Financial Research Journal

Print ISSN: 1024-8153
Online ISSN: 2423-5377

Ensemble Strategy for Algorithmic Trading Using Deep Reinforcement Learning

Meysam Amiri
Assistant Prof., Department of Finance and Banking, Faculty of Management and Accounting, Allameh Tabataba'i University, Tehran, Iran. E-mail: amiry@atu.ac.ir

Moslem Peymany Foroushany
Associate Prof., Department of Finance and Banking, Faculty of Management and Accounting, Allameh Tabataba'i University, Tehran, Iran. E-mail: m.peymany@atu.ac.ir

Hemo Boghosian*
*Corresponding Author, MSc., Department of Finance and Banking, Faculty of Management and Accounting, Allameh Tabataba'i University, Tehran, Iran. E-mail: hemoboghosian@gmail.com

Abstract

Objective

Trading strategies are crucial in investment companies as they guide decision-making processes and optimize returns. However, designing a profitable strategy within the complex and dynamic stock market environment poses significant challenges. The intricacies of market behavior and the multitude of influencing factors necessitate advanced modelling techniques. The growing availability of extensive data sets and increased computational power have facilitated the use of agent-based models, which have become essential tools for understanding economic and financial systems. The Tehran Stock Exchange often requires rapid adaptation due to severe volatility, regulatory changes, and sudden economic shifts. The choice to implement an ensemble strategy, consisting of deep reinforcement learning agents, arises from the unique challenges and opportunities of the Tehran Stock Exchange. Unlike traditional supervised learning models that make predictions solely based on historical data, agent-based models offer an adaptive approach that can respond to market changes in real-time. Another reason for selecting this strategy is its capacity to perform complex portfolio management operations. Combining multiple deep reinforcement learning agents, each with distinct strengths, the ensemble approach can leverage diverse strategies to

Citation: Amiri, Meysam; Peymany Foroushany, Moslem & Boghosian, Hemo (2026). Ensemble Strategy for Algorithmic Trading Using Deep Reinforcement Learning. *Financial Research Journal*, 28(2), 349-372. https://doi.org/10.22059/FRJ.2025.378736.1007620 (*in Persian*)

Financial Research Journal, 2026, Vol. 28, No.2, pp. 349-372
Published by University of Tehran, College of Management
https://doi.org/10.22059/FRJ.2025.378736.1007620
Article Type: Research Paper

Received: June 30, 2024
Revised: November 13, 2024
Accepted: March 15, 2025
Published online: June 20, 2026

optimize trades, manage risk, and enhance decision-making across different market conditions. Therefore, this research proposes an Ensemble strategy for algorithmic trading, leveraging deep reinforcement learning to optimize stock trading strategies that maximize returns while minimizing investment risk.

Methods

This study implements an ensemble trading strategy by modelling the stock market and employing five distinct deep reinforcement learning algorithms. This ensemble strategy synthesizes each algorithm's strengths and best features, making it adaptable to various market conditions. To achieve this, Data from stocks listed in the price index of the top 50 companies on the Tehran Stock Exchange are utilized to train and test these algorithms. The performance of the trading agent, using different reinforcement learning algorithms, is subsequently evaluated and compared against the benchmark index and a traditional minimum-variance portfolio allocation strategy. The comparative analysis helps thoroughly assess the effectiveness of the ensemble approach in real-world trading scenarios.

Results

From June 29, 2022, to January 20, 2024, the research implemented various trading models to gauge their performance. The ensemble strategy demonstrated a significant annual return of 47.13%, a cumulative return of 78.47%, and a risk-adjusted return of 1.56. These results indicate a superior performance over individual deep reinforcement learning algorithms, the benchmark price index of the 50 Tehran Stock Exchange companies, and the traditional minimum-variance portfolio allocation strategy. Among the individual algorithms, the Soft Actor-Critic (SAC) algorithm recorded the highest returns, with an annual return of 29.89% and a cumulative return of 47.89%. However, its higher annual volatility of 44.22% suggested weaker risk management. Conversely, the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm achieved a more balanced outcome with a risk-adjusted return of 0.92, highlighting its effective risk management alongside respectable returns. Therefore, the findings indicate that the ensemble strategy can effectively create a trading strategy that outperforms deep reinforcement learning algorithms, the price index of the top 50 companies on the Tehran Stock Exchange, and the minimum variance portfolio allocation strategy.

Conclusion

The Ensemble strategy offers a robust and adaptive framework for dynamic stock portfolio management by combining the strengths of multiple deep reinforcement learning algorithms. It is a reliable trading strategy that enhances returns and effectively manages investment risks. Future improvements to this strategy also involve further integrating fundamental and macroeconomic indicators to refine its predictive accuracy. Additionally, incorporating legal and regulatory constraints into the stock market modeling process, as well as considering market participants beyond investors, could improve the realism and performance of the model. This holistic approach would provide a more comprehensive understanding of market dynamics, potentially leading to more stable and robust trading outcomes.

Keywords: Algorithmic trading, Agent-based modeling, Deep reinforcement learning, Ensemble strategy.

تحقیقات مالی
شاپا چاپی: ۱۰۲۴-۸۱۵۳
شاپا الکترونیکی: ۲۴۲۳-۵۳۷۷

استراتژی گروهی برای معاملات الگوریتمی با یادگیری تقویتی عمیق

میثم امیری
استادیار، گروه مالی و بانکداری، دانشکدهٔ مدیریت و حسابداری، دانشگاه علامه طباطبائی، تهران، ایران. رایانامه: amiry@atu.ac.ir

مسلم پیمانی فروشانی
دانشیار، گروه مالی و بانکداری، دانشکده مدیریت و حسابداری، دانشگاه علامه طباطبائی، تهران، ایران. رایانامه: m.peymany@atu.ac.ir

همو بغوزیان*
* نویسنده مسئول، کارشناس ارشد، گروه مالی و بانکداری، دانشکده مدیریت و حسابداری، دانشگاه علامه طباطبائی، تهران، ایران. رایانامه: hemoboghosian@gmail.com

چکیده

هدف: استراتژی‌های معاملاتی در شرکت‌های سرمایه‌گذاری نقشی مهم دارند؛ با این حال، طراحی یک استراتژی سودآور در بازار سهامی که پیچیدگی و پویایی خاص خود را دارد، چالش‌برانگیز است. با افزایش دسترسی به داده‌ها و توان بیشتر محاسباتی، مدل‌های مبتنی بر عامل، برای درک اقتصاد و بازارهای مالی اهمیت بیشتری پیدا کرده‌اند. بازار بورس اوراق بهادار تهران، به‌دلیل نوسان‌های شدید، تغییرات قوانین نظارتی و تغییرات ناگهانی اقتصادی، اغلب به انطباق سریع نیاز دارد. انتخاب اجرای استراتژی گروهی، متشکل از عامل‌های یادگیری تقویتی عمیق، از چالش‌ها و فرصت‌های منحصربه‌فرد بورس اوراق بهادار تهران نشئت می‌گیرد. برخلاف مدل‌های یادگیری نظارت‌شده سنتی که پیش‌بینی‌ها را فقط بر اساس داده‌های تاریخی انجام می‌دهند، مدل‌های مبتنی بر عامل، یک رویکرد تطبیقی ارائه می‌کنند که می‌تواند بی‌درنگ به تغییرات بازار پاسخ دهد. یکی دیگر از دلایل انتخاب این استراتژی، ظرفیت آن برای اجرای عملیات پیچیدهٔ مدیریت پرتفوی است. با ترکیب چندین عامل یادگیری تقویتی عمیق که هر یک نقاط قوت متمایزی دارند، رویکرد گروهی می‌تواند از استراتژی‌های متنوعی برای بهینه‌سازی معاملات، مدیریت ریسک و بهبود تصمیم‌گیری در شرایط مختلف بازار استفاده کند. بنابراین، هدف از این پژوهش، پیشنهاد استراتژی گروهی برای معاملات الگوریتمی و استفاده از الگوریتم‌های یادگیری تقویتی عمیق برای معاملات سهام، به‌منظور به حداکثر رساندن بازده و کمینه‌کردن ریسک سرمایه‌گذاری است.

روش: در این پژوهش، با مدل‌سازی بازار سهام و استفاده از آموزش پنج الگوریتم یادگیری تقویتی عمیق، یعنی Advantage Actor Critic (A2C)، Deep Deterministic Policy Gradient (DDPG)، Proximal Policy Optimization (PPO)، Soft Actor-Critic (SAC) و Twin-Delayed Deep Deterministic (TD3)، یک استراتژی معاملاتی گروهی پیاده‌سازی می‌شود. این استراتژی، بهترین ویژگی‌های پنج الگوریتم را به ارث می‌برد و ادغام می‌کند؛ در نتیجه با موقعیت‌های مختلف بازار سازگاری بیشتری دارد. برای

استناد: امیری، میثم؛ پیمانی فروشانی، مسلم و بغوزیان، همو (۱۴۰۵). استراتژی گروهی برای معاملات الگوریتمی با یادگیری تقویتی عمیق. *تحقیقات مالی*، ۲۸(۲)، ۳۴۹- ۳۷۲.

تحقیقات مالی، ۱۴۰۵، دوره ۲۸، شماره ۲، صص. ۳۴۹- ۳۷۲
ناشر: دانشکدگان مدیریت دانشگاه تهران
نوع مقاله: علمی پژوهشی

تاریخ دریافت: ۱۴۰۳/۰۴/۱۰
تاریخ ویرایش: ۱۴۰۳/۰۸/۲۳
تاریخ پذیرش: ۱۴۰۳/۱۲/۲۵
تاریخ انتشار: ۱۴۰۵/۰۳/۳۰
doi: https://doi.org/10.22059/FRJ.2025.378736.1007620

آموزش و آزمایش الگوریتم‌ها، از سهام موجود در شاخص قیمتی ۵۰ شرکت بورس اوراق بهادار تهران استفاده شده است. در آخر، نتایج حاصل از معامله با استراتژی معاملاتی طراحی‌شده با الگوریتم‌های یادگیری تقویتی عمیق به‌صورت مجزا، شاخص قیمتی ۵۰ شرکت بورس اوراق بهادار تهران و استراتژی تخصیص پرتفوی حداقل واریانس مقایسه و به بحث گذاشته می‌شود.

یافته‌ها: با انجام معامله با استفاده از مدل‌های مختلف از تاریخ ۸ تیر ماه ۱۴۰۱ تا ۳۰ دی ماه ۱۴۰۲، استراتژی گروهی طراحی‌شده با بازده سالانه ۴۷/۱۳ درصد، بازده تجمعی ۷۸/۴۷ درصد، بازده تعدیل‌شده با ریسک ۱/۵۶ و حداکثر افت سرمایه ۱۸/۴۹ درصد، از الگوریتم‌های یادگیری تقویتی عمیق، شاخص قیمتی ۵۰ شرکت بورس اوراق بهادار تهران و استراتژی تخصیص پرتفوی حداقل واریانس از لحاظ بازدهی و مدیریت ریسک، عملکرد بهتری را از خود نشان داد. از بین الگوریتم‌های یادگیری تقویتی عمیق، SAC با بازده سالانه و بازده تجمعی ۲۹/۸۹ و ۸۹/۴۷ درصد از لحاظ بازدهی، بهترین عملکرد را داشت؛ اما نوسان‌های سالانه ۴۴/۲۲ درصدی آن موجب شد تا از لحاظ مدیریت ریسک عملکرد مطلوبی نداشته باشد. در مقابل، TD3 با بازده تعدیل‌شده با ریسک ۰/۹۲ از لحاظ بازدهی و مدیریت ریسک بهترین عملکرد را بین الگوریتم‌های یادگیری تقویتی عمیق داشت. بنابراین، یافته‌ها نشان می‌دهد که استراتژی گروهی می‌تواند به‌طور مؤثر یک استراتژی معاملاتی ایجاد کند که عملکردی بهتر از الگوریتم‌های یادگیری تقویتی عمیق و شاخص قیمت ۵۰ شرکت بورس اوراق بهادار تهران و استراتژی تخصیص پرتفوی حداقل واریانس از خود نشان دهد.

نتیجه‌گیری: با توجه به اینکه استراتژی گروهی بهترین ویژگی‌های هر یک از الگوریتم‌های یادگیری تقویتی عمیق را ترکیب می‌کند و به‌صورت پویا به مدیریت پرتفوی سهام می‌پردازد، می‌توان از آن به‌عنوان یک استراتژی معاملاتی قابل اتکا برای کسب بازدهی بیشتر و مدیریت ریسک سرمایه‌گذاری استفاده کرد. در پژوهش‌های آتی به‌منظور بهبود عملکرد این الگوریتم می‌توان متغیرهای بنیادی و اقتصاد کلان را نیز برای یادگیری عامل معامله‌گر به کار گرفت. همچنین در نظر گرفتن محدودیت‌های قانونی و نظارتی در مدل‌سازی بازار سهام و پیاده‌سازی عامل‌های دیگری به‌جز سرمایه‌گذاران، می‌تواند مدل را به واقعیت نزدیک‌تر کند و عملکرد آن را بهبود بخشد.

کلیدواژه‌ها: معاملات الگوریتمی، استراتژی گروهی، یادگیری تقویتی عمیق، مدل‌سازی مبتنی بر عامل.

مقدمه

اجرای استراتژی معاملاتی سودآور برای شرکت‌های سرمایه‌گذاری و نهادهای مالی امری حیاتی است. این استراتژی‌ها برای بهینه‌سازی تخصیص سرمایه و به حداکثر رساندن عملکرد سرمایه‌گذاری استفاده می‌شود. حداکثر کردن بازده را می‌توان بر اساس تخمین بازده و ریسک بالقوه میسر کرد؛ با این حال، برای تحلیلگران در نظر گرفتن همهٔ عوامل مرتبط در یک بازار پیچیده و پویا امری چالش‌برانگیز است (ژانگ و یانگ[1]، ۲۰۱۶).

تاکنون پژوهش‌های بسیاری در این خصوص انجام شده است. از نخستین مطالعات می‌توان به رویکرد سنتی پیشنهاد شده توسط مارکوئیتز اشاره کرد که ابتدا بازده مورد انتظار سهام و ماتریس کوواریانس قیمت سهام را محاسبه می‌کند؛ سپس، بهترین استراتژی تخصیص پرتفوی را با حداکثر کردن بازده برای یک نسبت ریسک معین یا به حداقل رساندن ریسک برای یک بازده از پیش تعیین‌شده به‌دست می‌آورد (مارکوئیتز[2]، ۱۹۵۲). با این حال، اجرای این رویکرد پیچیده و پُرهزینه است؛ زیرا مدیران پرتفوی ممکن است بخواهند در هر مرحله زمانی، روی تصمیم‌های خودشان تجدید نظر کنند و عوامل دیگری مانند هزینهٔ معاملات را در نظر بگیرند. روش دیگر برای معاملات سهام، مدل‌سازی آن به‌عنوان یک فرایند تصمیم‌گیری مارکوف[3] و استفاده از برنامه‌نویسی پویا[4] برای استخراج استراتژی بهینه است (نوینایر[5]، ۱۹۹۶ و ۱۹۹۷)؛ با این حال، به‌دلیل فضای حالت‌های بزرگ در مواجهه با بازار سهام مقیاس‌پذیری این مدل محدود است.

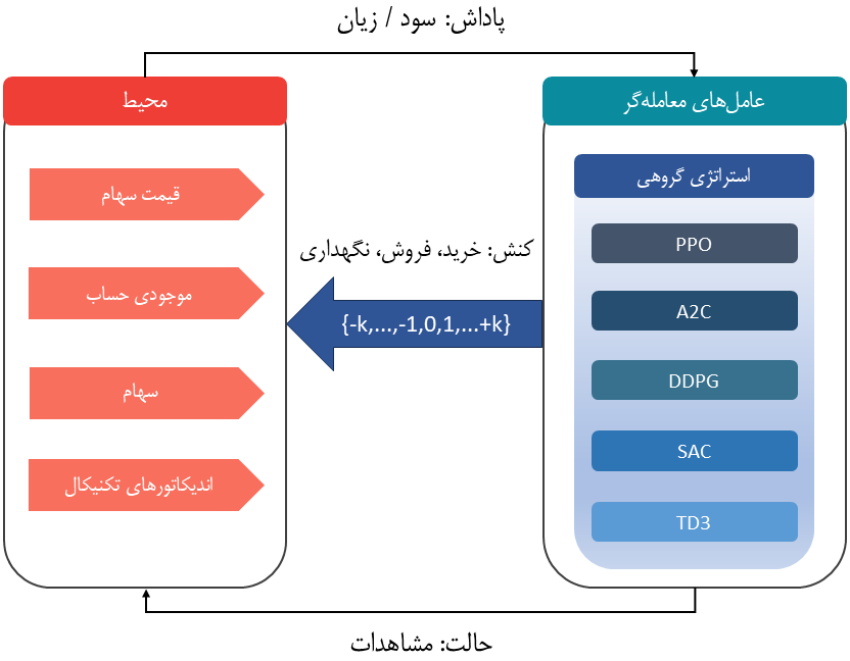
در سال‌های اخیر، الگوریتم‌های یادگیری ماشین و یادگیری عمیق، به‌طور گسترده برای ساخت مدل‌های پیش‌بینی و طبقه‌بندی در بازار مالی استفاده شده‌اند. داده‌های روزانهٔ سهام به همراه داده‌های دیگری مانند شاخص‌های بنیادی مستخرج از گزارش‌های دوره‌ای و اخبار بازار با الگوریتم‌های یادگیری ماشین ترکیب می‌شوند تا به بازده‌های سرمایه‌گذاری بیشتری دست پیدا کنند (ژانگ و اسکینا[6]، ۲۰۱۰). این رویکردها به‌جای تخصیص موقعیت‌های معاملاتی بین سهام انتخاب‌شده، تنها بر انتخاب سهام با عملکرد بالا متمرکزند؛ به عبارت دیگر، مدل‌های یادگیری ماشین برای مدل‌سازی موقعیت‌های معاملاتی آموزش داده نشده‌اند.

در این پژوهش، یک استراتژی گروهی[7] پیشنهاد می‌شود که پنج الگوریتم یادگیری تقویتی عمیق را ترکیب و استراتژی معاملاتی بهینه را در یک بازار سهام پیچیده و پویا پیدا می‌کند. پنج الگوریتم کنشگر ـ منتقد[8] (کندا و سیتسیکلیس[9]، ۲۰۰۱) استفاده‌شده عبارت‌اند از: Proximal Policy Optimization (PPO) (شولمن، ولسکی، داریوال،

1. Zhang & Yang
2. Markowitz
3. Markov Decision Process (MDP)
4. Dynamic Programming (DP)
5. Neuneier
6. Zhang & Skiena
7. Ensemble strategy
8. Actor-Critic
9. Konda & Tsitsiklis

رادفورد و کلیموف[1]، ۲۰۱۷)، Advantage Actor Critic (A2C) (سیلور و همکاران[2]، ۲۰۱۶)، Deep Deterministic Policy Gradient (DDPG) (سیلور و همکاران، ۲۰۱۶)، Soft Actor-Critic (SAC) (پاچکو آزنار[3]، ۲۰۲۳) و Twin Delayed DDPG (TD3) (یو[4]، ۲۰۲۳).

در شکل ۱، رویکرد یادگیری تقویتی عمیق طراحی‌شده ارائه شده است. این رویکرد از دو عنصر اصلی محیط[5] و عامل‌های معامله‌گر[6] تشکیل شده است که هر یک در تسهیل تصمیمات معاملاتی تطبیقی و بهینه نقش مهمی دارد.



شکل ۱. طرح کلی استراتژی گروهی مبتنی بر یادگیری تقویتی عمیق

در ادامه به توضیح هر یک از عناصر داخل شکل پرداخته می‌شود.

۱. محیط: این عنصر، بازار سهام را شبیه‌سازی می‌کند و مسئولیت ارائه داده‌ها و حالت‌ها[7] به عامل‌های معامله‌گر را برعهده دارد. محیط شامل قیمت و تعداد سهام، موجودی حساب و اندیکاتورهای تکنیکال است. محیط در نتیجهٔ کنش عامل‌های معاملاتی، مشاهداتی را جهت نمایش حالت فعلی بازار به آن‌ها ارائه می‌دهد.

۲. عامل‌های معاملاتی: هدف عامل‌های معاملاتی این است که با تطبیق پویای استراتژی خود با شرایط مختلف بازار، سود خود را به حداکثر برسانند. هر عامل معامله‌گر از یکی از الگوریتم‌های یادگیری تقویتی عمیق یعنی

1. Schulman, Wolski, Dhariwal, Radford & Klimov
2. Silver et al.
3. Pacheco Aznar
4. Yu
5. Environment
6. Trading Agents
7. States

A2C، DDPG، PPO، SAC و TD3 استفاده می‌کند و استراتژی گروهی از نقاط قوت منحصربه‌فرد هر یک در شرایط مختلف بازار بهره می‌جوید.

۳. کنش‌ها[1]: عامل می‌تواند سه کنش انجام دهد: خرید، نگهداری و فروش سهام. بر اساس وضعیت بازار، عامل تصمیم می‌گیرد که سهام بیشتری بخرد، سهام در دست خود را کاهش دهد یا موقعیت خود را حفظ کند.

۴. تابع پاداش[2]: تابع پاداش، موفقیت هر کنش را بر اساس سود یا زیان متحمل‌شده اندازه‌گیری می‌کند و به تصمیماتی که در طور زمان به بازده بالاتر منجر می‌شود، انگیزه می‌دهد.

استراتژی طراحی‌شده می‌تواند با موقعیت‌های مختلف بازار تطبیق داده شود و بازده را با توجه به محدودیت ریسک به حداکثر برساند. برای این کار، ابتدا یک محیط ساخته و فضای کنش[3]، فضای حالت[4] و تابع پاداش تعریف می‌شود؛ سپس، پنج الگوریتم یاد شده آموزش داده می‌شوند که کنش‌هایی را در محیط طراحی‌شده انجام دهند. در آخر، پنج عامل آموزش داده‌شده با استفاده از بازده تعدیل‌شده با ریسک با هم ترکیب می‌شوند. اثربخشی استراتژی گروهی با معیارهای عملکرد بهتر، نسبت به استراتژی تخصیص پرتفوی حداقل واریانس، شاخص قیمتی ۵۰ شرکت بورس اوراق بهادار تهران و ۵ رویکرد یادگیری تقویتی عمیق به‌صورت مجزا بررسی می‌شود.

پیشینهٔ پژوهش

پژوهش‌های اخیر در زمینه یادگیری تقویتی عمیق در بازارهای مالی، فضای حالت و کنش را به‌صورت گسسته یا پیوسته تعریف کرده‌اند و یکی از این رویکردهای یادگیری فقط منتقد[5]، فقط کنشگر[6] و کنشگر ـ منتقد را به‌کار گرفته‌اند (فیشر[7]، ۲۰۱۸).

رویکرد یادگیری فقط منتقد که رایج‌ترین آن‌هاست، یک مسئله فضای کنش گسسته را با استفاده از الگوریتم‌هایی مانند یادگیری عمیق کیو[8] حل می‌کند و یک عامل را روی یک سهم یا دارایی آموزش می‌دهد (جئونگ و کیم[9]، ۲۰۱۹). ایدهٔ رویکرد رویکرد فقط منتقد، استفاده از یک تابع ارزش ـ کیو[10] برای یادگیری خطمشی[11] انتخاب کنش بهینه است که پاداش مورد انتظار آینده را با توجه به حالت فعلی به حداکثر می‌رساند. به‌جای محاسبه جدول ارزش حالت ـ ارزش[12]، خطای بین ارزش ـ کیو تخمینی و ارزش ـ کیو هدف را در یک تابع انتقال به حداقل می‌رساند و از یک شبکه عصبی برای انجام تقریب تابع

1. Actions
2. Reward function
3. Action Space
4. State space
5. Critic-only
6. Actor-only
7. Fischer
8. Deep Q-Learning (DQN)
9. Jeong & Kim
10. Q-value
11. Policy
12. State-action

استفاده می‌کند. محدودیت اصلی رویکرد فقط منتقد این است که تنها با فضاهای حالت و کنش گسسته و محدود کار می‌کند که برای سبد بزرگی از سهام عملی نیست؛ زیرا قیمت‌ها پیوسته هستند (چن و گائو[1]، ۲۰۱۹).

ایده رویکرد فقط کنشگر این است که عامل، به‌طور مستقیم خودش خط‌مشی بهینه را یاد می‌گیرد و به‌جای داشتن یک شبکهٔ عصبی برای یادگیری ارزش ـ کیو، یک شبکه عصبی خط‌مشی را یاد می‌گیرد (دنگ، بائو، کنگ، رن و دای[2]، ۲۰۱۶). خط‌مشی یک توزیع احتمال است و می‌توان آن را یک استراتژی برای یک حالت معین (احتمال، انجام یک کنش مجاز) تعریف کرد. رویکرد فقط بازیگر می‌تواند محیط‌های فضای کنش پیوسته را مدیریت کند (ژیانگ و لیانگ[3]، ۲۰۱۷).

رویکرد کنشگر ـ منتقد اخیراً در کانون توجه پژوهشگران مالی قرار گرفته است (لی، رائو و شی[4]، ۲۰۱۸). ایدهٔ این رویکرد آن است که شبکهٔ کنشگر که نمایانگر خط‌مشی است و شبکه منتقد که تابع ارزش را نشان می‌دهد، به‌طور هم‌زمان به‌روز شوند. منتقد تابع ارزش را تخمین می‌زند، در حالی که کنشگر توزیع احتمال سیاست را که توسط منتقد با گرادیان‌های خط‌مشی هدایت می‌شود، به‌روز می‌کند. با گذشت زمان، کنشگر یاد می‌گیرد که کنش‌های بهتری انجام دهد و منتقد در ارزیابی آن کنش‌ها بهتر می‌شود. رویکرد کنشگر ـ منتقد ثابت کرده است که می‌تواند با محیط‌های بزرگ و پیچیده سازگار شود. بنابراین، رویکرد کنشگر ـ منتقد در معاملات با پرتفوی سهام بزرگ می‌تواند کارساز واقع شود (ژانگ، زوهرن و رابرتس[5]، ۲۰۱۹).

در سال‌های اخیر، علاقهٔ فزاینده‌ای به استفاده از رویکردهای هوش مصنوعی در بازار سرمایه ایران به‌وجود آمده است و مطالعات متعددی برای بررسی پتانسیل آن، به‌منظور بهبود پیش‌بینی سهام و استراتژی‌های معاملاتی منتشر شده است. محققان به‌دنبال استفاده از هوش مصنوعی برای بینش دقیق‌تر در خصوص بورس اوراق بهادار تهران، به‌ویژه از طریق یادگیری ماشین و یادگیری عمیق هستند.

در این راستا، حیدری و امیری (۱۴۰۱) به بررسی مدل‌های مختلف هوش مصنوعی برای پیش‌بینی روند قیمت سهام در بورس اوراق بهادار تهران پرداختند. آن‌ها با استفاده از داده‌های ۱۵۰ شرکت پذیرفته‌شده در بورس اوراق بهادار تهران، نشان دادند در حالی که مدل‌های یادگیری عمیق نسبت به روش‌های سنتی، بهبودهایی را در دقت پیش‌بینی ارائه می‌دهند، نتایج از نظر ارزش عملی برای سرمایه‌گذاران محدود باقی می‌ماند.

محبی، فدایی نژاد، اصولیان و حمیدی‌زاده (۱۴۰۱) با به‌کارگیری تکنیک‌های کاهش ابعاد[6]، دقت پیش‌بینی‌های روزانهٔ شاخص بورس اوراق بهادار تهران را بهبود دادند. آن‌ها با استفاده از الگوریتم‌های تفاضل اطلاعات متقابل[7] و

1. Chen & Gao
2. Deng, Bao, Kong, Ren & Dai
3. Jiang & Liang
4. Li, Rao & Shi
5. Zhang, Zohren & Roberts
6. Dimensionality Reduction
7. Mutual Information Difference (MID)

تجزیه‌وتحلیل مؤلفه‌های اصلی[1] برای شناسایی ویژگی‌های اقتصادی کلیدی، دقت مدل‌های مبتنی بر شبکهٔ عصبی را برای پیش‌بینی شاخص سهام، به‌ویژه از طریق مدل تابع پایه شعاعی[2] با استفاده از الگوریتم MID، بهبود بخشیدند.

نوراحمدی و نوراحمدی (۱۴۰۲) با هدف تخمین نسبت پوشش ریسک پویا برای ارائهٔ سیگنال‌های معامله در صنعت خودرو، در بازار بورس اوراق بهادار تهران، فیلتر کالمن را روی معاملات جفتی اعمال کردند. روش آن‌ها نسبت پوشش ریسک پویا را برای تولید سیگنال‌های معامله، بر اساس انحراف‌های اسپرد قیمت بین جفت سهام محاسبه کرد. آن‌ها در این پژوهش استفاده از فیلتر کالمن را به‌عنوان یک استراتژی سودآور برای معاملات جفتی معرفی کردند.

در نهایت، نوراحمدی، رحیمی و صادقی (۱۴۰۳) یک سیستم توصیه‌گر[3] سهام را برای بورس اوراق بهادار تهران بر اساس فیلتر مشارکتی پیشنهاد کردند. این مدل با استفاده از داده‌های تاریخی قیمت ۱۴۵ شرکت پذیرفته‌شده در بورس اوراق بهادار، سهام مناسب را با تجزیه‌وتحلیل الگوهای پنهان بازار شناسایی کرد و به هدف آن که کمک به سرمایه‌گذاران برای دستیابی به بازدهی بالاتر از شاخص بازار و ارزش در معرض خطر کمتر از آن بود، دست یافت.

پژوهش حاضر، یک سیستم معاملاتی پیشرفته را معرفی می‌کند که به‌طور فعال سبد سهام را مدیریت می‌کند و از روش‌های یادگیری ماشین سنتی پیشی می‌گیرد. برخلاف پژوهش‌های گذشته که بر پیش‌بینی قیمت‌ها، روندها و توصیه سهام تکی تمرکز داشتند، مدل استفاده‌شده در این پژوهش از مجموعه‌ای از عامل‌های یادگیری تقویتی عمیق استفاده می‌کند.

پژوهش‌های گذشته هنوز کاربرد الگوریتم‌های یادگیری تقویتی عمیق یا به‌طور کلی رویکردهای یادگیری تقویتی را برای معاملات الگوریتمی در بازار سرمایه ایران بررسی نکرده‌اند. برخلاف مدل‌های سنتی که بر پیش‌بینی‌های ایستا از داده‌های تاریخی متکی هستند، این سیستم به‌طور مداوم تکامل می‌یابد و استراتژی خود را برای بهینه‌سازی بازده و مدیریت مؤثر ریسک اصلاح می‌کند.

مدل معرفی‌شده با مدیریت کل یک سبد به‌جای تمرکز بر سهام منفرد، ابزاری عملی و قابل انطباق برای معامله به سرمایه‌گذاران ارائه می‌دهد. این رویکرد استاندارد جدیدی را برای معاملات مبتنی بر هوش مصنوعی در ایران ایجاد می‌کند و راه‌حلی قدرتمند برای مدیریت فعال پرفتوی ارائه می‌دهد.

روش‌شناسی پژوهش

مدل فرایند تصمیم‌گیری مارکوف برای معاملات سهام

در این پژوهش، معاملات سهام به‌عنوان یک فرایند تصمیم‌گیری مارکوف مدل می‌شود (یانگ، لیو، ژونگ و ولید[4]، ۲۰۲۰). برای مدل‌سازی ماهیت تصادفی بازار سهام، از فرایند تصمیم‌گیری مارکوف به شرح زیر استفاده شده است.

1. Principal Component Analysis (PCA)
2. Radial Basis function (RBF)
3. Recommender system
4. Yang, Liu, Zhong & Walid

- حالت $s = [p, h, b]$: برداری که قیمت سهام $p \in \mathbb{R}_+^D$، تعداد سهام $h \in \mathbb{Z}_+^D$ و موجودی حساب باقی‌مانده $b \in \mathbb{R}_+$ را شامل می‌شود. در اینجا D تعداد شرکت‌ها و $\mathbb{Z}_+$ اعداد صحیح غیرمنفی را نشان می‌دهد.
- کنش a: برداری از کنش‌ها در قبال D تعداد شرکت. کنش‌های مجاز برای سهم هر شرکت شامل خرید، فروش و نگهداری می‌شود که به‌ترتیب باعث افزایش، کاهش و ثابت‌ماندن تعداد سهام h می‌شود.
- پاداش $r(s, a, s')$: پاداش مستقیم انجام کنش a در حالت s و رفتن به حالت جدید s'.
- خط‌مشی $\pi(s)$: استراتژی معاملاتی در حالت s که توزیع احتمال کنش‌ها در حالت s است.
- ارزش ـ کیو $Q_\pi(s, a)$: پاداش مورد انتظار از انجام کنش a در حالت s با پیروی از خط‌مشی π.

در هر حالت، یکی از سه کنش ممکن روی سهم شرکت d $(d = 1, \dots, D)$ در پرتفوی سهام انجام می‌شود. به‌صورتی که:

- نتیجهٔ فروش تعداد $k[d] \in [1, h[d]]$ سهم: $h_{t+1}[d] = h_t[d] - k[d]$ به‌صورتی که $k[d] \in \mathbb{Z}_+$ و $d = 1, \dots, D$.
- نتیجهٔ نگهداری: $h_{t+1}[d] = h_t[d]$.
- نتیجهٔ خرید تعداد $k[d]$ سهم: $h_{t+1}[d] = h_t[d] + k[d]$.

در زمان t کنشی انجام می‌شود و قیمت سهام در $t + 1$ به‌روز می‌شود؛ بر همین اساس، ارزش پرتفوی ممکن است به سه حالت دیگر تغییر کند. شایان ذکر است که ارزش پرتفوی، $p^T h + b$ است و در صورت انجام کنش نگهداری نیز با توجه به تغییر قیمت سهام، ممکن است باعث تغییر ارزش پرتفوی شود.

محدودیت‌های معاملات سهام

در ادامه، به محدودیت‌ها و مفروض‌های در نظر گرفته‌شده در مدل‌سازی معاملات سهام اشاره می‌شود.

- نقدینگی بازار: سفارش‌ها به‌سرعت با قیمت پایانی اجرا می‌شوند. فرض می‌شود که بازار سهام تحت تأثیر عامل معاملاتی طراحی‌شده قرار نمی‌گیرد.
- موجودی حساب غیرمنفی $b \geq 0$: کنش‌های مجاز نباید به موجودی حساب منفی منجر شوند. بر اساس کنش انجام‌شده در زمان t، سهام به مجموعه‌هایی برای فروش $\mathcal{S}$ ، خرید $\mathcal{B}$ و نگهداری $\mathcal{H}$ تقسیم می‌شوند؛ به‌صورتی که $\mathcal{S} \cup \mathcal{B} \cup \mathcal{H} = \{1, \dots, D\}$ و با یکدیگر هم‌پوشانی ندارند. با فرض اینکه $p_t^B = [p_t^i : i \in \mathcal{B}]$ و $k_t^B = [k_t^i : i \in \mathcal{B}]$ بردار قیمت و تعداد سهام خریداری‌شده در سهام مجموعه خرید باشند، می‌توان p_t^S و k_t^S را برای سهام مجموعه فروش و p_t^H و k_t^H را برای سهام مجموعه نگهداری تعریف کرد. بنابراین، محدودیت برای موجودی حساب غیرمنفی را می‌توان به‌صورت زیر بیان کرد:

رابطهٔ ۱) $$b_{t+1} = b_t + (p_t^S)^T k_t^S - (p_t^B)^T k_t^B \geq 0$$

- هزینه معاملات: انواع مختلفی از هزینه‌های معاملاتی مانند کارمزد کارگزاری، کارمزد شرکت بورس، حق

نظارت سازمان، کارمزد شرکت سپرده‌گذاری و کارمزد شرکت فناوری بورس و مالیات وجود دارد. کارگزاران مختلف کارمزد متفاوتی دارند. با وجود این تفاوت در کارمزدها، در پژوهش حاضر، هزینه‌های معاملاتی بر اساس مجموع آخرین هزینه‌های معاملاتی برای خرید و فروش فرض می‌شود؛ به‌صورتی که برای خرید ۰/۳۷۱۲درصد و برای فروش ۰/۸۸ درصد ارزش هر معامله در نظر گرفته شده است.

- ریسک‌گریزی برای سقوط بازار: رویدادهایی ناگهانی‌ای وجود دارد که ممکن است باعث سقوط بازار سهام شود، مانند جنگ، فروپاشی حباب‌های بورس، نکول بدهی‌های دولتی و بحران مالی. برای کنترل ریسک در بدترین سناریو مانند بحران مالی، از شاخص تلاطم مالی[1] $turbulence_t$ استفاده می‌شود که حرکت شدید قیمت دارایی‌ها را اندازه‌گیری می‌کند و از طریق رابطهٔ ۲ محاسبه می‌شود (کریتزمن و لی[2]، ۲۰۱۰).

رابطهٔ ۲) $$turbulence_t = (y_t - \mu)\Sigma^{-1}(y_t - \mu)' \in \mathbb{R}$$

$y_t \in \mathbb{R}^D$ بازده سهام در دورهٔ کنونی t، $\mu \in \mathbb{R}^D$ میانگین بازده‌های تاریخی و $\Sigma \in \mathbb{R}^{D\times D}$ کوواریانس بازده‌های تاریخ است.

وقتی $turbulence_t$ بالاتر از یک آستانه تاریخی باشد که نشان‌دهنده شرایط شدید[3] بازار است، خرید متوقف می‌شود و عامل معامله‌گر، همهٔ سهام خود را می‌فروشد و زمانی که این شاخص به زیر آستانه بازگشت، معامله از سر گرفته می‌شود.

حداکثرسازی بازده به‌عنوان هدف معاملاتی

منظور از تابع پاداش، تغییر ارزش پرتفوی در زمانی است که کنش a در حالت s انجام می‌شود و به حالت جدید s' می‌رسد. هدف تابع پاداش زیر، طراحی یک استراتژی معاملاتی است که تغییر ارزش پرتفوی را به حداکثر برساند.

رابطهٔ ۳) $$r(s_t, a_t, s_{t+1}) = (b_{t+1} + p_{t+1}^T h_{t+1}) - (b_t + p_t^T h_t) - c_t$$

که در آن عبارت‌های اول و دوم، به‌ترتیب ارزش پرتفوی را در $t+1$ و t نشان می‌دهند. برای تجزیه بیشتر بازده، تابع انتقال سهام h_t به‌صورت زیر تعریف می‌شود.

رابطهٔ ۴) $$h_{t+1} = h_t - k_t^S + k_t^B$$

همچنین با توجه به رابطهٔ ۱ برای موجودی حساب b_t، رابطه ۳ را می‌توان به‌صورت زیر بازنویسی کرد.

رابطهٔ ۵) $$r(s_t, a_t, s_{t+1}) = r_H - r_S + r_B - c_t$$

که در آن داریم؛

1. Financial turbulence index
2. Kritzman & Li
3. Extreme

$$r_H = (p_{t+1}^H - p_t^H)^T h_t^H$$ رابطهٔ ۶)

$$r_S = (p_{t+1}^S - p_t^S)^T h_t^S$$ رابطهٔ ۷)

$$r_B = (p_{t+1}^B - p_t^B)^T h_t^B$$ رابطهٔ ۸)

r_H، r_S و r_B نشان‌دهندهٔ تغییر ارزش پرتفوی از زمان t به $t+1$ است که به‌ترتیب از نگهداری، فروش و خرید سهام ناشی می‌شود.

شاخص تلاطم مالی $turbulence_t$ همراه با تابع پاداش برای رسیدگی به ریسک‌گریزی معامله‌گران برای سقوط بازار ترکیب شده است. هنگامی که شاخص تلاطم در رابطه ۲ از یک آستانه تاریخی بالاتر می‌رود، رابطهٔ ۷ به معادله زیر تبدیل می‌شود.

$$r_{sell} = (p_{t+1} - p_t)^T k_t$$ رابطهٔ ۹)

که نشان می‌دهد عامل می‌خواهد با فروش تمام سهام نگهداری‌شده، تغییر منفی ارزش پرتفوی را به حداقل برساند؛ زیرا قیمت تمام سهام سقوط خواهد کرد.

مقداردهی اولیه بدین صورت است که p_0 روی قیمت سهام در زمان صفر تنظیم می‌شود. h و $Q_\pi(s,a)$ صفر هستند و $\pi(s)$ به‌طور یکنواخت بین تمام کنش‌ها برای هر حالت توزیع می‌شود. سپس، $Q_\pi(s_t,a_t)$ از طریق تعامل با محیط بازار سهام به‌روز می‌شود. استراتژی بهینه توسط معادله بلمن[1] ارائه می‌شود؛ به این ترتیب که پاداش مورد انتظار انجام عمل a_t در حالت s_t، جمع پاداش مورد انتظار مستقیم $r(s_t,a_t,s_{t+1})$ و پاداش آینده در حالت s_{t+1} است. در صورتی که پاداش‌های آینده با ضریب $0 < \gamma < 1$ با هدف همگرایی تنزیل شود، معادله زیر به‌دست می‌آید.

$$Q_\pi(s_t,a_t) = \mathbb{E}_{s_{t+1}}\left[r(s_t,a_t,s_{t+1}) + \gamma \mathbb{E}_{a_{t+1}\sim\pi(s_{t+1})}[Q_\pi(s_{t+1},a_{t+1})]\right]$$ رابطهٔ ۱۰)

هدف، طراحی یک استراتژی معاملاتی است که تغییر تجمعی مثبت ارزش پرتفوی $r(s_t,a_t,s_{t+1})$ را به حداکثر می‌رساند و از روش یادگیری تقویتی عمیق برای حل این مسئله استفاده می‌شود.

محیط بازار سهام

قبل از آموزش یک عامل معاملاتی تقویتی عمیق، محیطی برای شبیه‌سازی معاملات سهام دنیای واقعی ایجاد می‌شود که به عامل اجازه می‌دهد تعامل و یادگیری را انجام دهد. در معاملات عملی، اطلاعات مختلفی باید در نظر گرفته شود؛ برای مثال، قیمت‌های تاریخی سهام، سهام فعلی، اندیکاتورهای تکنیکال[2] و... . عامل معاملاتی باید چنین اطلاعاتی را از طریق محیط به‌دست آورد و کنش‌های تعریف‌شده در قسمت قبل را انجام دهد (بروکمن و همکاران[3]، ۲۰۱۶).

1. Bellman Equation
2. Technical indicators
3. Brockman et al.

محیطی شامل چند سهم

در این پژوهش، از یک فضای کنش پیوسته برای مدل‌سازی معاملات چند سهمی استفاده شده است. فرض می‌شود که بازار سهام در مجموع ۴۹ شرکت (شرکت‌های موجود در شاخص قیمتی ۵۰ شرکت بورس اوراق بهادار تهران) دارد.

فضای حالت

از یک بردار ۴۹۱ بُعدی متشکل از هفت بخش اطلاعات برای نشان‌دادن فضای حالت محیط معاملات چند سهمی $[b_t, p_t, h_t, M_t, R_t, C_t, X_t, U_t, L_t, T_t, S_t]$ استفاده شده است که هر جزء به‌صورت زیر تعریف می‌شود:

- $b_t \in \mathbb{R}_+$: موجودی در دسترس در مرحله زمانی کنونی t.
- $p_t \in \mathbb{R}_+^{49}$: قیمت پایانی تعدیل‌شده هر سهم.
- $h_t \in \mathbb{Z}_+^{49}$: تعداد سهام متعلق به عامل.
- $M_t \in \mathbb{R}^{49}$: میانگین متحرک همگرا/واگرا[1] یکی از رایج‌ترین اندیکاتورهای حرکتی است که میانگین‌های متحرک را شناسایی می‌کند (چونگ، نگ و لیو[2]، ۲۰۱۴).
- $R_t \in \mathbb{R}_+^{49}$: شاخص قدرت نسبی[3] میزان تغییرات اخیر قیمت را تعیین می‌کند. اگر قیمت در اطراف خط حمایت حرکت کند، نشان‌دهندهٔ فروش بیش از حد سهام است و می‌توان آن را خرید. اگر قیمت اطراف خط مقاومت حرکت کند، نشان می‌دهد که سهام بیش از حد خرید شده است و می‌توان آن را فروخت (چونگ و همکاران، ۲۰۱۴).
- $C_t \in \mathbb{R}_+^{49}$: شاخص کانال کالا[4] قیمت فعلی را با قیمت متوسط در یک بازه زمانی مقایسه می‌کند تا سیگنال خرید یا فروش را بدهد (میتا، پروچازکا، چرماک و شردل[5]، ۲۰۱۶).
- $X_t \in \mathbb{R}^{49}$: شاخص میانگین حرکتی[6] قدرت روند را با کمی‌کردن مقدار حرکت قیمت شناسایی می‌کند (گوریب[7]، ۲۰۱۸).
- $U_t \in \mathbb{R}_+^{49}$ و $L_t \in \mathbb{R}_+^{49}$: باندهای بولینجر[8] به سنجش نوسان‌های سهام کمک می‌کند تا تعیین شود که آیا ارزش آن‌ها بیش از قیمت آن‌هاست یا خیر (لائوگیسو و همکاران[9]، ۲۰۱۹).
- $T_t \in \mathbb{R}_+^{49}$ و $S_t \in \mathbb{R}_+^{49}$: میانگین متحرک ساده[10] ۳۰ روزه و ۶۰ روزه با استفاده از قیمت‌های پایانی محاسبه

1. Moving Average Convergence/Divergence (MACD)
2. Chong, NG & Lew
3. Relative Strength Index (RSI)
4. Commodity Channel Index (CCI)
5. Maitah, Procházka, Čermák & Šrédl
6. Average Directional Index (ADX)
7. Gurrib
8. Bollinger Bands (BB)
9. Lauguico et al.
10. Simple Moving Average (SMA)

شده‌اند. یک میانگین متحرک ساده، میانگین یک محدوده انتخاب‌شده از قیمت‌ها، معمولاً قیمت‌های پایانی را با تعداد دوره‌های آن محدوده محاسبه می‌کند (کرگ و پاربری[1]، ۲۰۰۵).

فضای کنش

برای یک سهم، فضای کنش به‌صورت $\{-k, \dots, -1, 0, 1, \dots, k\}$ تعریف شده است؛ به‌گونه‌ای که k و $-k$ به‌عنوان تعداد سهامی که می‌توان خرید و فروخت تعیین می‌شود و $k \leq h_{max}$ که در آن h_{max} (متغیر از پیش تعیین‌شده) گویای حداکثر مقدار سهامی است که می‌توان خرید. بنابراین اندازه کل فضای کنش [49] $(2k + 1)$ محاسبه می‌شود. در آخر فضای کنش به بازه $[-1, +1]$ نرمال می‌شود؛ زیرا الگوریتم‌های یادگیری تقویتی A2C و PPO خط‌مشی را مستقیماً روی یک توزیع گاوسی تعریف می‌کنند که باید نرمال و متقارن شود (هیل و همکاران[2]، ۲۰۱۸).

معامله مبتنی بر عامل با یادگیری تقویتی عمیق

همان طور که گفته شد، در این پژوهش از پنج الگوریتم مبتنی بر عامل کنشگر ـ منتقد برای پیاده‌سازی عامل معامله‌گر استفاده می‌شود. این پنج الگوریتم به‌ترتیب عبارت‌اند از: A2C، DDPG، PPO، TD3 و SAC. استراتژی گروهی با ترکیب این پنج عامل با یکدیگر برای ایجاد یک استراتژی معاملاتی با کارایی بیشتر پیاده‌سازی شده است.

Advantage Actor Critic (A2C)

الگوریتم A2C (سیلور و همکاران، ۲۰۱۶) یک الگوریتم کنشگر ـ منتقد است که برای بهبود به‌روزرسانی‌های گرادیان خط‌مشی معرفی شده است. از یک تابع مزیت[3] برای کاهش واریانس گرادیان خط‌مشی استفاده می‌کند. به‌روزرسانی گرادیان هماهنگ استفاده‌شده در A2C مقرون‌به‌صرفه‌تر و سریع‌تر است و با اندازه‌های برزگ دسته داده بهتر عمل می‌کند. این الگوریتم، به‌دلیل ثبات آن، یک مدل عالی برای معاملات سهام محسوب می‌شود. تابع هدف برای A2C به‌صورت زیر تعریف می‌شود.

$$\nabla J_\theta(\theta) = \mathbb{E}\left[\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\, A(s_t, a_t)\right] \qquad \text{رابطهٔ ۱۱)}$$

که در آن $\pi_\theta(a_t|s_t)$ شبکه خط‌مشی و $A(s_t, a_t)$ تابع مزیت است.

Deep Deterministic Policy Gradient (DDPG)

الگوریتم DDPG (سیلور و همکاران، ۲۰۱۵) چارچوب‌های یادگیری کیو (ساتون و بارتو[4]، ۱۹۹۸) و گرادیان خط‌مشی (ساتون، مک‌آلستر، سینگ و منصور[5]، ۲۰۰۰) را ترکیب می‌کند و از شبکه‌های عصبی به‌عنوان تقریب‌زننده تابع استفاده

1. Craig & Parbery
2. Hill et al.
3. Advantage function
4. Sutton & Barto
5. Sutton, Mcallester, Singh & Mansour

می‌کند. برخلاف یادگیری عمیق کیو که به‌طور غیرمستقیم از طریق جداول ارزش ـ کیو یاد می‌گیرد و دچار مسئله ابعاد است (بوسونیو، دی بروین، تولیچ، کوبر و پالونکو[1]، ۲۰۱۸)، این الگوریتم به‌طور مستقیم از مشاهدات خود و از طریق گرادیان خط‌مشی یاد می‌گیرد.

در هر مرحله زمانی، عامل DDPG در حالت s_t کنش a_t را انجام می‌دهد، پاداش r_t را می‌گیرد و به حالت s_{t+1} می‌رود. دسته‌ای از N انتقال از R گرفته می‌شود و ارزش ـ کیو y_i به‌صورت زیر به‌روزرسانی می‌شود.

$$y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'}, \theta^{Q'}), i = 1, \dots, N \quad \text{رابطهٔ ۱۲)}$$

سپس شبکه منتقد با به حداقل رساندن تابع زیان[2] $L(\theta^Q)$ که تفاوت مورد انتظار بین خروجی‌های شبکه منتقد هدف Q' و شبکه منتقد Q است، به‌روز می‌شود.

$$L(\theta^Q) = \mathbb{E}_{s_t, a_t, r_t, s_{t+1} \sim buffer}[(y_i - Q(s_t, a_t|\theta^Q))^2] \quad \text{رابطهٔ ۱۳)}$$

DDPG در مدیریت فضای کنش پیوسته مؤثر است؛ بنابراین برای معاملات سهام الگوریتمی مناسب تلقی می‌شود.

Proximal Policy Optimization (PPO)

در این پژوهش از PPO به‌عنوان یک جزء در استراتژی گروهی بررسی و استفاده می‌شود. PPO برای کنترل به‌روزرسانی گرادیان خط‌مشی و اطمینان از اینکه خط‌مشی جدید خیلی متفاوت از سیاست قبلی نیست، معرفی شده است. این الگوریتم سعی می‌کند هدف بهینه‌سازی خط‌مشی منطقه اعتماد[3] را با معرفی یک عبارت برش[4] به تابع هدف ساده کند (شولمن و همکاران، ۲۰۱۷).

دلیل انتخاب این الگوریتم برای معاملات سهام پیاده‌سازی و تنظیم سریع، ساده‌تر و پایدار آن است.

Soft Actor-Critic (SAC)

الگوریتم SAC (هارنوجا، ژو، ابیل و لوین[5]، ۲۰۱۸) یک الگوریتم خارج از خط‌مشی[6] است که برای یادگیری تقویتی حداکثر آنتروپی[7] توسعه یافته است. SAC در مقایسه با DDPG از خط‌مشی تصادفی استفاده می‌کند که مزایای خاصی نسبت به خط‌مشی قطعی[8] دارد. SAC به کنشگری نیاز دارد تا آنتروپی پاداش مورد انتظار و توزیع استراتژی را هم‌زمان به حداکثر برساند. معرفی حداکثر آنتروپی توانایی اکتشاف کنش را افزایش می‌دهد و امکان اکتشاف تصمیمات بیشتر سهام و دستیابی به عملکرد پایدارتر در شرایط پیچیده را فراهم می‌کند.

1. Busoniu, de Bruin, Tolić, Kober & Palunko,
2. Loss function
3. Trust Region Policy Optimization (TRPO)
4. Clipping term
5. Haarnoja, Zhou, Abbeel & Levine
6. Off-policy
7. Maximum Entropy
8. Deterministic

فرایند تکراری SAC به ارزیابی خط‌مشی نرم[1] و به‌روزرسانی خط‌مشی نرم تقسیم می‌شود. با استفاده از ارزیابی و به‌روزرسانی خط‌مشی نرم به‌طور مکرر و متناوب، خط‌مشی نهایی به مقدار بهینه همگرا می‌شود. هدف آموزشی SAC به شرح زیر است.

$$\pi^* = arg\max_{\pi} \sum E_{(s_t,a_t)\sim\rho_\pi}[r(s_t,a_t) + \alpha H(\pi(.|s_t))] \quad \text{رابطهٔ ۱۴)}$$

که در آن فرامتغیر α اهمیت نسبی آنتروپی را برای پاداش اندازه‌گیری می‌کند.

Twin-Delayed Deep Deterministic (TD3)

الگوریتم TD3 (فوجیموتو، هوف و مگر[2]، ۲۰۱۸)، جانشین مستقیم DDPG است که با بهبودهایی برای مقابله با مشکل تخمین بیش از حد همراه بوده است. این الگوریتم می‌تواند سوگیری تخمین بیش از حد را کاهش دهد؛ بنابراین با اضافه‌نمودن سه جزء شبکه‌های دو منتقدی بریده‌شده[3]، به‌روزرسانی‌های تأخیری و منظم‌سازی هموارسازی خط‌مشی هدف[4] به DDPG، انباشت خطاها را در فرایند یادگیری کاهش می‌دهد.

استراتژی گروهی

هدف این پژوهش، ایجاد یک استراتژی معاملاتی قدرتمند است. بنابراین، از یک استراتژی گروهی برای انتخاب خودکار بهترین عامل از بین پنج عامل مبتنی بر یادگیری عمیق تقویتی بر اساس بازده تعدیل‌شده با ریسک، برای معامله استفاده می‌شود. فرایند استراتژی گروهی به شرح زیر است.

مرحلهٔ ۱

از یک پنجره زمانی رو به رشد n ماهه برای آموزش مجدد پنج عامل خود به‌طور هم‌زمان استفاده می‌شود. در این پژوهش پنج عامل هر سه ماه یک بار آموزش داده می‌شوند.

مرحلهٔ ۲

هر پنج عامل با استفاده از یک پنجره متحرک اعتبارسنجی ۳ ماهه (۶۳ روزه) بعد از آموزش، اعتبارسنجی می‌شوند تا بهترین عامل با بالاترین بازده تعدیل‌شده با ریسک انتخاب شوند. بازده تعدیل‌شده با ریسک به‌صورت زیر محاسبه می‌شود و در آن، $\bar{r}_p$ بازده مورد انتظار پرتفوی و σ_p انحراف معیار پرتفوی است.

$$Risk - adjusted\ return = \frac{\bar{r}_p}{\sigma_p} \quad \text{رابطهٔ ۱۵)}$$

همچنین ریسک‌گریزی عامل‌ها با استفاده از شاخص تلاطم در مرحله اعتبارسنجی تنظیم می‌شود.

1. soft policy evaluation
2. Fujimoto, Hoof & Meger
3. Clipped Double Critic Networks
4. Target Policy Smoothing Regularization

مرحلهٔ ۳

پس از انتخاب بهترین عامل، از آن برای پیش‌بینی و معامله برای سه ماه بعدی (۶۳ روز) استفاده می‌شود. دلیل این انتخاب آن است که هر عامل معاملاتی به انواع مختلف روندها حساس است؛ یک عامل ممکن است در یک روند صعودی عملکرد خوبی داشته باشد؛ اما در یک روند نزولی خوب عمل نکند. عاملی دیگر ممکن است بیشتر با یک بازار بی‌ثبات سازگار باشد. در هر دوره معاملاتی، عاملی انتخاب می‌شود که بازده تعدیل‌شده با ریسک آن بالاتر باشد؛ زیرا می‌تواند بازده سرمایه‌گذاری را با افزایش ریسک، افزایش دهد.

یافته‌های پژوهش

در این بخش، ابتدا به داده‌های استفاده‌شده در مدل و نحوه پیش‌پردازش آن‌ها پرداخته و سپس ارزیابی عملکرد استراتژی معاملاتی طراحی‌شده ارائه می‌شود. برای پنج عامل مجزا و استراتژی گروهی آزمون پیشینه[1] انجام می‌شود. نتایج نشان می‌دهد که استراتژی گروهی عملکرد بهتری را نسبت به پنج عامل به‌صورت مجزا، شاخص قیمت ۵۰ شرکت بورس اوراق بهادار تهران و استراتژی تخصیص پرتفوی با حداقل واریانس به‌دست می‌آورد. در بخش مقایسه عملکرد، بیشتر به نتایج به‌دست‌آمده پرداخته می‌شود.

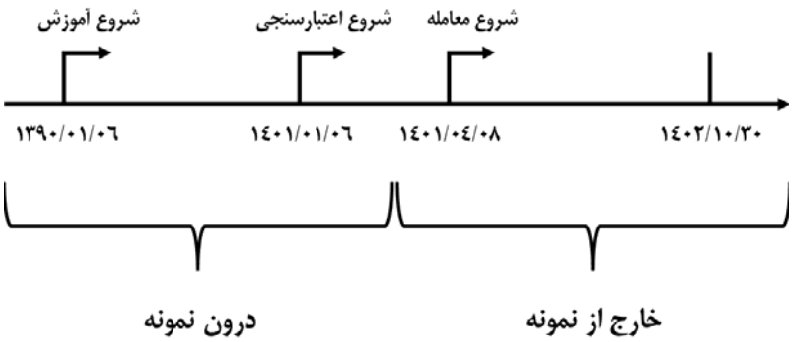
پیش‌پردازش داده‌ها

برای این پژوهش، از داده‌های تاریخی سهام موجود در شاخص قیمتی ۵۰ شرکت بورس اوراق بهادار استفاده شده است. در جدول ۱، نمادهای این شاخص به همراه صنعت آن‌ها آمده است. جهت جلوگیری از مشکل پراکندگی داده[2] از تابع interpolate موجود در کتابخانه pandas استفاده شده است. برای آزمون پیشینه الگوریتم‌ها، از داده‌های این سهام از تاریخ ۱۳۹۰/۰۱/۰۶ تا ۱۴۰۲/۱۰/۳۰ استفاده شده است. مجموعه داده پژوهش از دو دوره تشکیل شده است: ۱. دورهٔ درون نمونه[3] و ۲. دوره خارج از نمونه[4]. همان طور که در شکل ۲ نشان داده شده است، دوره درون نمونه شامل داده‌هایی برای مراحل آموزش و اعتبارسنجی و دوره خارج از نمونه شامل داده‌هایی برای مرحله معاملات است. در مرحله آموزش، پنج عامل یادگیری تقویتی عمیق آموزش داده می‌شود. سپس مرحله اعتبارسنجی پنج عامل با بازده تعدیل‌شده با ریسک و تنظیم پارامترهای کلیدی مانند نرخ یادگیری[5]، تعداد قسمت‌ها[6] و ... انجام و در نهایت در مرحله معامله، سودآوری هر یک از الگوریتم‌ها ارزیابی می‌شود. برای بهره‌برداری بهتر از داده‌های معاملاتی، آموزش عامل در مرحله معامله نیز ادامه دارد؛ زیرا به عامل کمک می‌کند تا سازگاری بیشتری با پویایی بازار داشته باشد. شایان ذکر است که آستانه شاخص تلاطم مالی بر اساس داده‌های تاریخی در دوره آموزش محاسبه و مقدار آن ۲۱۹/۷۸ تعیین شده است.

1. Backtest
2. Data Sparsity
3. In-sample period
4. Out-of-sample period
5. Learning rate
6. Number of episodes

جدول ۱. نمادهای شاخص قیمتی ۵۰ شرکت بورس اوراق بهادار تهران به همراه صنعت فعالیت آن‌ها

صنعت	نماد
استخراج نفت گاز و خدمات جنبی جزء اکتشاف	حفاری
استخراج کانه‌های فلزی	کچاد، کگل، کنور و معادن
بانک‌ها و مؤسسه‌های اعتباری	وبصادر، وپارس، وپاسار، وسینا، وکار، ونوین
حمل‌ونقل، انبارداری و ارتباطات	حکشتی
خدمات فنی و مهندسی	رمپنا
خودرو و ساخت قطعات	خبهمن، خساپا، خودرو
رایانه و فعالیت‌های وابسته به آن	رانفور
ساخت محصولات فلزی	فاراک
سرمایه‌گذاری‌ها	وتوسم، وخارزم، وساپا
سیمان، آهک و گچ	سفارس
شرکت‌های چندرشته‌ای صنعتی	وامید، وصندوق، وغدیر، ونیکی
فراورده‌های نفتی، کک و سوخت هسته‌ای	شبریز، شبندر، شبهرن، شسپا
فلزات اساسی	فاسمین، فخاس، فخوز، فملی، فولاد
محصولات شیمیایی	پارسان، تاپیکو، شاراک، شپدیس، شخارک، شفن، شیراز، شیران، فارس، کرماشا
محصولات غذایی و آشامیدنی جز قند و شکر	وبشهر
مخابرات	اخابر، همراه
مواد و محصولات دارویی	تیپیکو



شکل ۲. نحوه تقسیم‌بندی داده‌های سهام

مقایسهٔ عملکرد

پنج معیار برای ارزیابی نتایج استفاده شده است که در زیر هر یک شرح داده شده‌اند.

- بازده تجمعی[1]: با کم‌کردن ارزش نهایی پرتفوی از مقدار اولیه آن و سپس تقسیم بر ارزش اولیه محاسبه می‌شود که منعکس‌کنندهٔ بازده در پایان مرحله معامله است.

1. Cumulative return

- بازده سالانه[1]: میانگین هندسی پولی است که عامل هر سال در طول دوره زمانی کسب کرده است.
- نوسان‌های سالانه[2]: انحراف معیار سالانه بازده پرتفوی است.
- بازده تعدیل‌شده با ریسک: با تقسیم بازده سالانه بر نوسان‌های سالانه محاسبه می‌شود.
- حداکثر افت سرمایه[3]: حداکثر درصد زیان در طول دوره معاملات است.

با توجه به جدول ۲، مشاهده می‌شود که TD3 با بازده تعدیل‌شده با ریسک ۰/۳۶ بهترین عملکرد را در اولین دوره اعتبارسنجی (۱۴۰۱/۰۱/۰۶ تا ۱۴۰۱/۰۴/۰۸) دارد؛ بنابراین از آن برای معامله سه ماه بعدی (۱۴۰۱/۰۴/۰۸ تا ۱۴۰۱/۰۷/۱۶) استفاده می‌شود. DDPG با بازده تعدیل‌شده با ریسک ۰/۵۶ بهترین عملکرد را در سومین دوره اعتبارسنجی (۱۴۰۱/۰۷/۱۶ تا ۱۴۰۱/۱۰/۱۴) دارد؛ بنابراین از آن برای معامله سه ماه بعدی (۱۴۰۱/۱۰/۱۴ تا ۱۴۰۲/۰۱/۲۸) استفاده می‌شود.

جدول ۲. مدل‌های منتخب در هر دوره بر حسب بازده تعدیل‌شده با ریسک

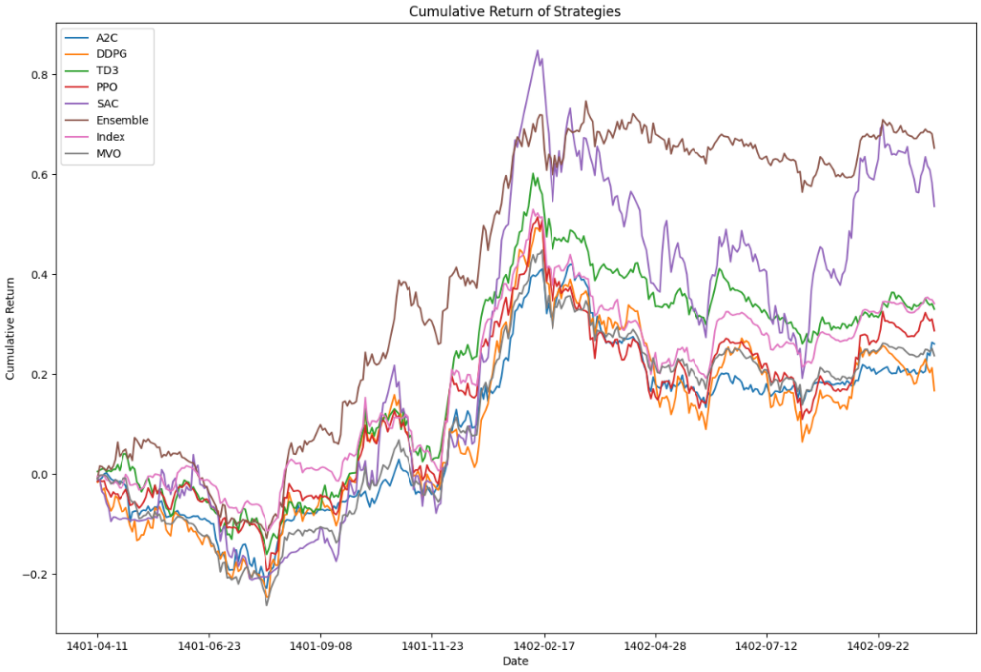
شروع دوره	پایان دوره	مدل منتخب	A2C	PPO	DDPG	SAC	TD3
۱۴۰۱/۰۴/۰۸	۱۴۰۱/۰۷/۱۶	TD3	۰/۰۳	−۰/۳۴	۰/۲۳	−۰/۳۳	**۰/۳٦**
۱۴۰۱/۰۷/۱۶	۱۴۰۱/۱۰/۱۴	TD3	−۰/۴۴	−۰/۶۹	−۰/۳۵	−۱/۱۷	**−۰/۲۵**
۱۴۰۱/۱۰/۱۴	۱۴۰۲/۰۱/۲۸	DDPG	۰/۴۰	۰/۰۱	**۰/۵٦**	−۰/۰۹	۰/۱۹
۱۴۰۲/۰۱/۲۸	۱۴۰۲/۰۴/۳۱	TD3	۰/۵۰	۰/۲۳	۰/۴۷	۰/۲۵	**۰/۷۲**
۱۴۰۲/۰۴/۳۱	۱۴۰۲/۰۸/۰۲	A2C	**−۰/۱۳**	−۰/۵۲	−۰/۲۸	−۰/۶۷	−۰/۲
۱۴۰۲/۰۸/۰۲	۱۴۰۲/۱۰/۳۰	TD3	−۰/۱۴	−۰/۶۶	−۰/۰۹	−۰/۳۳	**۰/۱۳**

در جدول ۳ و شکل ۳ می‌توان مشاهده کرد که عامل A2C با ریسک سازگارتر است و کمترین حداکثر افت سرمایه (۲۵/۴۷− درصد) و نوسان‌های سالانه (۲۲/۴۱ درصد) را در بین پنج عامل دارد؛ بنابراین A2C در مدیریت بازار نزولی بهتر است. عامل SAC در دنبال‌کردن روند بهتر است و در ایجاد بازده بیشتر به خوبی عمل می‌کند؛ این عامل بالاترین بازده سالانه (۲۹/۸۹درصد) و بازده تجمعی (۴۷/۸۹درصد) را در بین پنج عامل دارد. بنابراین SAC هنگام مواجهه با بازار صعودی ترجیح داده می‌شود. TD3 نیز عملکرد مشابهی دارد اما به‌خوبی SAC نیست؛ بنابراین می‌تواند به‌عنوان یک استراتژی مکمل برای SAC در بازار صعودی استفاده شود.

همان طور که در جدول ۳ مشاهده می‌شود، استراتژی گروهی به بازده تعدیل‌شده با ریسک ۱/۵۶ درصدی دست یافته است که بالاتر از استراتژی تخصیص پرتفوی حداقل واریانس و شاخص قیمت ۵۰ شرکت بورس اوراق بهادار تهران است و نشان می‌دهد که این استراتژی به جهت متعادل‌کردن ریسک و بازده، عملکرد قابل قبولی از خود نشان

1. Annualized return
2. Annualized volatility
3. Maximum drawdown

داده است. همچنین با ثبت بازده سالانه ۴۷/۱۳ درصدی و بازده تجمعی ۷۸/۴۷ درصدی، از سایر استراتژی‌ها بـه جهـت بازده نیز عملکرد بهتری داشته است.



شکل ۳. منحنی بازده تجمعی استراتژی گروهی، الگوریتم‌های یادگیری تقویتی عمیق، شاخص قیمت ۵۰ شرکت بورس اوراق بهادار تهران و روش تخصیص پرتفوی حدقل واریانس (ارزش اولیه پرتفوی ۱ میلیون ریال، از ۱٤۰۱/۰٤/۰۸ تا ۱٤۰۲/۱۰/۳۰)

حداکثر افت سرمایه ۱۸/۴۹ درصدی استراتژی گروهی نیز کمتر از رقبای خود بود که انعطاف‌پذیری آن را در طـول رکود بازار برجسته می‌کند. این نتیجه نشان می‌دهد که استراتژی گروهی می‌تواند بـه‌طـور مـؤثری نوسـان‌هـای بـازار را مدیریت کند و از سرمایه‌گذاران در برابر ضررهای قابل توجه محافظت کنـد کـه ایـن ویژگـی در محـیط پرنوسـان بـازار سرمایه ایران بسیار مهم است. به‌طور خلاصه، استراتژی طراحی‌شده نه‌تنها به بازده بالاتری دست یافت، بلکه قابلیت‌های مدیریت ریسک برتر را در مقایسه با معیارهای سنتی و الگوریتم‌های یادگیری تقویتی نشـان داد. ایـن نتیجـه اثربخشـی رویکرد ارائه‌شده را در ارائه ابزاری عملی، متعادل و تطبیقی برای مدیریت پرتفوی در بازار پر نوسان ایران نشان می‌دهد.

بنابراین، یافته‌ها نشان می‌دهد که استراتژی گروهی می‌تواند به‌طور مؤثر یک استراتژی معـاملاتی ایجـاد کنـد کـه عملکردی بهتر از الگوریتم‌های یادگیری تقویتی عمیق و شاخص قیمت ۵۰ شرکت بورس اوراق بهادار تهران و استراتژی تخصیص پرتفوی حداقل واریانس از خود نشان دهد.

جدول ۳. مقایسه عملکرد الگوریتم‌ها در دوره معاملات

۱٤۰۱/۰٤/۰۸ تا ۱٤۰۲/۱۰/۳۰	استراتژی گروهی	شاخص قیمت ۵۰ شرکت	MVO	A2C	DDPG	TD3	PPO	SAC
بازده سالانه	۱۳/۴۷	۲۱/۷۴	۳۹/۱۴	۱۶/۰۸	۶/۲۱	۲۱/۱۵	۱۶/۹۲	۲۹/۸۹
بازده تجمعی	۷۸/۴۷	۳۴/۳۲	۲۲/۲۹	۲۵	۹/۴۳	۳۳/۲۵	۲۶/۳۴	۴۷/۸۹
نوسانات سالانه	۲۷/۱۱	۲۲/۱۴	۲۱/۹۵	۲۲/۴۱	۳۲/۱۷	۲۴/۱۲	۲۶/۷۶	۴۴/۲۲
بازده تعدیل‌شده با ریسک	۱/۵۶	۱/۰۰	۰/۷۲	۰/۷۸	۰/۳۵	۰/۹۲	۰/۷۲	۰/۸۱
حداکثر افت سرمایه	−۱۸/۴۹	−۲۹/۴۹	−۲۷/۶۷	−۲۵/۴۷	−۳۷/۶۷	−۲۹/۷۵	−۳۴/۶۹	−۵۱/۰۵

نتیجه‌گیری و پیشنهادها

در این پژوهش، پتانسیل ترکیب الگوریتم‌های یادگیری تقویتی عمیق، برای مدیریت پویای پرتفوی بررسی شد. به‌منظور تطبیق با موقعیت‌های مختلف بازار، از یک استراتژی گروهی استفاده شد تا به‌طور خودکار بر اساس بازده تعدیل‌شده با ریسک، بهترین عامل را برای معاملهٔ سهام در هر دوره انتخاب کند. طبق نتایج به‌دست‌آمده و مقایسهٔ معیارهای مربوط به بازده و ریسک سرمایه‌گذاری، عملکرد استراتژی گروهی از پنج الگوریتم یادگیری تقویتی عمیق به‌صورت مجزا، شاخص قیمت ۵۰ شرکت بورس اوراق بهادار تهران و روش تخصیص پرتفوی با حداقل واریانس بهتر بوده است.

برای پژوهش‌های آینده پیشنهاد می‌شود که متغیرهای بیشتری برای فضای حالت در نظر گرفته شود و متغیرهایی مانند شاخص‌های بنیادی، متغیرهای اقتصاد کلان، تحلیل پردازش زبان طبیعی اخبار بازار مالی و... به مدل افزوده شوند. همچنین در نظر گرفتن محدودیت‌های قانونی و نظارتی بیشتر در مدل‌سازی بازار سهام و پیاده‌سازی عامل‌های دیگری به‌جز سرمایه‌گذاران می‌تواند مدل را به واقعیت نزدیک‌تر کند و دقت پیش‌بینی‌ها را بهبود بخشد.

منابع

حیدری، مهدی و امیری، حمیدرضا (۱۴۰۱). بررسی قدرت مدل‌های مبتنی بر هوش مصنوعی در پیش‌بینی روند قیمت سهام بورس اوراق بهادار تهران. *تحقیقات مالی*، ۲۴(۴)، ۶۰۲- ۶۲۳.

محبی، سمیه؛ فدائی نژاد، محمد اسماعیل؛ اصولیان، محمد و حمیدیزاده، محمدرضا (۱۴۰۱). انتخاب ویژگی‌های مناسب برای مدل پیش‌بینی شاخص بورس اوراق بهادار تهران بر مبنای تکنیک کاهش ابعاد. *تحقیقات مالی*، ۲۴(۴)، ۵۷۷- ۶۰۱.

نوراحمدی، محمدجواد و نوراحمدی، مرضیه (۱۴۰۲). کاربرد فیلتر کالمن برای تخمین نسبت پوشش ریسک پویا در استراتژی معاملات زوجی (مطالعه موردی: صنعت خودرو). *تحقیقات مالی*، ۲۵(۱)، ۶۳-۸۷.

نوراحمدی، مرضیه؛ رحیمی، علی و صادقی، حجت الله (۱۴۰۳). طراحی سیستم توصیه‌کننده سهام مبتنی بر الگوریتم فیلترینگ مشارکتی برای بورس اوراق بهادار تهران، *تحقیقات مالی*، ۲۶(۲)، ۳۰۲-۳۳۰.

References

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J. & Zaremba, W. (2016). OpenAI Gym. *arXiv:arXiv:1606.01540*

Busoniu, L., de Bruin, T., Tolić, D., Kober, J. & Palunko, I. (2018). Reinforcement learning for control: Performance, stability, and deep approximators. *Annual Reviews in Control.*

Chen, L. & Gao, Q. (2019). Application of deep reinforcement learning on automated stock trading. *2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)*, 29-33.

Chong, T., Ng, W.-K. & Liew, V. (2014). Revisiting the performance of MACD and RSI oscillators. *Journal of Risk and Financial Management*, 1-12.

Craig A., E. & Parbery, S. A. (2005). Is smarter better? A comparison of adaptive, and simple moving average trading strategies. *Research in International Business and Finance*, 399-411.

Deng, Y., Bao, F., Kong, Y., Ren, Z. & Dai, Q. (2016). Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems*, 1-12.

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

Fujimoto, S., Hoof, H. & Meger, D. (2018). Addressing function approximation error in actor-critic methods. *International conference on machine learning*, 1587-1596.

Gurrib, I. (2018). Performance of the average directional index as a market timing tool for the most actively traded USD based currency pairs. *Banks and Bank Systems*, 58-70.

Haarnoja, T., Zhou, A., Abbeel, P. & Levine, S. (2018). Soft actor critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *International conference on machine learning*, 1861-1870.

Heidari, M. & Amiri, H. (2022). Inspecting the Predictive Power of Artificial Intelligence Models in Predicting the Stock Price Trend in Tehran Stock Exchange. *Financial Research Journal, 24*(4), 602-623. (*in Persian*)

Hill, A., Raffin, A., Ernestus, M., Gleave, A., Kanervisto, A., Traore, R., … Wu, Y. (2018). *Stable baselines.* https://github.com/hill-a/stable-baselines.

Jeong, G. & Kim, H. (2019). Improving financial trading decisions using deep Q-learning: predicting the number of shares, action strategies, and transfer learning. *Expert Systems with Applications*, 117, 125- 138.

Jiang, Z. & Liang, J. (2017). Cryptocurrency portfolio management with deep reinforcement learning. *In 2017 Intelligent systems conference (IntelliSys)* (pp. 905-913). IEEE.

Konda, V. & Tsitsiklis, J. (2001). Actor-critic algorithms. *Society for Industrial and Applied Mathematics*. 12.

Kritzman, M. & Li, Y. (2010). Skulls, financial turbulence, and risk management. *Financial Analysts Journal*, 66(5), 30-41.

Lauguico, S., Concepcion II, R., Alejandrino, J., Macasaet, D., Tobias, R. R., Bandala, A. & Dadios, E. (2019). A fuzzy logic-based stock market trading algorithm using bollinger bands. *International conference on humanoid, nanotechnology, information technology, communication and control, environment, and management (HNICEM)*, 1-6.

Li, J., Rao, R. & Shi, J. (2018). Learning to Trade with Deep Actor Critic Methods. *11th International Symposium on Computational Intelligence and Design*, 66-71.

Maitah, M., Procházka, P., Čermák, M. & Šrédl, K. (2016). Comodity Channel index: evaluation of trading rule of agricultural Commodities. *International Journal of Economics and Financial*, 176-178.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 77-91.

Mohebi, S., Fadaeinejad, M. E., Osoolian, M. & Hamidizadeh, M. R. (2022). Feature Selection for the Prediction Model of the Tehran Stock Exchange Index by Dimensionality Reduction Techniques. *Financial Research Journal, 24*(4), 577-601. (*in Persian*)

Neuneier, R. (1996). Optimal asset allocation using adaptive dynamic programming. *Conference on Neural Information Processing Systems*.

Neuneier, R. (1997). Enhancing Q-learning for optimal asset allocation. *Coference on Neural Information Processing Systems*.

Nourahmadi, M. J. & Nourahmadi, M. (2023). Application of Kalman Filter to Estimate Dynamic Hedge Ratio in Pairs Trading Strategy: A Case Study of the Automobile Industry. *Financial Research Journal, 25*(1), 63-87. (*in Persian*)

Nourahmadi, M., Rahimi, A. & Sadeqi, H. (2024). Designing a Stock Recommender System Using the Collaborative Filtering Algorithm for the Tehran Stock Exchange. *Financial Research Journal, 26*(2), 302-330. (*in Persian*)

Pacheco Aznar, D. (2023). P*ortfolio Management: A Deep Distributional RL Approach*. SSRN.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. (2017). *Proximal policy optimization algorithms*. arXiv:1707.06347.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484-489.

Sutton, R. & Barto, A. (1998). *Reinforcement learning: an introduction*. IEEE Transactions on Neural Networks, 1054.

Sutton, R., Mcallester, D., Singh, S. & Mansour, Y. (2000). Policy gradient methods for

reinforcement learning with function approximation. *Conference on Neural Information Processing Systems (NeurIPS)*.

Yang, H., Liu, X.-Y., Zhong, S. & Walid, A. (2020). Deep reinforcement learning for automated stock trading: An ensemble strategy. *In Proceedings of the first ACM international conference on AI in finance*, 1-8.

Yu, K. (2023). Quantitative Trading of Stocks Based on TD3 Algorithm. *Highlights in Science, Engineering and Technology*, 224-231.

Zhang, Y. & Yang, X. (2017). Online portfolio selection strategy based on combining experts' advice. *Computational Economics*, 50(1), 141-159.

Zhang, Z., Zohren, S. & Roberts, S. (2019). Deep reinforcement learning for trading. *arXiv preprint arXiv:1911.10107*.