Journal of Applied And Basic Machine Intelligence Research (ABMIR)

پژوهش‌های نظری و کاربردی هوش ماشینی

سال دوم، شماره اول، بهار و تابستان ۱۴۰۳، صفحات۱۰۳ - ۹۲

معرفی یک روش مبتنی بر یادگیری تقویتی برای تعیین زمان و تعداد مناسب خرید سهام

فاطمه دره‌زرشکی، ولی درهمی*

دانشکده مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران

مقاله پژوهشی

تاریخ دریافت:
۱۴۰۲/۱۲/۲۰
تاریخ پذیرش:
۱۴۰۳/۰۳/۰۶

کلیدواژه‌ها:
بازار سهام، بهینه‌سازی هزینه‌های اجرایی سهام، یادگیری تقویتی، یادگیری کیو

نویسنده‌مسئول:
vderhami@yazd.ac.ir

چکیده

نوسان قیمت و عدم اطمینان موجود در بازار، تعیین استراتژی بهینه برای خرید سهام را به یک فرایند پیچیده تبدیل کرده است. عدم تکرار شرایط یک معامله، لزوم یادگیری به‌صورت تعاملی را ایجاب می‌کند. یادگیری تقویتی یک روش یادگیری تعاملی است که تنها با استفاده از یک سیگنال اسکالرِ راندمان، می‌تواند پارامترهای سیستم را تنظیم نماید. در این مقاله با تعریف مناسب حالت‌های سیستم شامل گام زمانی، تعداد کل سهام خریداری‌شده تا گام زمانی فعلی، میزان انحراف معیار قیمت سهام از گام نخست تا گام زمانی موردنظر و میزان تغییرات قیمت نسبت به گام زمانی قبل و همچنین تعریف مناسب سیگنال تقویتی، از روش یادگیری کیو به‌عنوآن‌یکی از معروف‌ترین الگوریتم‌های یادگیری تقویتی برای تقریب توابع ارزش حالت-عمل استفاده می‌شود. در این پژوهش، بازار سهام با توجه به روابط ریاضی موجود، مدل شده و روش ارائه‌شده در آن به کار گرفته‌شده است. عملکرد استراتژی حاصل از مدل پیشنهادی با استراتژی بازگشت به میانگین در ۵۰۰۰ بازارِ شبیه‌سازی‌شده مورد مقایسه قرارگرفته است. نتایج نشان‌دهنده آن است که بهره‌گیری از مدل پیشنهادی در مقایسه با استراتژی بازگشت به میانگین نه‌تنها هزینه متوسط پایین‌تر، بلکه قابلیت اطمینان بسیار بالاتری نیز دارد.

: 10.22034/ABMIR.2024.21357.1050

E-ISSN: 2821-2037

۱- مقدمه

به‌علت دینامیک پیچیده معاملات در بازارهای سهام[1]، توسعه سیستم‌های معاملاتی خودکار همواره موردتوجه معامله‌گران و تحلیل‌گران مالی بوده است. تصمیم‌گیری در بازارهای سهام با درجه بالایی از ریسک و عدم اطمینان همراه است. محیط بازار سرشار از عدم قطعیت است و نوسانات موجود درروند قیمت، ارزیابی شرایط آتی بازار را بسیار چالش‌برانگیز می‌کند. با توجه به نوسان‌ها و غیریکنواخت‌بودن روند قیمت سهام، اتخاذ تصمیمات معاملاتی در بازارهای سهام یک فرایند پیچیده است و تضمینی برای دستیابی به بهینه سراسری وجود ندارد. تاکنون کارهای زیادی در حوزه پیش‌بینی قیمت سهام انجام‌شده است تا زمینه تصمیم‌گیری بهینه را برای سرمایه‌گذاران فراهم آورد. به‌ویژه، برخی از سیستم‌های اخیر که از روش‌های یادگیری ماشین مانند شبکه‌های عصبی مصنوعی[2] استفاده می‌کنند [۳-۱]؛ اما این سیستم‌ها دارای این محدودیت هستند که عمدتاً مبتنی بر یادگیری با ناظر[3] می‌باشند. باوجود قابلیت بالای این روش یادگیری، استفاده از آن به‌تنهایی در کاربردهایی که داده‌های آموزشی در دسترس نیست و نیازمند یادگیری از طریق تعامل با محیط هستند، مناسب نیست [٤]. هرچند پیش‌بینی قیمت‌های آتی سهام یکی از مسائل مهم در بازارهای سهام محسوب می‌شود، مسائل مهم دیگری نیز وجود دارند که هدف آن‌ها اتخاذ اقدامات مناسب با توجه به روند قیمت سهام است.

بهینه‌سازی هزینه‌های اجرایی سهام[4] (OEC) یک مسئله مهم و کاربردی است که در آن‌یک معامله‌گر می‌خواهد هزینه خرید یک تعداد از پیش تعریف‌شده سهام را در یک افق زمانی ثابت به حداقل برساند. مسئله بهینه‌سازی هزینه‌های اجرایی سهام برای اولین بار در سال ۱۹۹۸ [٥] معرفی شد. این مقاله چند مدل برنامه‌سازی پویا[5] پیشنهاد می‌دهد؛ مدل اول فرض می‌کند که قیمت‌های آتی تنها تحت تأثیر قیمت‌های گذشته و فعالیت‌های تجاری شخص معامله‌گر قرار می‌گیرند. مدل‌های دوم و سوم اجزایی را به رابطه قیمت اضافه می‌کنند تا اطلاعات مربوط به فعالیت‌های تجاری سایر معامله‌گران در بازار و نویز خارجی را منعکس کنند. از آن زمان، اصلاحات و الحاقات مختلفی در مطالعات بعدی ارائه شد. به‌عنوان‌مثال، انواع مختلفی از ریسک‌ها در معاملات گنجانده‌شده است [۹-٦]؛ هر دو نوع معاملات گسسته و پیوسته ترکیب‌شده‌اند [۱۰] و مسئله OEC با استفاده از رویکردهای دیگر مانند معادلات دیفرانسیل موردمطالعه قرارگرفته است [۱٤-۱۱]. پژوهش‌های مطرح‌شده، از رویکرد برنامه‌سازی پویا به‌منظور استخراج استراتژی‌های معاملاتی استفاده می‌کنند. این رویکرد محدودیت‌های خاص خود را دارد، زیرا دانش کاملی را از مدل محیط فرض می‌کند که معمولاً در موقعیت‌های واقعی وجود ندارد. در بازارهای مالی، عامل باید یک محیط ناشناخته را کاوش کند و تصمیمات را به‌صورت برخط[6] اتخاذ نماید.

یادگیری تقویتی[7] یک چارچوب برای حل مسائل کنترل بهینه به‌صورت تعاملی با کمترین اطلاعات است که می‌تواند به‌عنوان‌یک رویکرد جایگزین در این مسئله مورداستفاده قرار گیرد. موفقیت‌های یادگیری تقویتی به‌طور گسترده در کاربردهایی ازجمله مسیریابی روبات‌ها [۱٥]، بازی آتاری[8] [۱٦] و کنترل هلیکوپتر [۱۷] نشان داده‌شده است. حتی در برخی آزمایش‌ها، یادگیری تقویتی در اجرای سیاست‌های کنترل بهینه از متخصصان انسانی بهتر عمل می‌کند [۱٦ و ۱۸]. سوالی که مطرح می‌شود این است که آیا می‌توان در بازارهای مالی یک مدل یادگیری تقویتی را به‌منظور شکست دادن معامله‌گران انسانی آموزش داد؟

در مقایسه با سایر مسائل یادگیری تقویتی، بهینه‌سازی هزینه‌ها در بازارهای مالی امری چالش‌برانگیز است. بازارهای مالی محیط‌هایی نامطمئن و پرخطر هستند. سری‌های زمانی قیمت، بسیار تصادفی و حاوی مقدار زیادی نویز و پرش می‌باشند. بنابراین، توزیع

[1] Stock markets
[2] Artificial neural networks
[3] Supervised learning
[4] Optimizing execution costs
[5] Dynamic programming
[6] Online
[7] Reinforcement learning
[8] Atari game

پاداش‌ها ذاتاً تصادفی است. در چنین شرایطی تعادل میان کاوش[1] و بهره‌گیری[2]، از اهمیت ویژه‌ای برخوردار است. چگونگی تعریف حالت[3] محیط بر دقت و عملکرد الگوریتم‌های یادگیری تقویتی بسیار مؤثر است. از دیگر چالش‌هایی که در این مسئله با آن روبه‌رو هستیم، چگونگی تعریف حالت محیط است؛ به‌گونه‌ای که بتواند ویژگی مارکوف[4] را برآورده کند.

یادگیری کیو[5] یک روش مستقل از مدل[6] و مبتنی بر ارزش[7] است که برای حل مسائل تعریف‌شده در فضای گسسته مورداستفاده قرار می‌گیرد. در این روش، یادگیری مستقیماً از طریق تجربه و بدون نیاز به یک مدل کامل از دینامیک محیط صورت می‌گیرد و همگرایی آن در هر دو محیط قطعی[8] و تصادفی[9] تضمین‌شده است [۱۹ و ۲۰].

در این مقاله، مسئله بهینه‌سازی هزینه‌های اجرایی سهام به‌عنوان‌یک فرایند تصمیم‌گیری مارکوف[10] (MDP) بیان می‌شود و یک مدل جدید یادگیری تقویتی بر اساس یادگیری کیو برای حل مسئله موردنظر ارائه می‌شود. مدل پیشنهادی منجر به یک سیاست معاملاتی می‌شود که نسبت به استراتژی بازگشت به میانگین[11] بهبود قابل‌توجهی دارد. سهم علمی مقاله به شرح زیر است:

۱- تعریف مناسب فضای حالت و عمل برای مسئله بهینه‌سازی هزینه‌های اجرایی خرید سهام

۲- تعریف سیگنال تقویتی در جهت رسیدن به هدف اصلی بهینه‌سازی

۳- به‌کارگیری روش یادگیری کیو در مسئله مذکور.

ادامه مقاله بدین شرح سازمان‌دهی شده است: در بخش ۲ مفاهیم اولیه‌ای که برای هدف ما موردتوجه است، شرح داده می‌شود. در بخش ۳، مدل پیشنهادی برای مسئله موردنظر ارائه می‌گردد. در بخش ٤، نتایج تجربی و در بخش ۵، نتیجه‌گیری و پیشنهادات برای تحقیقات آینده بیان می‌شود.

۲- مفاهیم اولیه

۲-۱ بهینه‌سازی هزینه‌های اجرایی سهام

در کنار هزینه خرید سهام، هزینه‌های دیگری مانند کمیسیون، اختلاف قیمت پیشنهادی و قیمت فروش و تأثیر تصمیمات معاملاتی فعلی خریدار بر قیمت‌های آتی وجود دارد. چنین هزینه‌هایی به‌عنوان هزینه‌های اجرایی شناخته می‌شوند [۵]. در مسئله بهینه‌سازی هزینه‌های اجرایی سهام، یک معامله‌گر می‌خواهد هزینه خرید تعداد کل $\bar{S}$ سهام را در بازه زمانی ثابت T به حداقل برساند. اگر تعداد سهام خریداری‌شده در دوره t به قیمت P_t را با S_t نشان دهیم به‌طوری‌که $t = ۱,۲, \dots, T$ ، هدف سرمایه‌گذار از به حداقل رساندن هزینه‌های اجرایی به‌صورت رابطه (۱) بیان می‌شود [۵].

$$\begin{aligned} &\text{Minimize} \sum_{t=1}^{T} P_t S_t \\ &\text{s.t.} \sum_{t=1}^{T} S_t = \bar{S}, \\ &S_t \geq 0, t = 1,2, \dots, T. \end{aligned} \qquad (۱)$$

برای اینکه تعریف مسئله را کامل کنیم، نیاز داریم که قانون حرکت P_t را مشخص نماییم. فرض کنید برای قیمت، یک قیمت تعادلی[12] P_e وجود دارد به‌طوری‌که $x_t = \log\frac{P_t}{P_e}$ دارای دینامیکی بر اساس رابطه (۲) است.

$$dx_t = -\lambda x_t + \sigma\xi_t \qquad (۲)$$

در رابطه (۲)، $\xi_t \sim N(۰,۱)$ یک مقدار نویز گوسی[13] است که نشان‌دهنده شوک‌ها و رویدادهای پیش‌بینی‌نشده در بازار سهام است. رابطه (۲) یک گسسته سازی استاندارد از فرایند اورنشتاین-اولنبک[14] است و بیان‌کننده این است که قیمت P_t تمایل دارد با نرخ λ به قیمت تعادلی P_e بازگردد [۲۱].

[1] Exploration
[2] Exploitation
[3] State
[4] Markov property
[5] Q-learning
[6] Model-free
[7] Value-based
[8] Deterministic
[9] Stochastic
[10] Markov decision process
[11] Mean reversion strategy
[12] Equilibrium price
[13] Gaussian noise
[14] Ornstein-Uhlenbeck process

۲-۲- فرایند تصمیم‌گیری مارکوف

یک فرایند تصمیم‌گیری مارکوف، توسط چندگانه $\{S, A, R, T, \gamma\}$ تعریف می‌شود. این فرایند تصمیم‌گیری، یک مسئله تصمیم‌گیری گسسته و تصادفی را بیان می‌کند که از یک مجموعه متناهی شامل حالت‌ها $S = \{s_0, s_1, \dots, s_{n-1}\}$، مجموعه متناهی $A(s)$ شامل عمل‌های[1] مجاز در هر حالت s، یک تابع پاداش[2] $R: S \times A \times S \to \mathbb{R}$ و یک تابع احتمال انتقال[3] به‌صورت $T: S \times A \times S \to [۰,۱]$ تشکیل‌شده است. $\gamma \in [۰,۱]$ فاکتور تخفیف[4] برای پاداش‌های آینده است که ارزش نسبی پاداش‌های باتاخیر[5] را در مقابل پاداش‌های لحظه‌ای تعیین می‌کند. خروجی این فرایند تصمیم‌گیری، یک سیاست[6] (یک نگاشت از حالت‌ها به عمل‌ها) است که پاداش مورد انتظار را به حداکثر می‌رساند [٤]. ویژگی مارکوف بیان می‌کند که تصمیم‌گیری در یک حالت به حالت یا حالت‌های قبل بستگی ندارد و یک حالت به‌تنهایی، یک ساختار داده به‌اندازه کافی غنی برای تصمیم‌گیری بهینه است [٤]. در گام زمانی t در حالت s_t، عامل عمل $a_t \in A(s_t)$ را انتخاب می‌کند، در پاسخ عمل a_t، فرایند با احتمال $T \to [۰,۱] \in T$ به حالت s_{t+1} می‌رود و عامل پاداش $r_{t+1} \in R$ را دریافت می‌کند.

۲-۳ یادگیری تقویتی

یادگیری تقویتی یکی از روش‌های یادگیری است که برای حل مسئله تصمیم‌گیری مارکوف ارائه‌شده است. در یادگیری تقویتی، یادگیری از طریق تعامل مستقیم با محیط و بدون نیاز به ناظر اتفاق می‌افتد. این روش یادگیری بر پایه مفهوم پاداش یا جریمه برای عمل‌های انجام‌شده بناشده است. در یادگیری تقویتی، یادگیری درنتیجه تفسیر عامل از تعامل با محیط اطراف و پیامدهای سیگنال پاداش اتفاق می‌افتد [٤]. عامل در محیطی عمل می‌کند که اطلاعات جزئی از آن دارد. برای اینکه عامل بتواند به بهترین رفتار دست یابد، یعنی پاداش خود را به حداکثر برساند، باید از طریق تجربه یاد بگیرد و دانش خود را به‌روز کند. یادگیری تقویتی با یادگیری با ناظر که در آن یادگیری بر اساس داده‌های آموزشی و توسط یک ناظر خارجی خبره صورت می‌گیرد، متفاوت است. یادگیری با ناظر نوع مهمی از یادگیری است؛ اما استفاده از آن در مسائل تعاملی که اغلب به‌دست‌آوردن داده‌های آموزشی دشوار است، مناسب نیست. بنابراین، در محیط‌های ناشناخته (مانند بازارهای مالی) عامل باید بتواند از تجربیات گذشته خود یاد بگیرد.

در چارچوب یادگیری تقویتی، فرایند تصمیم‌گیری، مارکوف در نظر گرفته می‌شود. به‌طورکلی، هدف عامل بیشینه‌سازی بازگشت[7] مورد انتظار است که با نماد R_t و به‌صورت تابعی از پاداش آنی و پاداش‌های تخفیف‌یافته آتی تعریف می‌شود. این تابع را می‌توان به‌صورت رابطه (۳) نوشت.

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+(k+1)} \qquad (۳)$$

به‌طوری‌که $۰ \le \gamma \le ۱$ است.

یک سیاست $\pi: S \to A$، یک نگاشت از مجموعه حالت‌ها به مجموعه عمل‌ها است و نشان می‌دهد که در هر حالت کدام عمل باید انتخاب شود. $\pi(s_t) = a_t$، انتخاب عمل a_tدر حالت s_t را نشان می‌دهد. سیاست بهینه، سیاستی است که بازگشت مورد انتظار از یک حالت را برای همه حالت‌ها بیشینه می‌کند. وظیفه یادگیری تقویتی، جستجو برای یافتن یک سیاست نزدیک به سیاست بهینه است.

۲-٤ توابع ارزش[8]

رویکردهای یادگیری تقویتی عموماً مبتنی بر تخمین توابع ارزش هستند. توابع ارزش حالت (حالت-عمل) به هر حالت (حالت-عمل) مقداری را متناسب با امید دریافت پاداش‌های آتی نسبت می‌دهند. این توابع میزان خوب بودنِ قرار گرفتنِ عامل در یک حالت معین (میزان خوب بودن انجام یک عمل معین در یک حالت معین) را تخمین می‌زنند. میزان خوب بودن، بر اساس بازگشت مورد انتظار R_t تعیین می‌شود. به‌طور خاص، ارزش حالت $s_t = s$

[1] Action
[2] Reward function
[3] Transaction probability function
[4] Discount factor
[5] Delayed rewards
[6] Policy
[7] Return
[8] Value functions

تحت سیاست π برابر با بازگشت مورد انتظاری است که با شروع از $s_t = s$ و پیروی از سیاست π به دست می‌آید و به‌صورت $V^{\pi}(s)$ نمایش داده می‌شود و بیان ریاضی آن مطابق رابطه (٤) است.

$$V^{\pi}(s) = E_{\pi}[R_t | s_t = s] \quad (٤)$$

به‌طور مشابه، ارزش انجام عمل $a_t = a$ در حالت $s_t = s$ تحت سیاست π برابر با بازگشت مورد انتظاری است که با شروع از حالت $s_t = s$، انجام عمل $a_t = a$ و پیروی از سیاست π به دست می‌آید و به‌صورت $Q^{\pi}(s,a)$ نمایش داده می‌شود. رابطه (٥) تعریف ریاضی آن را بیان می‌کند.

$$Q^{\pi}(s,a) = E_{\pi}[R_t | s_t = s, a_t = a] \quad (٥)$$

یک خصوصیت مهم توابع ارزش این است که آن‌ها رابطه بازگشتی خاصی را ارضا می‌کنند. تحت هر سیاست π و برای هر حالت $s_t = s$، رابطه (٦) بین ارزش حالت s_t و ارزش حالت s_{t+1} برقرار است.

$$V^{\pi}(s) = E_{\pi}[r_{t+1} + \gamma V^{\pi}(s_{t+1}) | s_t = s] \quad (٦)$$

رابطه (٦)، معادله بلمن[1] برای $V^{\pi}(s_t)$ نامیده می‌شود و اثبات می‌شود که $V^{\pi}(s)$، راه‌حلی یکتا برای معادله بلمن متناظر است [٤].

۲-۵ یادگیری کیو

یادگیری کیو، یک روش مستقل از مدلِ برون‌سیاست[2] برای حل مسئله کنترل است که اولین بار در سال ۱۹۹۲ [۲۲] معرفی شد. این الگوریتم راه‌حلی را برای مسائل تصمیم‌گیری که به‌عنوان فرایند تصمیم‌گیری مارکوف بیان می‌شوند، ارائه می‌دهد. الگوریتم (۱)، شبه کد الگوریتم یادگیری کیو را نشان می‌دهد.

قبل از شروع فرایند یادگیری، مقادیر Q به‌صورت تصادفی مقداردهی اولیه می‌شوند. فرایند یادگیری به‌صورت تکراری پیش می‌رود. در هر تکرار، عامل حالت $s_t = s$ را مشاهده می‌کند، سپس عمل $a_t = a$ را با استفاده از یکی از روش‌های انتخاب عمل (به‌عنوان‌مثال، روش شبه‌حریصانه[3]) انتخاب و به محیط اعمال می‌کند. در پاسخ، محیط به حالت s_{t+1} می‌رود و عامل پاداش r_{t+1} را دریافت می‌کند. در پایان هر تکرار، عامل تابع ارزش Q را با استفاده از رابطه (۷) به‌روزرسانی می‌کند.

$$Q(s_t,a_t) = Q(s_t,a_t) + \alpha\left(r_{t+1} + \gamma\, max_{a_{t+1}} Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)\right) \quad (۷)$$

$\alpha \in (۰,۱]$ نرخ یادگیری[4] و $\gamma \in [۰,۱]$ فاکتور تخفیف است.

الگوریتم (۱): الگوریتم یادگیری کیو [٤]

1: Input parameters: learning rate $\alpha \in (0,1]$, $\varepsilon \in [0,1]$.
2: Initialize $Q(s,a)$, for all $s \in S$, $a \in A(s)$ arbitrarily.
3: for each episode:
4: Initialize s.
5: while (notIsTerminal(s)):
6: Choose action a from $A(s)$ using policy derived from Q.
7: Take action a, observe r and s'.
8: $Q(s,a) = Q(s,a) + \alpha(r + \gamma\, max_{a'}\, Q(s',a') - Q(s,a))$.
9: $s = s'$
10: end while
11: end for
12: for each $s \in S$:
13: $\pi(s) = argmax_a\, Q(s,a)$
14: end for
15: Output: $\pi(s) \in A(s), \forall s \in S$

عبارت $r_{t+1} + \gamma max_{a_{t+1}} Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)$ خطای تفاضل موقتی[5] نامیده می‌شود. این مقدار، اختلاف بین پاداش دریافتی و پاداش تخمینی عمل را نشان می‌دهد. نرخ یادگیری α، تأثیر مقدار خطای تفاضل موقتی را در به‌روزرسانی مقادیر Q تعیین

[1] Bellman equation
[2] Off-policy
[3] ε-greedy
[4] Learning rate
[5] Temporal difference error

می‌کند. به‌عبارت‌دیگر، نرخ یادگیری میان تجربه‌های قدیمی و تجربه‌های جدید تعادل ایجاد می‌کند. فاکتور تخفیف γ، تعیین می‌کند که تا چه اندازه مقادیر Q تحت تأثیر پاداش‌های آتی قرار خواهند گرفت. هدف از تخمین تابع ارزشِ حالت–عمل، یافتن سیاست است.

در بخش بعد، جزئیات مدل یادگیری کیوی پیشنهادی برای حل مسئله OEC بیان می‌شود.

۳- مدل پیشنهادی مبتنی بر یادگیری کیو

در حل مسئله به کمک روش‌های یادگیری تقویتی، چگونگی تعریف MDP مدل از اهمیت ویژه‌ای برخوردار است. حالت‌ها، مجموعه عمل‌های مجاز در هر حالت و تابع پاداش باید به‌گونه‌ای تعریف شوند که هدف مسئله محقق شود. در این بخش، ابتدا چگونگی تعریف حالت محیط، مجموعه عمل‌ها، سیگنال پاداش، تابع احتمال انتقال حالت محیط و فاکتور تخفیف بیان می‌شود و سپس به چگونگی تأثیر فعالیت تجاریِ شخص خریدار بر قیمت آتی پرداخته می‌شود.

۳-۱ حالت محیط

نحوه نمایش حالت محیط بر کارایی الگوریتم‌های یادگیری بسیار مؤثر است؛ زیرا عامل در هرلحظه بازار را تنها از طریق حالت محیط در آن لحظه درک می‌کند. بنابراین، حالت محیط باید یک ساختار داده به‌اندازه کافی غنی برای اتخاذ تصمیمات معاملاتی مناسب باشد. در این مسئله، حالت محیط را در هر گام زمانی با استفاده از چهار متغیر تعریف می‌نماییم:

- گام زمانی t
- تعداد سهام خریداری‌شده تا گام زمانی t
- انحراف معیار قیمت‌ها ($\{P_1, P_2, \dots, P_t\}$) تا گام t
- $\Delta P_t = P_t - P_{t-1}$

گام زمانی و تعداد سهام خریداری‌شده مقادیری گسسته هستند؛ اما مقادیر انحراف معیار قیمت و تغییرات قیمت مقادیری پیوسته می‌باشند و برای پیاده‌سازی الگوریتم یادگیری کیو باید گسسته‌سازی شوند. الگوریتم (۲) و الگوریتم (۳)، به ترتیب، روش گسسته‌سازی انحراف معیار قیمت و تغییرات قیمت را نشان می‌دهند.

الگوریتم (۲): گسسته‌سازی مقادیر انحراف معیار قیمت

گام اول

1: Input parameters:
parameters of the problem dynamics, equation (2).
2: std_list = []
3: for $counter = 1{:}\,1000$
4: for $i = 1{:}\,T$
5: GenerateP_i using equation (2).
6: Calculate the standard deviation of
$[P_1, P_2, \dots, P_i]$ and append it to std_list
7: end for
8: end for
9: Remove the outliers from std_list using boxplot.
10: $Median$ = median(std_list)
11: Max = max(std_list)
12: Output: Max and $Median$.

گام دوم

1: Function std_Discretization(Inputs) Return Output.
2: Inputs: std, Max, and $Median$
3: if $std < \frac{Median}{2}$:
4: $std = 0$
5: elseif $std < Median$:
6: $std = 1$
7: elseif $std < \frac{Median+Max}{2}$:
8: $std = 2$
9: else:
10: $std = 3$
11: end if
12: Output: std

[1] Outliers

گسسته‌سازی متغیرهای موردنظر، بر اساس قیمت‌های به‌دست‌آمده از دینامیک مسئله (رابطه (۲)) انجام می‌شود.

بر اساس الگوریتم (۲)، ابتدا مقادیر متغیر انحراف معیار قیمت به کمک دینامیک مسئله تولید می‌شود، سپس پس از حذف داده‌های پرت[1] به کمک نمودار جعبه‌ای، مقدار میانه و مقدار بیشینه این متغیر تعیین می‌شود و بر اساس میانه و مقدار بیشینه، مقدار انحراف معیار قیمت به چهار وضعیت کم، متوسط، زیاد و خیلی زیاد گسسته می‌شود.

به‌منظور گسسته‌سازی مقدار تغییرات قیمت، بر اساس الگوریتم (۳)، مقادیر متغیر تغییرات قیمت با استفاده از دینامیک مسئله تولید و پس از حذف داده‌های پرت، مقدار کمینه و بیشینه این متغیر تعیین می‌شود. سپس با به‌کارگیری تابع خطی، متغیر موردنظر در بازه [۱,۱-] نرمال می‌گردد و در پایان به سه وضعیت منفی، صفر و مثبت گسسته‌سازی می‌شود.

۳-۲ عمل‌ها

بدون از دست دادن کلیت مسئله، فرض می‌شود که یک خریدار می‌خواهد تعداد $\bar{S} = ۱۰۰۰$ سهام را در $T = ۲۰$ روز خریداری نماید. عمل‌ها را تعداد سهامی که در یک گام زمانی می‌تواند خریداری شود، در نظر می‌گیریم. برای سادگی، فضای عمل‌ها را گسسته و مجموعه عمل‌ها در هر حالت را به‌صورت زیرمجموعه‌ای از $A = \{۰,۱۰۰,۲۰۰,۳۰۰,۴۰۰\}$ تعریف می‌کنیم؛ بدین معنا که در یک گام زمانی عامل می‌تواند تعداد صفر، ۱۰۰، ۲۰۰، ۳۰۰ و حداکثر ۴۰۰ سهام را خریداری نماید.

۳-۳ تابع پاداش

سیگنال پاداش را هزینه خرید سهام و به‌صورت رابطه (۸) تعریف می‌نماییم.

$$r_{t+1} = -a_t P_t - 0.01(a_t P_t) - C \qquad (۸)$$

پارامتر C، یک هزینه ثابت است که به‌عنوان هزینه معامله[2] در نظر گرفته‌شده است و a_t تعداد سهام خریداری‌شده در گام زمانی t به قیمت P_t را نشان می‌دهد.

الگوریتم (۳): گسسته‌سازی مقادیر تغییرات قیمت

گام اول

1: Input parameters:
parameters of the problem dynamics, equation (2).
2: deltaP_list = []
3: for $counter = 1:1000$
4: for $i = 1:T$
5: GenerateP_i using equation (2).
6: Calculate$P_i - P_{i-1}$ and append it to deltaP_list.
7: end for
8: end for
9: Remove the outliers from deltaP_list using boxplot
10: $Min = \min$(deltaP_list)
11: $Max = \max$(deltaP_list)
12: Output: Min and Max.

گام دوم

1: Function detaP_Discretization(Inputs) Return Output.
2: Inputs: ΔPt, Min, and Max
3: $m = \frac{2}{Max-min}$
4: $y = m(\Delta P_t - Max) + 1$
5: if $y < -0.1$
6: $\Delta P_t = 0$
7: elseif $y > 0.1$
8: $\Delta P_t = 2$
9: else
10: $\Delta P_t = 1$
11: end if
12: Output: ΔP_t

[1] Outliers
[2] Transaction cost

۳-٤ تابع احتمال انتقال حالت محیط

در مسئله موردنظر، یک سیستم پویا داریم که می‌توان آن را در نقاط زمانی گسسته مشاهده نمود. در این سیستم، بردار حالت شامل دو جزء قطعی و دو جزء تصادفی است. به همین علت، انتقال بین حالت‌ها، نه‌تنها به عمل انجام‌شده، بلکه به تغییر در اجزاء تصادفی حالت نیز بستگی دارد. هنگامی‌که عامل در یک حالت خاص، عملی را به محیط اعمال می‌کند، محیط به حالتی جدید منتقل می‌شود و عامل پاداشی را دریافت می‌نماید؛ اما فرایند تصمیم‌گیری مارکوف موردنظر، احتمال واقعی انتقال حالت را در اختیار عامل قرار نمی‌دهد. بنابراین در این مسئله، تابع احتمال انتقال حالت در دسترس نمی‌باشد و برای آموزش مدل یادگیری تقویتی، روش مستقل از مدل یادگیری کیو مورداستفاده قرار می‌گیرد.

۳-۵ فاکتور تخفیف

ازآنجاکه هدف، بهینه‌سازی هزینه‌ها در بلندمدت است، مقدار فاکتور تخفیف برابر ۱ در نظر گرفته می‌شود. با در نظر گرفتن $\gamma = ۱$، با تمام پاداش‌های آینده بدون توجه به فاصله زمانی از گام فعلی، به‌طور مساوی رفتار می‌شود و پاداش تخفیف‌یافته وجود نخواهد داشت. بنابراین، $\gamma = ۱$ برای اهداف بلندمدت مناسب‌تر است.

۳-٦ تأثیر فعالیت تجاری خریدار بر قیمت

رابطه (۲)، به‌عنوان سری زمانی قیمت در نظر گرفته‌شده است. در این رابطه، تأثیر فعالیت تجاری شخص خریدار بر قیمت آتی لحاظ نشده است؛ درحالی‌که ممکن است پس‌ازاینکه خریدار اقدام به خرید سهام می‌کند، قیمت به علت کاهش عرضه افزایش یابد. به‌منظور لحاظ کردن تأثیر خرید شخص خریدار، پس از خرید، قیمت P_t بر اساس رابطه (۹) افزایش می‌یابد و سپس در محاسبه P_{t+1} مورداستفاده قرار می‌گیرد.

$$P_t = P_t + \frac{factor}{100} a_t P_t \qquad (۹)$$

پارامتر $factor \in [۰,۱]$، درصد افزایش قیمت را تعیین می‌کند. در مدل موردنظر، پارامتر $factor$ برابر ۰/۰۱ در نظر گرفته‌شده است؛ بدین معنا که به ازای خرید هر صد سهام، به‌اندازه یک درصد مقدار قیمت افزایش می‌یابد.

٤- آزمایش تجربی و نتایج

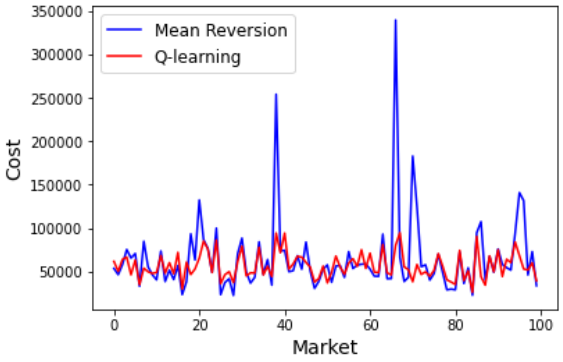
بدون از دست دادن کلیات مسئله، الگوریتم یادگیری کیو، برای بهینه‌سازی هزینه خرید $\bar{S} = ۱۰۰۰$ سهام در دوره‌ای به طول $T =$ ۲۰ روز به کار می‌رود. در مدل پیشنهادی، بازار را اعداد تصادفی تولیدشده از یک توزیع تصادفی گوسی با میانگین صفر و واریانس یک در طول یک دوره (ξ_tها به ازای $t = ۱,۲,\dots,T$ در رابطه (۲)) تعریف می‌کنیم؛ پارامترهای سری زمانی قیمت (رابطه (۲))، به‌صورت $P_e = ۵۰$، $\lambda = ۰٦/۰$، $\sigma = ۱/۰$ و $P_0 = ۵۳$ و ثابت در نظر گرفته‌شده‌اند. به‌منظور کاوش مناسب، فرایند یادگیری در ۱۰۰۰ بازار و هر بازار به تعداد ۱۰۰۰ مرحله[1] انجام می‌شود. برای تعادل میان کاوش و بهره‌گیری، روش شبه‌حریصانه با $\varepsilon =$ ۱/۰ به‌عنوان روش انتخاب عمل در یادگیری کیو مورداستفاده قرار می‌گیرد و نرخ یادگیری $\alpha = ۵/۰$ در نظر گرفته می‌شود.

به‌منظور ارزیابی عملکرد استراتژی معاملاتی حاصل از مدل پیشنهادی، سیاست معاملاتی به‌دست‌آمده از مدل یادگیری کیو با استراتژی بازگشت به میانگین مقایسه می‌شود. استراتژی بازگشت به میانگین یکی از رایج‌ترین استراتژی‌های معاملاتی در بازارهای مالی است که بر پایه فرضیه "بازگشت به میانگین" عمل می‌کند. این فرضیه بر این اساس است که قیمت یک دارایی، مانند سهام، در طول زمان به میانگین خود بازمی‌گردد. براساس این استراتژی، ابتدا میانگین قیمت‌های قبلی سهام در طول بازه زمانی طی‌شده محاسبه می‌شود و هرگاه قیمت سهام از قیمت میانگین کمتر شد، خرید انجام می‌شود؛ اگر قیمت سهام از قیمت میانگین کمتر نشود، خریدار مجبور به خرید سهام در روزهای پایانی دوره می‌شود.

در ارزیابی، عامل هوشمند و خریداری که از استراتژی بازگشت به میانگین پیروی می‌کند در ۵۰۰۰ بازار تولیدشده از رابطه (۲)، به بهره‌گیری از استراتژی معاملاتی خود می‌پردازند. شکل (۱)، هزینه‌های عامل یادگیری کیو در رقابت با استراتژی بازگشت به میانگین را در ۱۰۰ بازار و جدول (۱)، هزینه میانگین و انحراف

[1] Episode

معیار هزینه‌های آن‌ها را در ۵۰۰۰ بازار نشان می‌دهد. شکل (۲)، نمودار قیمت سهام طی ۲۰ روز در چهار بازار مختلف و عملکرد دو استراتژی مذکور را در رقابت با یکدیگر به تصویر می‌کشد.

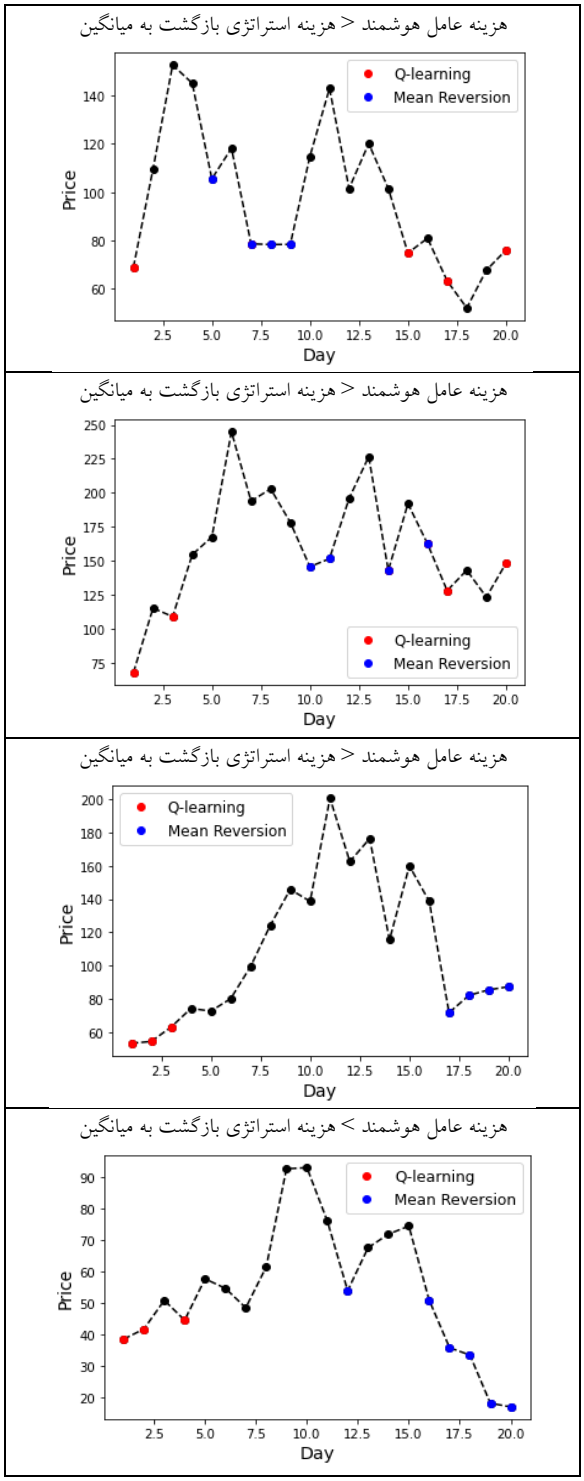


شکل (۱): نمودار هزینه‌های عامل هوشمند و استراتژی بازگشت به میانگین در ۱۰۰ بازار. منحنی قرمزرنگ هزینه‌های عامل هوشمند و منحنی آبی‌رنگ هزینه‌های حاصل از به‌کارگیری استراتژی بازگشت به میانگین را نشان می‌دهد.

جدول (۱): مقایسه استراتژی حاصل از یادگیری کیو و استراتژی بازگشت به میانگین در ۵۰۰۰ بازار

استراتژی‌ها	هزینه میانگین	انحراف معیار هزینه‌ها
یادگیری کیو	۵۷۰۶۵/۶۶	۱۵۹۲۲/۷۶
بازگشت به میانگین	۶۰۳۷۰/۵۸	۳۲۱۲۵/۴۸

بر اساس شکل (۱) و نتایج جدول (۱)، استراتژی بازگشت به میانگین انحراف معیار بسیار بالایی دارد. انحراف معیار زیاد، قابلیت اطمینان این استراتژی را کاهش می‌دهد و این استراتژی را به یک استراتژی پرخطر تبدیل می‌کند؛ با توجه به شکل (۱)، خریداری که از استراتژی بازگشت به میانگین پیروی می‌کند ممکن است هزینه بسیار زیادی متضرر شود. استفاده از سیاست حاصل از یادگیری کیو نسبت به استراتژی بازگشت به میانگین ۵/۴۷ درصد، میانگین هزینه‌ها و ۵۰/۴۴ درصد، انحراف معیار هزینه‌ها را در ۵۰۰۰ بازار کاهش می‌دهد. استراتژیِ به‌دست‌آمده توسط مدل پیشنهادی، به‌علت انحراف معیار کم نسبت به استراتژی بازگشت به میانگین، ۵۰/۴۴ درصد قابلیت اطمینان بیشتری دارد و پایدارتر است.



شکل (۲): نمودار قیمت سهام در ۲۰ روز، در چهار بازار مختلف. نقاط قرمزرنگ و نقاط آبی‌رنگ به ترتیب نقاطی را نشان می‌دهند که

عامل یادگیری کیو و خریداری که از استراتژی بازگشت به میانگین پیروی می‌کند، اقدام به خرید سهام می‌کنند.

۵- نتیجه‌گیری

در این مقاله مدلی مبتنی بر یادگیری کیو با یک ساختار MDP ساده برای به‌کارگیری یادگیری تقویتی در مسئله OEC ارائه شد. فضای حالت-عمل و سیگنال تقویتی مناسب برای مسئله مذکور تعریف شد. آزمایش‌های انجام‌شده در بازارهای مختلفِ شبیه‌سازی‌شده نشان داد که با استفاده از دانش کسب‌شده از روش پیشنهادی می‌توان استراتژی‌ای به دست آورد که در مقایسه با استراتژی بازگشت به میانگین نه‌تنها هزینه متوسط پایین‌تری دارد بلکه قابلیت اطمینان بسیار بالاتری نیز دارد. بر اساس نتایج آزمایش‌ها نتیجه می‌شود که یادگیری تقویتی می‌تواند به‌عنوان‌یک ابزار مهم برای کمک به تصمیم‌گیری در معاملات سهام مورداستفاده قرار گیرد.

در شبیه‌سازی انجام‌شده، تأثیر فعالیت‌های تجاری سایر خریداران در بازار لحاظ نشده است. پیشنهاد می‌شود که در پژوهش‌های آینده، بازاری شبیه‌سازی شود که علاوه بر عامل، چندین خریدار در آن حضورداشته باشند و با فعالیت‌های تجاری خود، قیمت سهام را تحت تأثیر قرار دهند؛ در حضور سایر خریداران، مفهوم آموزش تجلی بیش‌تری می‌یابد. همچنین پیشنهاد می‌شود که برای تطبیق سری زمانی قیمت با شرایط یک بازار دلخواه، پارامترهای سریِ زمانی با استفاده از داده‌های بازار موردنظر تنظیم گردد.

References

[۱] ولی درهمی، فریناز اعلمیان هرندی، محمدباقر دولت‌شاهی، "یادگیری تقویتی"، دانشگاه یزد، چاپ اول، ۱۳۹۶.

[2] K. Chaudhari and A. Thakkar, “Neural network systems with an integrated coefficient of variation-based feature selection for stock price and trend prediction.” Expert Systems with Applications, Vol. 219, p. 119527, 2023.

[3] Y. Zhao and G. Yang, “Deep Learning-based Integrated Framework for stock price movement prediction.” Applied Soft Computing, Vol. 133, p. 109921, 2023.

[4] A. Chudziak, “Predictability of stock returns using neural networks: Elusive in the long term.” Expert Systems with Applications, Vol. 213, p. 119203, 2023.

[5] D. Bertsimas and A. W. Lo, “Optimal control of execution costs,” Journal of Financial Markets, Vol.1, No.1, pp. 1-50, 1998.

[6] R. Almgren and N. Chriss, “Optimal Execution of Portfolio Transactions,” Journal of Risk, Vol.3, No.2, pp. 5-40, 2001.

[7] R. F. Almgren, “Optimal execution with nonlinear impact functions and trading-enhanced risk,” Applied Mathematical Finance, Vol.10, No.1, pp. 1-18, 2003.

[8] J. Lorenz and R. Almgren, “Mean–Variance Optimal Adaptive Execution,” Applied Mathematical Finance, Vol.18, No.5, pp. 395-422, 2011.

[9] G. Huberman and W. Stanzl, “Optimal Liquidity Trading,” Review of finance, Vol.9, No.2, pp. 165-200, 2005.

[10] A. A. Obizhaeva and J. Wang, “Optimal trading strategy and supply/demand dynamics,” *Journal of Financial Markets*, Vol.16, No.1, pp. 1-32, 2013.

[11] A. Schied and T. Schöneborn, “Risk aversion and the dynamics of optimal liquidation strategies in illiquid markets,” *Finance and Stochastics*, Vol.13, No.2, pp. 181-204, 2009.

[12] R. Almgren, “Optimal Trading with Stochastic Liquidity and Volatility,” SIAM Journal on Financial Mathematics, Vol.3, No.1, pp. 163-181, 2012.

[13] P. Forsyth, J. Kennedy, S. Tse, and H. Windcliff, “Optimal trade execution: A mean quadratic variation approach,” Journal of Economic Dynamics and Control, Vol.36, No.12, pp. 1971-1991, 2012.

[14] O. Guéant, “Optimal Execution and Block Trade Pricing: A General Framework,” Applied Mathematical Finance, Vol.22, No.4, pp. 336-365, 2015.

[15] Z. Liu, Y. Zhai, J. Li, G. Wang, Y. Miao, and H. Wang, “Graph Relational Reinforcement Learning for Mobile Robot Navigation in Large-Scale Crowded Environments.” IEEE Transactions on Intelligent Transportation Systems, Vol. 24, No. 8, pp. 8776-8787, 2023.

[16] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, … and D. Hassabis, “Human-level control through deep reinforcement learning,” nature, Vol.518, No.7540, pp. 529-533, 2015.

[17] B. Xian, X. Zhang, H. Zhang, and X. Gu, “Robust Adaptive Control for a Small Unmanned Helicopter Using Reinforcement Learning.” *IEEE*

Transactions on Neural Networks and Learning Systems, Vol. 33, No. 12, pp. 7589-7597, 2022.

[18] Y. D. Song, Q. Song, and W. C. Cai, "Fault-Tolerant Adaptive Control of High-Speed Trains Under Traction/Braking Failures: A Virtual Parameter-Based Approach," IEEE Transactions on Intelligent Transportation Systems, Vol.15, No.2, pp. 737-748, 2014.

[19] F. S. Melo, "Convergence of Q-learning: A simple proof," Institute of Systems and Robotics, Tech. Rep., pp. 1-4, 2001.

[20] T. Jaakkola, M. I. Jordan, and S. P. Singh, "Convergence of stochastic iterative dynamic programming algorithms," Advances in Neural Information Processing Systems, pp. 703–710, 1994.

[21] G. Ritter, "Machine Learning for Trading," SSRN Electronic Journal, 2017.

[22] J. C. H. Watkins and P. Dayan, "Q-learning," Machine learning, Vol.8, No.3, pp. 279–292, 1992.

Journal of Applied And Basic Machine Intelligence Research (ABMIR)

Applied and Basic Machine Intelligence Research

Vol. 2, No. 1, Spring & Summer 2024, pp. 92- 103

A Reinforcement Learning Approach to Determine When and How Many Stocks to Buy in Stock Trading

Fatemeh Darezereshki, Vali Derhami*

Computer Engineering Department, Yazd University, Yazd, Iran

Article Information

Original Research Paper

Received:
2024 February 9

Accepted:
2024 Mary 26

Keywords:
Stock Market, Optimizing Execution Costs of Shares, Reinforcement Learning, Q-learning

Corresponding Author*:
vderhami@yazd.ac.ir

Abstract

Due to the volatility and uncertainty inherent in the stock market, devising an optimal trading strategy is a complex endeavor. Given the non-repetitive nature of trading circumstances, learning through interactions becomes imperative. Reinforcement learning emerges as an interactive learning approach capable of adjusting system parameters based solely on a scalar efficiency signal. This paper introduces a methodology wherein the states of the system are defined by the time step, the total number of shares purchased thus far, the standard deviation of stock prices from the beginning to the current step, and the difference between the current price and the price at the previous step. By defining a suitable reinforcement signal, the paper employs one of the most popular reinforcement learning algorithms, Q-learning, to approximate state-action value functions. The stock market is simulated using a set of equations, and the proposed method is applied. Performance evaluation is conducted by comparing the proposed model against mean reversion trading strategy across 5000 simulated markets. The experimental results demonstrate that the trading strategy derived from the Q-model not only yields lower average cost but also exhibits greater reliability compared to mean reversion strategy.

: 10.22034/ABMIR.2024.21357.1050

E-ISSN: 2821-2037