giá trị tối ưu

Về cơ bản, việc giải quyết một nhiệm vụ học tập tăng cường có nghĩa là tìm ra một chính sách đạt được nhiều phần thưởng trong thời gian dài. Đối với MDP hữu hạn, chúng ta có thể xác định chính xác một chính sách tối ưu theo cách sau. Các hàm giá trị xác định thứ tự từng phần đối với các chính sách. Chính sách π được xác định là tốt hơn hoặc bằng chính sách π' nếu lợi tức kỳ vọng của nó lớn hơn hoặc bằng π' đối với tất cả các trạng thái. Nói cách khác, $\pi \geq \pi'$ nếu và chỉ khi $v_\pi(s) \geq v_{\pi'}(s)$ với mọi $s \in S$. Luôn có ít nhất một chính sách tốt hơn hoặc bằng tất cả các chính sách khác. Đây là một chính sách tối ưu. Mặc dù có thể có nhiều hơn một, chúng tôi biểu thị tất cả các chính sách tối ưu bằng π^*. Chúng chia sẻ cùng một hàm giá trị trạng thái, được gọi là hàm giá trị trạng thái tối ưu, được ký hiệu là v^* và được định nghĩa là: $v^*(s) = \max_\pi v_\pi(s), \forall \mathrm{s} \in \mathrm{S}$

Các chính sách tối ưu cũng chia sẻ cùng một hàm giá trị hành động tối ưu, được ký hiệu là q^* và được định nghĩa là: $q^*(s,a) = \max_\pi q_\pi(s,a), \forall \mathrm{s} \in \mathrm{S}, a \in A$

Đối với (các cặp trạng thái - hành động, a), hàm này cung cấp lợi nhuận kỳ vọng cho việc thực hiện hành động a ở trạng thái s và sau đó tuân theo một chính sách tối ưu. Do đó, chúng ta có thể viết q^* dưới dạng v^* như sau: $q^*(s,a) = E_\pi[R_{t+1} + \gamma v^*(\mathrm{S}_{t+1}) \mid S_t = s, A_t = \mathrm{a}]$ (1.11)

Vì v^* là hàm giá trị cho một chính sách, nó phải thỏa mãn điều kiện tự nhất quán được đưa ra bởi phương trình Bellman đối với các giá trị trạng thái (1.11). Tuy nhiên, vì đây là hàm giá trị tối ưu, điều kiện nhất quán của v^* có thể được viết ở dạng đặc biệt mà không cần tham chiếu đến bất kỳ chính sách cụ thể nào. Đây là phương trình Bellman cho v^*, hoặc phương trình tối ưu Bellman. Một cách trực quan, phương trình tối ưu Bellman diễn tả thực tế rằng giá trị của một trạng thái theo một chính sách tối ưu phải bằng với lợi tức kì vọng cho hành động tốt nhất từ trạng thái đó:

$$\begin{aligned} v^*(s) &= \max_{a \in A(s)} q_{\pi^*}(s,a) \\ &= \max_a E_{\pi^*}[G_t \mid S_t = s, A_t = \mathrm{a}] \\ &= \max_a E_{\pi^*}[\mathrm{R}_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = \mathrm{a}] \\ &= \max_a E[\mathrm{R}_{t+1} + \gamma v^* G_{t+1} \mid S_t = s, A_t = \mathrm{a}] \\ &= \max_a \sum_{s',r} p(\mathrm{s'},\mathrm{r}|\mathrm{s},\mathrm{a})\,[r + \quad v^*(s')] \end{aligned} \tag{1.12}$$

Hai phương trình cuối cùng là hai dạng của phương trình tối ưu Bellman cho v^* Phương trình tối ưu Bellman cho q^* là $q^*(s,a) = E_\pi[R_{t+1} + \gamma \max_{a'} q^*(\mathrm{S}_{t+1}, \mathrm{a}') \mid S_t = s, A_t = \mathrm{a}]$

$$= \sum_{s',r} p(\mathrm{s}', \mathrm{r} | \mathrm{s}, \mathrm{a})[r + \gamma \max_{a'} q^*(s', \mathrm{a}') \quad (1.13)$$

3. Q-Learning và Zap Q-learning

3.1. Q-Learning

Một trong những bước đột phá ban đầu trong việc học củng cố là thuật toán Q-learning (Watkins, 1989), được định nghĩa bởi:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)] \quad (2.1)$$

Trong trường hợp này, hàm giá trị hành động đã học, Q, gần đúng trực tiếp với q_* là hàm giá trị hành động tối ưu, độc lập với chính sách đang được tuân thủ. Điều này đơn giản hóa đáng kể việc phân tích thuật toán và kích hoạt các bằng chứng hội tụ sớm. Chính sách vẫn có một tính năng trong đó xác định cặp trạng thái-hành động nào được truy cập và cập nhật. Tuy nhiên, tất cả những gì cần thiết để hội tụ chính xác là tất cả các cặp tiếp tục được cập nhật. Theo giả định này và một biến thể của các điều kiện xấp xỉ ngẫu nhiên thông thường trên chuỗi các tham số kích thước bước, Q đã được chứng minh là hội tụ với xác suất 1 đến q_*. Thuật toán Q-learning được đưa ra sau đây ở dạng thủ tục.

Thuật toán Q-learning

Khởi tạo $Q(s,a)$ ***với mọi*** $s \in S^+$, $a \in A(s)$ ***ngoại trừ*** $Q(ter\min al,.) = 0$

Vòng lặp cho mỗi tập (episode):

Khởi tạo S

Vòng lặp cho mỗi bước của tập (episode):

Chọn A ***từ*** S ***bằng cách sử dụng chính sách bằng nguồn từ***

Thực hiện hành động A***, quan sát*** R , S'

$$Q(S,A) \leftarrow Q(S,A) + \alpha[R + \gamma \max_a Q(S',a) - Q(S,A)]$$

$$S \leftarrow S'$$

Cho tới khi S ***là kết thúc***

3.2. Zap Q-Learning

Hãy xem xét một mô hình MDP với không gian trạng thái X, không gian hành động U, hàm chi phí $\mathrm{c}: \mathrm{X} \times \mathrm{U} \rightarrow \mathbb{R}$ và hệ số chiết khấu $\beta \in (0,1)$. Giả định rằng trạng thái và không gian hành động

là hữu hạn: kí hiệu $l=|X|$, $l_u=|U|$ và P_u là ma trận xác suất chuyển có điều kiện cỡ $l\times l$, điều kiện $u\in U$. Quá trình trạng thái hành động (X,U) thích nghi với một bộ lọc $\{F_n: n\geq 0\}$ và Q1 được giả định trong suốt: Q1 là Quá trình chung (X,U) là một chuỗi Markov bất khả quy, với pmf ϖ bất biến duy nhất.

Hàm giá trị nhỏ nhất là lời giải duy nhất cho phương trình tối ưu chiết khấu chi phí:

$$h^*(x)=\min_{u\in \mathrm{U}} Q^*(x,u):=\min_{u\in \mathrm{U}}\left\{c(x,u)+\beta\sum_{x'\in \mathrm{X}}P_u\left(x,x'\right)h^*\left(x'\right)\right\},\quad x\in \mathrm{X} \tag{2.2}$$

"Q-function" là nghiệm của phương trình sau:

$$Q^*(x,u)=c(x,u)+\beta\sum_{x'\in X}P_u\left(x,x'\right)\underline{Q}^*\left(x'\right),\quad x\in X,u\in U \tag{2.3}$$

trong đó: $\underline{Q}(x):=\min_{u\in U}Q(x,u)$ với mọi $Q: X\times U\to R$

Giải thuật Zap Q - learning như sau:

Đầu vào: $\theta_0\in R^d,\zeta_0=\psi\left(X_0,U_0\right),\hat{A}_0\in R^{d\times d},n=0,T\in \mathsf{Z}^+$ ***(bước khởi tạo)***

Lặp:

$$\phi_n\left(X_{n+1}\right):=\arg\min_u Q^{\theta_n}\left(X_{n+1},u\right)$$
$$d_{n+1}:=c\left(X_n,U_n\right)+\beta Q^{\theta_n}\left(X_{n+1},\phi_n\left(X_{n+1}\right)\right)-Q^{\theta_n}\left(X_n,U_n\right);$$
$$A_{n+1}:=\zeta_n\left[\beta\psi\left(X_{n+1},\phi_n\left(X_{n+1}\right)\right)-\psi\left(X_n,U_n\right)\right]^T$$
$$\hat{A}_{n+1}=\hat{A}_n+\gamma_{n+1}\left[A_{n+1}-\hat{A}_n\right]$$
$$\theta_{n+1}=\theta_n-\alpha_{n+1}\hat{A}_{n+1}^{-1}\zeta_n d_{n+1}$$
$$\zeta_{n+1}:=\lambda\beta\zeta_n+\psi\left(X_{n+1},U_{n+1}\right)$$
$$n=n+1$$

cho tới khi $n\geq T$

3.3. Zap Q-learning cho thời điểm dừng tối ưu

Xem xét một chuỗi Markov thời gian rời rạc X = { X_n : n 0 } phát triển trên một không gian trạng thái X. Mục tiêu trong các vấn đề thời gian dừng tối ưu là cực tiểu hóa trên tất cả các thời gian dừng, kỳ vọng chi phí kết hợp là:

$$E\left[\sum_{n=0}^{\tau}\beta^n c(X_n)+\beta^\tau c_s(X_\tau)\right. \tag{2.4}$$

với $c: X\to R$ ký hiệu cho chi phí của mỗi trạng thái, $c_s: X\to\mathbb{R}$ ký hiệu cho chi phí cuối cùng, và $\beta\in[0,1]$ là hệ số chiết khấu. Ví dụ về các vấn đề như vậy phát sinh chủ yếu trong các ứng dụng tài chính như phân tích phái sinh. Thời điểm mua hoặc bán một tài sản và nói chung là trong các vấn đề như vậy thì liên quan đến phân tích tuần tự.

Trong công việc này, quy tắc quyết định tối ưu được tính gần đúng bằng cách sử dụng các kỹ thuật học tăng cường. Chúng tôi đề xuất và phân tích một thuật toán phương sai tối ưu để xấp xỉ hàm giá trị

được liên kết với quy tắc dừng tối ưu.

Chúng tôi giả định rằng không gian trạng thái $X \subset R^m$ là compact. Ký hiệu B là sig-ma đại số Borel. Chuỗi Markov thuần nhất được định nghĩa trong một không gian xác suất (Ω; F; P) và xác định phân phối ban đầu μ : X → [0; 1], và một hạt nhân chuyển tiếp P: cho mỗi $x \in X$ và $A \in B$:

$$P(x, A) = \Pr(\mathrm{X}_{n+1} \in \mathrm{A} \mid X_n = x)$$

Giả sử rằng X là ergodic thống nhất: Tồn tại một phép đo xác suất bất biến duy nhất π, một hằng số D nhỏ hơn vô cùng, và 0 < ρ < 1, như vậy, cho tất cả $x \in X$ và $A \in B$:

$$\left\| P^n(x, \mathrm{A}) - \pi(\mathrm{A}) \right\| \le D\rho^n$$

Kí hiệu $\{F_n : n \ge 0\}$ quá trình lọc liên quan đến X. Tính chất Markov khẳng định rằng đối với các hàm đo có giới hạn $h : X \to R$

$$E[\mathrm{h}(X_{n+1}) | F_n, X_n = x] = \int P(x, dy) h(y)$$

Trong bài báo này, thời gian dừng τ: Ω → [0; ∞] là một biến ngẫu nhiên nhận các giá trị trong các số nguyên không âm, với tính chất được định nghĩa $\{\omega : \tau(\omega) \le n ; \omega \in \Omega\} \in F_n$, với mọi n ≥ 0. Chính sách dừng được định nghĩa là một hàm đo được $\phi: X \to \{0,1\}$ xác định thời điểm dừng:

$$\tau^{\phi} = \min\{n \ge 0 : \phi(X_n) = 1\} \tag{2.5}$$

Hàm giá trị tối ưu được định nghĩa là cực tiểu của (2.4) trong tất cả các lần dừng, với mọi $x \in X$:

$$h^*(x) := \inf_{\tau} \mathrm{E}\left[\sum_{n=0}^{\tau} \beta^n c(X_n) + \beta^{\tau} c_s(X_{\tau}) \mid X_0 = x \right] \tag{2.6}$$

Tương tự, hàm Q liên quan được định nghĩa là:

$$Q^*(x) := c(x) + \beta E[h^*(X_1) \mid X_0 = x] \tag{2.7}$$

Theo đó, Q* là nghiệm của phương trình Bellman, với mọi $x \in X$

$$Q^*(x) = c(x) + \beta E\left[min\left(c_s(X_1), Q^*(X_1)\right) \mid X_0 = x \right] \tag{2.8}$$

và quy tắc dừng tối ưu được xác định bởi chính sách dừng tương ứng

$$\phi^*(x) = \mathrm{I}\left\{ c_s(x) \le Q^*(x) \right\} \tag{2.9}$$

Trong đó $I\{.\}$ là hàm chỉ thị. Sử dụng định nghĩa chung (3), thời gian dừng tối ưu thỏa mãn:

$$\tau^* = \tau^{\phi^*}$$

Phương trình Bellman (2.8) có thể được biểu diễn dưới dạng phương trình điểm cố định với Q* = FQ*, trong đó F biểu thị toán tử lập trình động: cho bất kỳ hàm Q : X → R và $x \in X$

$$FQ(x) = c(x) + \beta E\left[min\left(c_s(X_1), Q(X_1)\right) \mid X_0 = x \right] \tag{2.10}$$

Phân tích được đóng khung trong không gian Hilbert $L_2(\pi)$ thông thường của các hàm đo có giá trị

thực trên X với tích trong, và chuẩn (xem [4]) như sau:

$$\langle f, g \rangle_\pi = \mathrm{E}[f(X)g(X)] \text{ và } \|f\|_\pi = \sqrt{\langle f, f \rangle_\pi} \tag{2.11}$$

Trong đó kỳ vọng ở (2.10) liên quan đến trạng thái phân phối ổn định π. Giả định rằng các hàm chi phí c và c_s nằm trong $L_2(\pi)$

Mục tiêu trong công việc này là ước tính Q* bằng cách sử dụng một họ hàm số được tham số hóa Q^θ, trong đó $\theta \in R^d$ biểu diễn vector tham số. Chúng tôi giới hạn tham số hóa tuyến tính xuyên suốt, do đó: $Q^\theta(x) := \theta^T \psi(x)$, $x \in X$

Trong đó: $\psi := [\psi_1, \ldots, \psi_d]^T$ với $\psi_i : X \to R$, $\psi_i \in L_2(\pi)$, $1 \le i \le d$ biểu thị hàm cơ bản. Với bất kỳ vector tham số $\theta \in R^d$, chúng tôi ký hiệu sai số Bellman là: $B_\varepsilon^\theta = FQ^\theta - Q^\theta$

Giả định rằng các hàm cơ bản là độc lập tuyến tính, Ma trận hiệp phương sai d × d chiều: $\sum^\psi$ có hạng đầy đủ trong đó: $\sum^\psi (i,j) = \langle \psi_i, \psi_j \rangle_\pi$, $1 \le i, j \le d$

Thiết lập trong không gian trạng thái hữu hạn, có thể xây dựng một thuật toán nhất quán tính toán chính xác hàm Q. Mục tiêu trong phần này là tìm θ* sao cho:

$$E[\mathrm{B}_\varepsilon^{\theta^*}(X_n)\psi_i(X_n)] = 0, \ 1 \le i \le d$$

Hoặc tương đương: $\langle FQ^{\theta^*} - Q^{\theta^*}, \psi_i \rangle = 0$, $1 \le i \le d$

Trong [4], các tác giả đã chứng minh được:

$$Q^{\theta^*} - Q^*{}_\pi \le \frac{1}{1-\beta^2}\left[\min_\theta Q^\theta - Q^*{}_\pi\right]$$

Với mỗi $\theta \in R^d$ ký hiệu $\phi^\theta : X \to \{0,1\}$ có chính sách tương ứng

$$\phi^\theta(x) := I\{c_s(x) \le Q^\theta(x)\}$$

Đối với bất kỳ hàm f nào có tập xác định X, các toán tử S_θ, S_θ^c được định nghĩa như sau:

$$\mathsf{S}_\theta f(x) := I\{Q^\theta(x) < c_s(x)\} f(x)$$

$$\mathsf{S}_\theta^c f(x) := I\{c_s(x) \le Q^\theta(x)\} f(x)$$

Dễ thấy, với mỗi $x \in X$ thì $\mathsf{S}_\theta f(x) := (1 - \phi^\theta(x) f(x)$

Các tác giả trong [1] đã chứng minh được θ^* là nghiệm của phương trình:

$$A(\theta^*)\theta^* + \beta \bar{c}_s(\theta^*) + b^* = 0 \tag{2.12}$$

trong đó $\theta \in R^d$, $A(\theta)$ là ma trận d x d, và b^*, $\bar{c}_s$ là vector d chiều:

$$A(\theta) := \mathrm{E}[\psi(X_n)\beta \mathsf{S}_\theta \psi^T(X_{n+1}) - \psi(X_n)\psi^T(X_n)] \tag{2.13}$$

$$b^* := \mathrm{E}[\psi(X_n)c(X_n)]$$

$$\bar{c}_s(\theta) := \mathrm{E}[\psi(X_n)\mathsf{S}_\theta^c c_s(X_{n+1})]$$

Cho một dãy các ma trận thu hoạch d x d $\{G_n : n \ge 0\}$ và một chuỗi step-size vô hướng $\{\alpha_n : n \ge 0\}$,

thuật toán Q-learning tương ứng cho dừng tối ưu được đưa ra bởi thủ tục đệ quy sau:

$$\theta_{n+1} = \theta_n + \alpha_{n+1} G_{n+1} \psi(X_n) d_{n+1}$$

Với $\{d_n\}$ ký hiệu là chuỗi sai khác tạm thời:

$$d_{n+1} = c(X_n) + \beta \min\left(c_s(X_{n+1}), Q^{\theta_n}(X_{n+1})\right) - Q^{\theta_n}(X_n)$$

Thuật toán bộ lọc Kalman điểm cố định cũng có thể được viết dưới dạng trường hợp đặc biệt: Chúng ta có $G_n \equiv \left[\hat{\Sigma}_n^{\psi}\right]^{\dagger}$ với $M^{\dagger}$ kí hiệu giả nghịch đảo của ma trận M bất kỳ, $\hat{\Sigma}_{n+1}^{\psi}$ là ước lượng của trung bình Σ_{ψ}. Ước lượng có thể được đệ quy bằng cách sử dụng đệ quy Monte-Carlo tiêu chuẩn:

$$\hat{\Sigma}_{n+1}^{\psi} = \hat{\Sigma}_n^{\psi} + \alpha_{n+1}\left[\psi(X_n)\psi^T(X_n) - \hat{\Sigma}_n^{\psi}\right]$$

Trong thuật toán Zap-Q, dãy ma trận thu hoạch {Gn} được thiết kế sao cho hiệp phương sai tiệm cận của thuật toán kết quả được tối thiểu. Nó sử dụng ma trận thu hoạch $G_n = -\hat{A}_{n+1}^{\dagger}$ với $\hat{A}_{n+1}$ là một ước lượng của $A(\theta_n)$ với $A(\theta_n)$ được định nghĩa trong (2.13)

Số hạng bên trong kỳ vọng trong (2.13), sau thay thế θ = θn, được ký hiệu là:

$$A_{n+1} := \psi(X_n)\left[\beta \mathsf{S}_{\theta_n}\psi(X_{n+1}) - \psi(X_n)\right]^T \tag{2.14}$$

Sử dụng (2.14), ma trận $A(\theta_n)$ được ước tính đệ quy bằng cách sử dụng xấp xỉ ngẫu nhiên trong thuật toán Zap-Q:

Đầu vào: Khởi tạo $\theta_0 \in R^d$, $\hat{A}_0 : d \times d$ ***xác định âm; chuỗi step-size*** $\{\alpha_n\}$ ***và*** $\{\gamma_n\}$ ***và*** $n = 0$

Lặp lại:

Thu được số hạng Temporal Difference:

$$d_{n+1} = c(X_n) + \beta \min\left(c_s(X_{n+1}), Q^{\theta_n}(X_{n+1})\right) - Q^{\theta_n}(X_n)$$

Cập nhật ước lượng ma trận thu hoạch $\hat{A}_n$ ***của*** $A(\theta_n)$***, với*** A_{n+1} ***được định nghĩa trong (2.14):***

$$\hat{A}_{n+1} = \hat{A}_n + \gamma_{n+1}\left[A_{n+1} - \hat{A}_n\right]$$

Cập nhật vector tham số:

$$\theta_{n+1} = \theta_n - \alpha_{n+1}\hat{A}_{n+1}^{\dagger}\psi(X_n)d_{n+1}$$

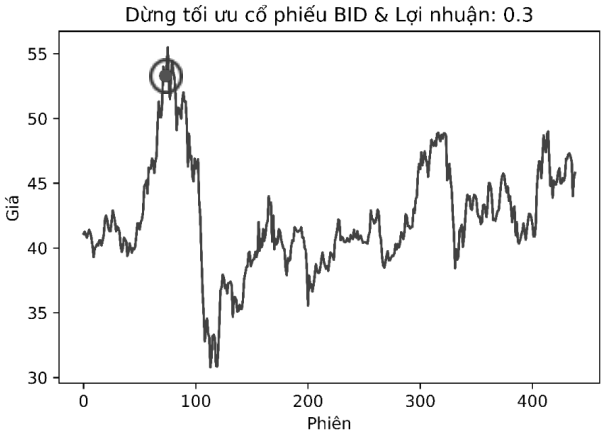
$$n = n+1$$

Tới khi $n \geq N$

Đầu ra: $\theta = \theta_N$

4. Kết quả thực nghiệm và kết luận

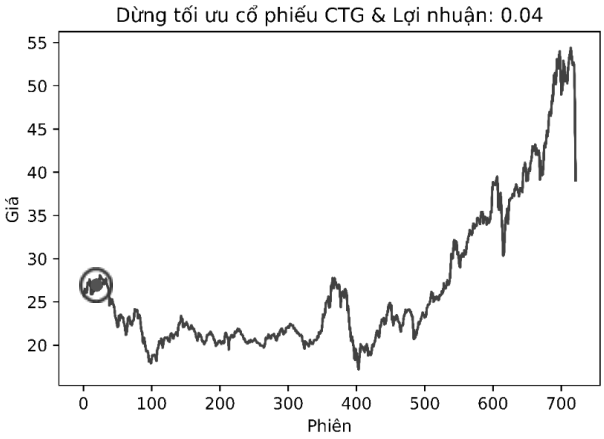
Chúng tôi đã cho thuật toán xác định thời điểm dừng tối ưu đối với 30 mã cổ phiếu của thị trường chứng khoán Việt Nam. Dữ liệu của quá khứ được dùng để huấn luyện các tham số. Sau khi có được các tham số đã huấn luyện, cho thuật toán chạy với dữ liệu thực, và giả sử rằng thời điểm mua là thời điểm bắt đầu chạy với dữ liệu thực. Khi điều kiện dừng thỏa mãn thì một lệnh bán được thực hiện.



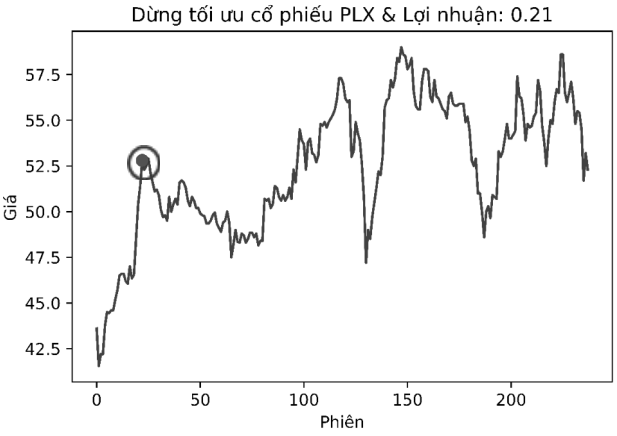
Hình 2. Kết quả dừng tối ưu với mã cổ phiếu BID.

Trong Hình 2 ta thấy, giá mua cổ phiếu là khoảng 41 và giá bán khoảng 53.5 và lợi nhuận sau khi chiết khấu khoảng 30%.

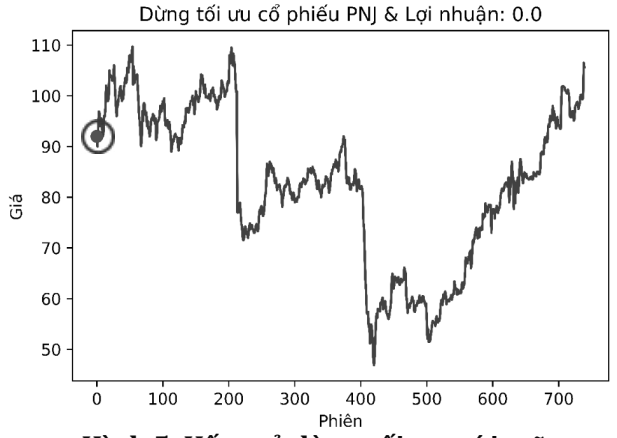
Dưới đây là kết quả dừng tối ưu và lợi nhuận sau chiết khấu đối với một số mã cổ phiếu:



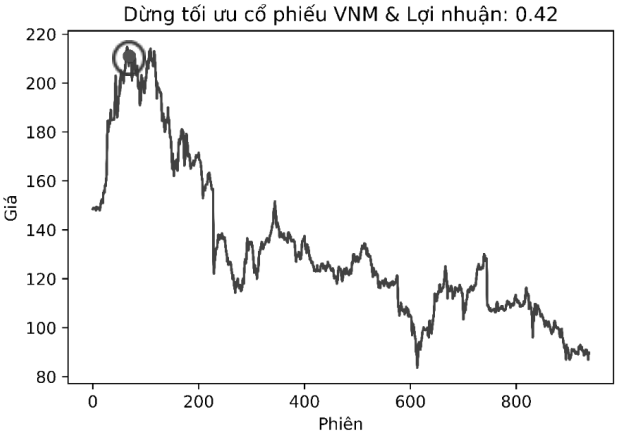
Hình 3. Kết quả dừng tối ưu với mã cổ phiếu CTG.



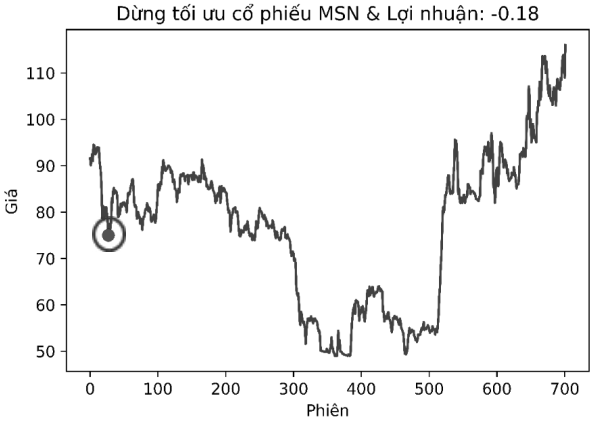
Hình 4. Kết quả dừng tối ưu với mã cổ phiếu PLX.



Hình 5. Kết quả dừng tối ưu với mã cổ phiếu PNJ.



Hình 6. Kết quả dừng tối ưu với mã cổ phiếu VNM.



Hình 7. Kết quả dừng tối ưu với mã cổ phiếu MSN.

Thống kê cho 30 cổ phiếu, sau khi đã chiết khấu lãi suất ngân hàng kết quả thu được lợi nhuận khoảng 3%. Đây là một kết quả khá khiêm tốn. Một điều chúng tôi nhận thấy quá trình đầu tư theo thuật toán trên đây thường có kết quả dừng rất sớm. Chúng tôi sẽ cải tiến thuật toán để hy vọng có kết quả lợi nhuận cao hơn.

Tài liệu tham khảo

[1] Chen, S., Devraj, A. M., Busic, A., Meyn, S., (2019), Zap Q-Learning for Optimal Stopping Time Problems, arXiv:1904.11538v3.

[2] Choi, D. and Van Roy, B., (2006), A generalized Kalman filter for fixed point approximation and efficient temporal-difference learning, Discrete Event Dynamic Systems: Theory and Applications, 16(2):207–239.

[3] Sutton, R. S. and Barto, A. G., (2018), Reinforcement Learning: An introduction, The MIT Press, Cambridge, Massachusetts.

[4] Tsitsiklis, J. N. and Van Roy, B., (1999), Optimal stopping of Markov processes: Hilbert space theory, approximation algorithms, and an application to pricing high-dimensional financial derivatives, IEEE Trans. Automat. Control, 44(10):1840–1851.

[5] Đặng Hùng Thắng, (2007), Giáo trình xác suất: Quá trình ngẫu nhiên và tính toán ngẫu nhiên, NXB Đại học quốc gia Hà Nội, Hà Nội.

[6] Nguyễn Duy Tiến, (2000), Các mô hình xác suất và ứng dụng: Phần I – Xích Markow và ứng dụng, NXB Đại học quốc gia Hà Nội, Hà Nội.