

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC & KỸ THUẬT MÁY TÍNH**



LUẬN VĂN TỐT NGHIỆP ĐẠI HỌC

**XÂY DỰNG MÔ HÌNH
HỌC TĂNG CƯỜNG TRONG
HOẠT ĐỘNG ĐẦU TƯ TÀI CHÍNH**

Ngành: KHOA HỌC MÁY TÍNH

**HỘI ĐỒNG : KHOA HỌC MÁY TÍNH
GVHD : TS. NGUYỄN AN KHƯƠNG
GVPB : TS. TRẦN TUẤN ANH
SVTH : TRẦN HOÀNG ĐĂNG QUÂN (1813716)**

TP. HỒ CHÍ MINH, 10/2023

KHOA: KH & KT Máy tính
BỘ MÔN: KHMT

NHIỆM VỤ LUẬN VĂN/ ĐỒ ÁN TỐT NGHIỆP
Chú ý: Sinh viên phải dán tờ này vào trang nhất của bản thuyết trình

HỌ VÀ TÊN: TRẦN HOÀNG ĐĂNG QUÂN
NGÀNH: KHMT

MSSV: 1813716
LỚP:

1. Đầu đề luận văn/ đồ án tốt nghiệp: “XÂY DỰNG MÔ HÌNH HỌC TĂNG CƯỜNG TRONG HOẠT ĐỘNG ĐẦU TƯ TÀI CHÍNH”

2. Nhiệm vụ (yêu cầu về nội dung và số liệu ban đầu):

- Tìm hiểu sâu kiến thức nền tảng về thị trường chứng khoán và các nghiệp vụ đầu tư tài chính
- Tìm hiểu sâu về các phương pháp học máy, học sâu và học tăng cường và các ứng dụng của nó cho bài toán dự đoán xu hướng giá cổ phiếu.
- Nghiên cứu tổng quan và trình bày một số công trình nghiên cứu đã có về đề tài này.
- Đề xuất mô hình tổng hợp giải quyết bài toán dự báo xu hướng giá cổ phiếu.
- Thu thập và xử lý, phân tích dữ liệu về một số cổ phiếu quan trọng nhất trong TTCK VN.
- Hiện thực mô hình và đánh giá kết quả trên dữ liệu tự thu thập được.

3. Ngày giao nhiệm vụ: .../.../2023

4. Ngày hoàn thành nhiệm vụ: 30/07/2023

5. Họ tên giảng viên hướng dẫn:

Phản hướng dẫn:

Nguyễn An Khương

HD chính

Nội dung và yêu cầu LVTN/ ĐATN đã được thông qua Bộ môn.

Ngày tháng năm

CHỦ NHIỆM BỘ MÔN

(Ký và ghi rõ họ tên)

GIẢNG VIÊN HƯỚNG DẪN CHÍNH

(Ký và ghi rõ họ tên)

Nguyễn An Khương

PHẦN DÀNH CHO KHOA, BỘ MÔN:

Người duyệt (chấm sơ bộ):

Đơn vị:

Ngày bảo vệ:

Điểm tổng kết:

Nơi lưu trữ LVTN/ĐATN:

Ngày 29 tháng 09 năm 2023

PHIẾU ĐÁNH GIÁ LUẬN VĂN/ ĐỒ ÁN TỐT NGHIỆP
(Dành cho người hướng dẫn)

1. Họ và tên SV: **Trần Hoàng Đăng Quân** MSSV: **1813716** Ngành (chuyên ngành): **KHMT**
2. Đề tài: **"XÂY DỰNG MÔ HÌNH HỌC TĂNG CƯỜNG TRONG HOẠT ĐỘNG ĐẦU TƯ TÀI CHÍNH"**

3. Họ tên người hướng dẫn: **Nguyễn An Khương**

4. Tổng quát về bản thuyết minh:

Số trang: **79**

Số bảng số liệu

Số tài liệu tham khảo: **35**

Hiện vật (sản phẩm)

Số chương: **06**

Số hình vẽ:

Phần mềm tính toán:

5. Những ưu điểm chính của ĐATN:

- Đăng Quân rất chịu khó và kiên nhẫn học để hiểu tương đối rõ cơ sở lý thuyết rất khó trong học tăng cường.
- Đăng Quân nắm vững kiến thức kỹ thuật trong các hoạt động đầu tư chứng khoán.
- Luận văn được trình bày khá công phu, kết quả thực nghiệm của LV khả quan.

6. Những thiếu sót chính của ĐATN: Bộ dữ liệu thu thập chưa đủ "mịn", mô hình chỉ mới hoạt động cho các cổ phiếu riêng rẽ.

7. Đề nghị: Được bảo vệ ☒

Bổ sung thêm để bảo vệ ☐

Không được bảo vệ ☐

8. Các câu hỏi SV phải trả lời trước Hội đồng:

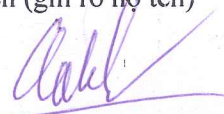
a. Hãy phân tích và so sánh chi tiết cách tiếp cận học sâu và học tăng cường đối với bài toán dự đoán xu hướng giá cổ phiếu?

b. Kế hoạch khái quát hóa mô hình cho đồng thời nhiều cổ phiếu?

9. Đánh giá chung (bằng chữ: Xuất sắc, Giỏi, Khá, TB): **Giỏi**

Điểm : **8.8/10**

Ký tên (ghi rõ họ tên)


Nguyễn An Khương

Ngày tháng năm

PHIẾU ĐÁNH GIÁ LUẬN VĂN/ ĐỒ ÁN TỐT NGHIỆP
(Dành cho người phản biện)

1. Họ và tên SV: Trần Hoàng Đăng Quân
MSSV: 1813716
Ngành (chuyên ngành): KHMT
2. Đề tài: XÂY DỰNG MÔ HÌNH HỌC TĂNG CƯỜNG TRONG HOẠT ĐỘNG ĐẦU TƯ TÀI CHÍNH
3. Họ tên người phản biện: Trần Tuấn Anh
4. Tổng quát về bản thuyết minh:
Số trang: Số chương:
Số bảng số liệu: Số hình vẽ:
Số tài liệu tham khảo: Phần mềm tính toán:
Hiện vật (sản phẩm)
5. Những ưu điểm chính của LV/ ĐATN:
 - Luận văn trình bày một nghiên cứu về dự đoán giá cổ phiếu trong hoạt động đầu tư tài chính sử dụng nền tảng học máy, mà cụ thể là học tăng cường. Đây là một trong những cách tiếp cận mới, nhiều triển vọng trong hoàn cảnh dữ liệu có sự biến thiên mạnh mẽ và liên tục.
 - Luận văn có tham khảo các công trình nghiên cứu khác và có biện luận hướng tiếp cận cụ thể.
 - Luận văn đã có đánh giá các nghiên cứu xung quanh và có so sánh với kết quả của tác giả. Kết quả cũng cho thấy luận văn đã có hiệu quả cao hơn theo các độ đo định sẵn. Luận văn có demo kết quả.
6. Những thiếu sót chính của LV/ĐATN:
 - Dữ liệu sử dụng cho quá trình học còn bị giới hạn khi chỉ sử dụng rất ít thông tin ảnh hưởng tới kết quả trong khi chỉ số chứng khoán thường bị ảnh hưởng rất nhiều bởi các thông tin xung quanh khác.
 - Hướng tiếp cận là mới, sáng tạo, tuy nhiên nên xem xét kỹ hơn tính phù hợp giữa cách tiếp cận và đầu ra mong muốn.
 - Luận văn có các lỗi về chính tả, cú pháp cần chỉnh sửa, ví dụ như công thức toán học, hay sử dụng ngôn ngữ không đồng nhất.
7. Đề nghị: Được bảo vệ ☐ Bỏ sung thêm đề bảo vệ ☐ Không được bảo vệ ☐
8. Các câu hỏi SV phải trả lời trước Hội đồng:
 - a. Mặc dù kết quả đã có những cải tiến nhất định, các sai số trong ứng dụng thực tế còn cao, liệu có an toàn khi sử dụng hay không?
 - b. Phân tích thông số Recall cùng với các thông số khác.
9. Đánh giá chung (bằng chữ: Xuất sắc, Giỏi, Khá, TB): Xuất sắc Điểm : 9.2 /10

Ký tên (ghi rõ họ tên)

Trần Tuấn Anh

LỜI CAM ĐOAN

Chúng tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi. Tất cả những nội dung nghiên cứu và trình bày trong luận văn là trung thực và chưa từng được công bố trước đây. Trong luận văn có sử dụng số liệu và dẫn chứng từ nhiều nguồn khác nhau sẽ được chỉ rõ trong phần trích dẫn tài liệu tham khảo. Nếu phát hiện bất kỳ sự gian lận nào, chúng tôi xin chịu hoàn toàn trách nhiệm đối với luận văn của mình. Trường Đại học Bách khoa Thành phố Hồ Chí Minh và tập thể hướng dẫn không liên quan đến bất kỳ vi phạm về tác quyền và bản quyền do chúng tôi gây ra trong quá trình thực hiện luận văn.

LỜI CẢM ƠN

Sau năm năm học tập tại khoa Khoa học và Kỹ thuật Máy tính, Trường Đại học Bách khoa Thành phố Hồ Chí Minh, chúng tôi đã áp dụng toàn bộ kiến thức đã học được để thực hiện luận văn tốt nghiệp, chủ đề cuối cùng quyết định kết quả học tập và quá trình tốt nghiệp.

Trên hết và trước tiên, chúng tôi xin bày tỏ lòng biết ơn đến các thầy cô giáo của chúng tôi đã cung cấp kiến thức hữu ích trong suốt quãng thời gian học tập trên giảng đường, đặc biệt là Tiến sỹ Nguyễn An Khương đã giúp chúng tôi trong quá trình hoàn thành luận văn. Sự hỗ trợ, động viên và ý tưởng đáng tin cậy của thầy đã góp phần lớn trong việc hoàn thành luận văn.

Chúng tôi cũng muốn gửi lời cảm ơn đặc biệt đến các anh chị ở công ty chứng khoán Vietcap vì những lời khuyên hữu ích của họ trong quá trình thực hiện nghiên cứu.

Điều này sẽ không thể hoàn chỉnh nếu không có gia đình, nguồn năng lượng lớn nhất của chúng tôi. Họ là nguồn hỗ trợ tinh thần lớn nhất trong hành trình của chúng tôi đến với luận văn này.

Chúng tôi đã cố gắng hết sức để tránh sai sót, nhưng nếu có bất kỳ thông tin phát hiện, chúng tôi rất trân trọng nếu được các bạn đưa ra lời khuyên hoặc sửa chữa để chúng tôi có thể cải thiện luận văn của mình. Cuối cùng, chúng tôi xin chân thành cảm ơn và chúc mọi người những điều tốt đẹp nhất.

TÓM TẮT LUẬN VĂN

Từ những năm 2000, việc sử dụng kỹ thuật học máy để dự đoán xu hướng giá cổ phiếu đã trở nên khá phổ biến trên thế giới. Tuy nhiên, phương pháp này vẫn chưa quen thuộc với các nhà đầu tư và nhà kinh tế Việt Nam. Sự bùng nổ của học máy trong nhiều ứng dụng thực tế gần đây đã thúc đẩy việc áp dụng các phương pháp học máy trên thị trường chứng khoán Việt Nam. Do đó, chúng tôi đã quyết định chọn chủ đề này cho luận văn của mình.

Với mục đích xây dựng một bot giao dịch tự động trong tương lai, chúng tôi đã quyết định phát triển các bộ dự đoán xu hướng giá cổ phiếu cho luận văn của mình như là bước đầu tiên. Khác với các nghiên cứu đi trước tập trung vào các kỹ thuật của học giám sát, ở nghiên cứu này chúng tôi sẽ đề xuất một phương pháp sử dụng học tăng cường làm trọng tâm. Kết quả đạt được cạnh tranh về hiệu suất so với các nghiên cứu khác.

MỤC LỤC

Danh sách bảng	vi
Danh sách hình vẽ	vii
1 Giới thiệu	1
1.1 Đặt vấn đề	1
1.2 Mục tiêu và phạm vi của đề tài	3
1.3 Nhiệm vụ của đề tài	4
1.4 Cấu trúc luận văn	4
2 Kiến thức nền tảng	5
2.1 Các khái niệm cơ bản trong chứng khoán	5
2.1.1 Định nghĩa về chứng khoán, cổ phiếu	5
2.1.2 Thị trường chứng khoán là gì?	6
2.2 Một số khái niệm trong phân tích kỹ thuật	6
2.2.1 Biểu đồ nến	6
2.2.2 Đường trung bình động hội tụ phân kỳ	7
2.2.3 Đường trung bình động giản đơn	9
2.2.4 Dải Bollinger	10
2.2.5 Chỉ số sức mạnh tương đối	11
2.2.6 Chỉ số kênh hàng hóa	13
2.2.7 Chỉ báo chuyển động định hướng trung bình	14
2.3 Học tăng cường	16
2.3.1 Giới thiệu	16
2.3.2 Quá trình quyết định Markov	22

2.3.3	Xấp xỉ hàm giá trị	24
2.3.4	Tham số hóa chiến lược	28
3	Các công trình liên quan	37
3.1	Tổng quan các quá trình nghiên cứu trên thế giới về đề tài	37
3.2	Hướng tiếp cận dựa trên CNN	39
3.3	Hướng tiếp cận dựa trên RNN	41
3.4	Hướng tiếp cận dựa trên học tăng cường	44
4	Phương pháp đề xuất	47
4.1	Phát biểu bài toán	47
4.2	Kiến trúc mô hình đề xuất	48
4.3	Chuẩn bị dữ liệu	49
4.3.1	Thu thập dữ liệu	49
4.3.2	Xử lý dữ liệu	52
4.3.3	Chuẩn hóa dữ liệu	53
4.3.4	Dán nhãn dữ liệu	54
4.4	Hiện thực mô hình	54
4.4.1	Không gian trạng thái	54
4.4.2	Không gian hành động	55
4.4.3	Hàm phần thưởng	56
4.4.4	Chiến lược	57
5	Thực nghiệm và kết quả	60
5.1	Bộ dữ liệu	60
5.2	Các độ đo được sử dụng	61
5.3	Thí nghiệm đánh giá	63
5.4	Thí nghiệm mô phỏng dự đoán xu hướng giá	68
6	Tổng kết	69
6.1	Các kết quả đạt được	69
6.2	Hạn chế và hướng phát triển	70
6.2.1	Hạn chế của mô hình	70
6.2.2	Hướng phát triển trong tương lai	70
	Tài liệu tham khảo	70

DANH SÁCH BẢNG

4.1	Danh sách 30 cổ phiếu của chỉ số VN30 đầu năm 2023	51
4.2	Một vài ngày dữ liệu giao dịch của cổ phiếu ACB	52
4.3	Danh sách các đặc trưng dữ liệu sẽ được thêm vào đầu vào của mô hình	53
4.4	Cấu trúc của actor	58
4.5	Cấu trúc của critic	58
4.6	Các tham số được thêm vào nhằm điều chỉnh giá trị phương sai . .	59
5.1	Bộ dữ liệu được sử dụng	60
5.2	Phân bố nhân trên tập kiểm thử của VN30	61
5.3	Môi trường và thư viện sử dụng	63
5.4	Thiết lập siêu tham số cho từng thí nghiệm	64
5.5	Kết quả độ đo trong thí nghiệm của 30 mã chứng khoán trên trong giai đoạn huấn luyện	65
5.6	Kết quả độ đo trong thí nghiệm của 30 mã chứng khoán trong giai đoạn kiểm thử	67
5.7	Kết quả độ đo so với các mô hình khác	68

DANH SÁCH HÌNH VẼ

1.1	Thống kê tài khoản nhà đầu tư chứng khoán	1
2.1	Các thành phần của một biểu đồ nến	7
2.2	Bảng tính ví dụ tính MACD	8
2.3	Bảng tính ví dụ tính SMA(10)	9
2.4	Bảng tính cách tính dải Bollinger	11
2.5	Bảng tính chỉ số RSI	12
2.6	Bảng tính chỉ số CCI	13
2.7	Bảng tính chỉ số ADX	15
2.8	Trò chơi Mario	18
2.9	Một thế cờ trong cờ vua	19
2.10	Xe tự hành đang quan sát cảnh vật xung quanh	20
2.11	Ví dụ về biểu diễn MDP dưới dạng đồ thị	22
2.12	Các lựa chọn khi tạo hàm xấp xỉ giá trị	25
2.13	Trò chơi tìm kho báu	28
2.14	Vòng lặp vô hạn xảy ra khi vào ô vuông xám	29
2.15	Vòng lặp có thể không xảy ra khi đi vào ô vuông xám	29
2.16	Liên tưởng phương pháp Actor Critic với việc chơi trò chơi	33
3.1	Phân bố các bài báo theo các mô hình	38
3.2	Trực quan hóa của mô hình 2D-CNNpred	39
3.3	Kết quả độ đo Macro Average F-measure của mô hình CNNpred	40
3.4	Kiến trúc mô hình KDTCN	41
3.5	Kiến trúc mô hình Time-Weighted LSTM	42

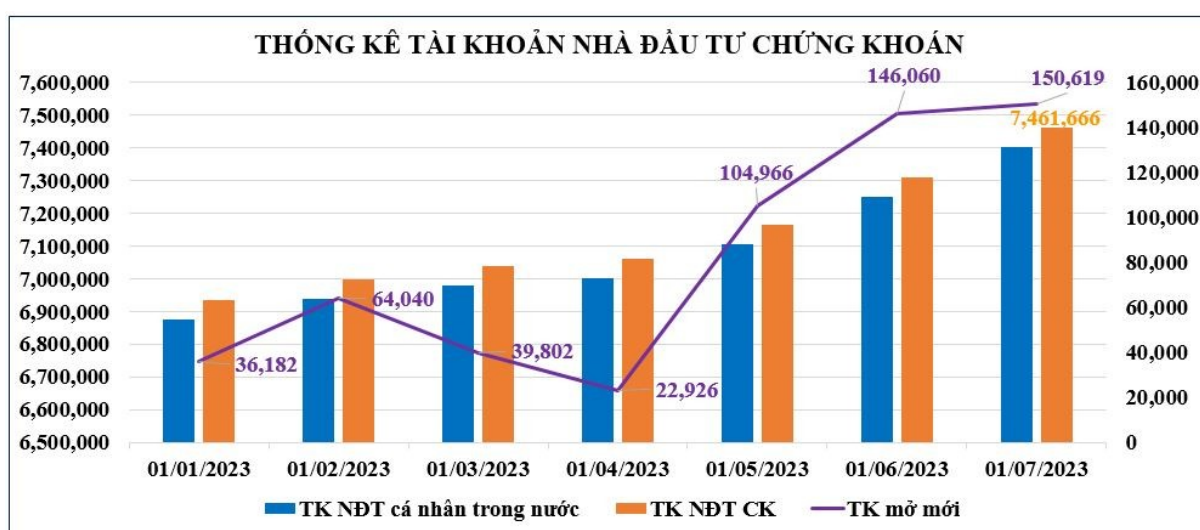
3.6	Kiến trúc của mô hình ALSTM nhằm học đầu vào cho việc huấn luyện đối nghịch	43
3.7	Kiến trúc của mô hình sử dụng huấn luyện đối nghịch	44
3.8	Kết quả trên thang đo độ chính xác và MCC của mô hình Adv-ALSTM và một số mô hình khác	44
3.9	Bộ khung của mô hình AlphaStock	45
3.10	Tổng quan mô hình giao dịch chứng khoán bằng học tăng cường . .	46
3.11	Kết quả so sánh hiệu suất của mô hình	46
4.1	Tổng quan của mô hình dự đoán chứng khoán sử dụng học tăng cường	48
4.2	Một vài quan sát của môi trường trả về cho tác nhân	55
4.3	Ý tưởng của hàm phần thưởng	56
5.1	Trò chơi mô phỏng dự đoán xu hướng giá	68

CHƯƠNG 1

GIỚI THIỆU

1.1 Đặt vấn đề

Thị trường chứng khoán là một trong những chủ đề hấp dẫn nhất hiện nay. Theo số liệu từ Trung tâm Lưu ký Chứng khoán Việt Nam (VSD), trong tháng 7/2023, số lượng tài khoản nhà đầu tư chứng khoán tính đến 31/7 đạt hơn 7,46 triệu tài khoản, tăng khá mạnh so với thời điểm cuối tháng trước, trong đó số lượng tài khoản nhà đầu tư mở mới đã tăng 150.619 tài khoản.



Hình 1.1: Thống kê tài khoản nhà đầu tư chứng khoán¹

Dự đoán xu hướng của thị trường chứng khoán là việc cố gắng tiên đoán liệu giá trị tương lai của cổ phiếu của một công ty hoặc các công cụ tài chính khác được giao dịch trên sàn tăng lên hay giảm xuống? Kể từ khi thị trường chứng khoán ra đời, đã có rất nhiều lý thuyết đầu tư được nghiên cứu nhằm thu được lợi nhuận cao thỏa mãn mong muốn của các nhà đầu tư. Những lý thuyết nổi tiếng nhất trong đầu tư cổ phiếu đã bắt đầu từ gần 100 năm trước và vẫn giữ được ảnh hưởng đến thế hệ các nhà đầu tư ngày nay. Năm 1934, *Security Analysis* [1] được xuất bản lần đầu tiên là một trong những cuốn sách tài chính ảnh hưởng nhất đặt nền móng cho thuyết ngày nay được gọi là đầu tư giá trị (Value Investing), nó được cho ảnh hưởng đến triết lý đầu tư của một trong những nhà đầu tư thành công nhất mọi thời đại Warren Buffett [2]. Năm 1965, giả thuyết thị trường hiệu quả (Efficient Market Hypothesis - EMH) của Fama [3] khẳng định rằng thị trường tài chính là hiệu quả, nói rằng giá cả của các tài sản tài chính đã phản ánh đầy đủ các thông tin hiện có, do đó không thể đánh bại thị trường bằng cách phân tích các biến động giá cả trong quá khứ. Từ đó các mô hình dự đoán chứng khoán tiếp tục phát triển từ sự đóng góp của các chuyên gia và nhà đầu tư tài chính.

Nhờ sự phát triển nhanh chóng của máy tính và truyền thông, giao dịch trực tuyến đã trở thành hoạt động giao dịch chính thay thế cho giao dịch truyền thống trực tiếp, đồng thời cũng tạo điều kiện cho giao dịch thuật toán được hỗ trợ bởi trí tuệ nhân tạo (Artificial Intelligence - AI). Theo một cuộc khảo sát được thực hiện bởi J. P. Morgan vào năm 2023, 53% các nhà giao dịch tin rằng trí tuệ nhân tạo và học máy là công nghệ triển vọng nhất đối với giới đầu tư trong tương lai². Khảo sát về các ứng dụng gần đây của học sâu trong ngành tài chính [4] đưa ra một cái nhìn tổng quát về các cách tiếp cận mới nhất của các nhà nghiên cứu học sâu về bài toán dự đoán cổ phiếu. Dựa trên khảo sát, tác giả cho biết hướng tiếp cận chủ yếu với bài toán là sử dụng các mô hình thuộc lớp học giám sát trong đó mạng thần kinh hồi quy (Recurrent Neural Network - RNN) được sử dụng nhiều nhất vì các ứng dụng của nó trong các bài toán về xử lý chuỗi thời gian. Ngoài cách tiếp cận truyền thống, ngày nay các nhà nghiên cứu đã chú ý đến các hướng tiếp cận khác, chẳng hạn như học bán giám sát, mạng đối nghịch tạo sinh (Generative Adversarial Network - GAN), học chuyển tiếp (Transfer Learning) hay học tăng cường (Reinforcement Learning - RL), tuy nhiên các phương pháp này vẫn chưa phát triển và cần được tiếp tục nghiên cứu.

Mặc dù thị trường chứng khoán Việt Nam chỉ mới phát triển được hơn 20

¹Nguồn ảnh: <https://thoibaotaichinhvietnam.vn/tai-khoan-nha-dau-tu-chung-khoan-mo-moi-thang-7-dat-muc-cao-nhat-tu-dau-nam-133495.html>

²<https://www.jpmorgan.com/solutions/cib/markets/etrading-trends>

năm, nhưng sức hút của nó ngày nay đã hấp dẫn rất nhiều các nhà đầu tư từ mọi lứa tuổi. Trên thế giới, nhiều thành tựu trong lĩnh vực khoa học máy tính hiện đang được áp dụng vào thị trường tài chính, đặc biệt là dự báo giá cổ phiếu và các thuật toán giao dịch. Tuy nhiên, việc sử dụng phương pháp học máy trong tài chính vẫn chưa phổ biến trong giới đầu tư Việt Nam. Do đó, nghiên cứu của chúng tôi được thực hiện với hy vọng cung cấp một số thông tin về hiệu suất của các mô hình học máy trong việc dự đoán xu hướng giá cổ phiếu sử dụng học tăng cường. Chúng tôi hy vọng nghiên cứu này có thể được xem như là bước đầu tiên trong việc xây dựng một công cụ áp dụng học tăng cường hữu ích cho các nhà giao dịch, sẽ được hoạt động như một bot giao dịch tự động trong tương lai.

1.2 Mục tiêu và phạm vi của đề tài

Mục tiêu của đề tài là hiện thực một mô hình dự đoán xu hướng giá cổ phiếu Việt Nam dựa trên kiến thức tài chính và khoa học máy tính với dữ liệu đầu vào là lịch sử giá của một cổ phiếu và dữ liệu đầu ra dự đoán giá ngày mai của cổ phiếu đó sẽ tăng lên hay giảm xuống?

Vì phạm vi nghiên cứu rộng và đa dạng cũng như nguồn lực hạn chế, trong phạm vi của luận văn, sinh viên thực hiện sẽ thêm vào một số hạn chế sau:

- Bỏ qua tác động của các yếu tố kinh tế khác, các thông tin lan truyền trên mạng xã hội,
- Bỏ qua tác động của chi phí giao dịch trong thiết kế mô hình,
- Việc đưa ra các quyết định giao dịch sẽ không ảnh hưởng đến giá cả trên thị trường chứng khoán.

Chúng tôi hy vọng rằng nghiên cứu này có thể đóng góp vào chủ đề về giao dịch thuật toán sử dụng học máy. Khác với các nghiên cứu trước đây tập trung vào các mô hình học giám sát, ở đây chúng tôi tạo ra mô hình dự đoán xu hướng giá cả sử dụng học tăng cường. Kỳ vọng của chúng tôi là có thể xây dựng một mô hình có hiệu suất đáp ứng được trong quá trình huấn luyện và kiểm thử. Ngoài ra, luận văn cũng có thể được xem như là bước đầu tiên để phát triển một bot giao dịch tự động toàn diện trong nghiên cứu tiếp theo, có thể được triển khai trên thị trường giao dịch thực tế.

1.3 Nhiệm vụ của đề tài

Trong luận văn này, sinh viên thực hiện đặt ra bốn nhiệm vụ cần đạt được sau khi thực hiện xong đề tài, gồm:

- Tìm hiểu sâu kiến thức nền tảng về tài chính và phương pháp học tăng cường trong việc ứng dụng cho bài toán dự đoán xu hướng giá cổ phiếu,
- Tìm hiểu tổng quan và chi tiết về một số công trình nghiên cứu đã được thực hiện về đề tài này,
- Đề xuất mô hình tổng hợp giải quyết bài toán,
- Hiện thực mô hình.

1.4 Cấu trúc luận văn

Dựa trên khối lượng công việc, luận văn văn có bố cục như sau:

Chương 1. Giới thiệu: Giới thiệu vắn tắt về vấn đề nghiên cứu, mục tiêu, nhiệm vụ và cấu trúc của luận văn.

Chương 2. Kiến thức nền tảng: Trình bày tóm tắt các đơn vị kiến thức có liên quan hoặc được sử dụng trong luận văn.

Chương 3. Các công trình liên quan: Thảo luận về các thành tựu của các ứng dụng học giám sát, học tăng cường trong vấn đề nghiên cứu và các phương pháp được đề xuất trong việc giải quyết vấn đề.

Chương 4. Phương pháp đề xuất: Trình bày về việc thu thập và xử lý dữ liệu cũng như cách huấn luyện mô hình.

Chương 5. Thực nghiệm và kết quả: Giới thiệu các độ đo sử dụng trong luận văn và tiến hành thí nghiệm nhằm phân tích đánh giá kết quả của mô hình.

Chương 6. Tổng kết: Tổng kết các kết quả đã đạt được trong quá trình nghiên cứu và phát triển luận văn, thảo luận những hạn chế và hướng phát triển trong tương lai.

CHƯƠNG 2

KIẾN THỨC NỀN TẢNG

2.1 Các khái niệm cơ bản trong chứng khoán

2.1.1 Định nghĩa về chứng khoán, cổ phiếu

Chứng khoán là tài sản, bao gồm các loại sau đây¹:

- Cổ phiếu, trái phiếu, chứng chỉ quỹ;
- Chứng quyền, chứng quyền có bảo đảm, quyền mua cổ phần, chứng chỉ lưu ký;
- Chứng khoán phái sinh;
- Các loại chứng khoán khác do Chính phủ quy định.

Cổ phiếu là loại chứng khoán xác nhận quyền và lợi ích hợp pháp của người sở hữu đối với một phần vốn cổ phần của tổ chức phát hành. Có hai loại cổ phiếu là *cổ phiếu thường* và *cổ phiếu ưu đãi*. Cổ phiếu ưu đãi bao gồm các loại sau đây²:

- Cổ phiếu ưu đãi cổ tức,
- Cổ phiếu ưu đãi hoàn lại,
- Cổ phiếu ưu đãi biểu quyết,
- Cổ phiếu ưu đãi khác theo quy định tại điều lệ công ty và pháp luật về chứng

¹<https://luatvietnam.vn/chung-khoan/luat-chung-khoan-2019-179050-d1.html>

²<https://luatvietnam.vn/doanh-nghiep/luat-doanh-nghiep-2020-186272-d1.html>

khoán.

2.1.2 Thị trường chứng khoán là gì?

Thuật ngữ “thị trường chứng khoán” ám chỉ tới những hoạt động trao đổi ở đó các cổ phiếu của các công ty đã phát hành ra công chúng được mua bán. Các hoạt động trao đổi này diễn ra trên một thị trường trao đổi thường hoạt động dưới dạng hệ thống broker-dealer, ở đó cá nhân hay tổ chức vận hành sẽ thực hiện các dịch vụ mua bán cho chính họ hoặc đại diện cho nhà đầu tư. Ngoài hệ thống broker-dealer, thị trường cũng thể hoạt động thông qua quá trình đấu giá mà broker và nhà đầu tư tập trung lại một địa điểm để tương tác trực tiếp. Ngày nay, thị trường chủ yếu hoạt động trực tuyến nhưng ở một số nơi vẫn còn các thị trường hoạt động theo quy trình đấu giá, ví dụ cho loại hình này tiêu biểu là **New York Stock Exchange (NYSE)**³.

Thị trường chứng khoán là một bộ phận của nền kinh tế thị trường tự do. Thị trường cho phép các công ty huy động tiền bằng cách chào bán cổ phiếu và trái phiếu công ty, đồng thời cho phép các nhà đầu tư tham gia vào hoạt động kinh doanh của công ty, kiếm lợi nhuận thông qua sự tăng giảm giá trị của cổ phiếu và kiếm thu nhập thông qua chia cổ tức. Thị trường chứng khoán hoạt động như một kênh ở đó các khoản tiết kiệm và đầu tư của các cá nhân được chuyển thành các cơ hội đầu tư hiệu quả và bổ sung vào quá trình hình thành vốn và tăng trưởng kinh tế của đất nước.

Ngoài các cổ phiếu riêng lẻ, các nhà đầu tư còn quan tâm đến các chỉ số về thị trường. Chỉ số thị trường chứng khoán là một giá trị thống kê phản ánh tình hình của thị trường cổ phiếu. Nó được tổng hợp từ danh sách các cổ phiếu theo phương pháp tính toán tốt nhất. Ở Việt Nam, các chỉ số thị trường phổ biến thường được các nhà đầu tư đề cập là VN-INDEX hoặc VN30-INDEX.

2.2 Một số khái niệm trong phân tích kỹ thuật

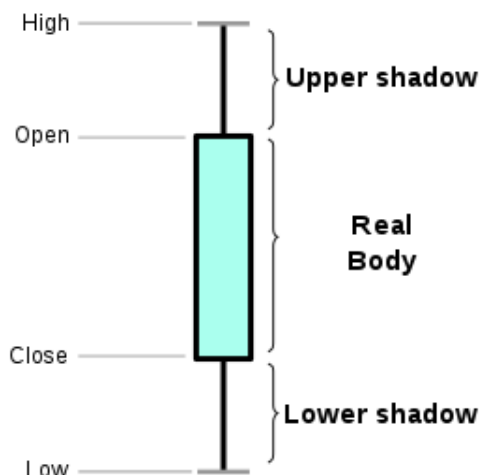
2.2.1 Biểu đồ nến

Biểu đồ nến bắt nguồn từ Nhật Bản hơn 100 năm trước khi phương Tây phát triển biểu đồ thanh và biểu đồ điểm và hình. Vào những năm 1700, một người Nhật Bản tên là Homma đã phát hiện rằng có mối liên hệ giữa giá cả và cung cầu của gạo, thị trường đã bị ảnh hưởng mạnh mẽ bởi cảm xúc của các thương nhân

³<https://www.nyse.com/article/nyse-closing-auction-insiders-guide>

[5].

Biểu đồ nến thể hiện cảm xúc đó bằng cách thể hiện trực quan các chuyển động giá với các màu sắc khác nhau. Các nhà đầu tư sẽ dựa vào đó để đưa ra quyết định giao dịch dựa trên các mẫu nến thường xuyên xảy ra nhằm dự đoán xu hướng giá trong ngắn hạn.



Hình 2.1: Các thành phần của một biểu đồ nến⁴

Hình 2.1 minh họa về các thành phần có trong biểu đồ nến. Biểu đồ nến hiển thị giá cao nhất (high), thấp nhất (low), mở cửa (open), đóng cửa (close) của một cổ phiếu tại một thời điểm cụ thể. Vùng không gian nằm giữa giá đóng cửa và giá mở cửa được gọi là thân nến (real body), nếu thân nến có màu xanh tức là giá đóng cửa cao hơn giá mở cửa thì gọi là nến tăng, ngược lại thì thân nến có màu đỏ còn gọi là nến giảm. Vùng không gian nằm giữa giá cao nhất và giá đóng cửa (nếu là nến tăng) hoặc giá mở cửa (nếu là nến giảm) gọi là bóng nến trên (upper shadow). Vùng không gian nằm giữa giá thấp nhất và giá mở cửa (nếu là nến tăng) hoặc giá đóng cửa (nếu là nến giảm) gọi là bóng nến dưới (lower shadow).

2.2.2 Đường trung bình động hội tụ phân kỳ

Đường trung bình động hội tụ phân kỳ, hay đường MACD (Moving Average Convergence Divergence) ra đời từ năm 1979 bởi nhà phát minh Gerald Appel [6]. Đây được coi là một trong những chỉ báo kỹ thuật phổ biến và thông dụng trong phân tích đầu tư chứng khoán.

MACD là giá trị tìm được khi lấy đường trung bình động lũy thừa (EMA) 12

⁴Nguồn ảnh: https://en.wikipedia.org/wiki/Candlestick_chart

CHƯƠNG 2. KIẾN THỨC NỀN TẢNG

ngày trừ đi đường EMA 26 ngày

$$MACD = EMA(12) - EMA(26),$$

trong đó,

$$EMA(n)_t = \frac{2}{n+1} * (Close_t - EMA(n)_{t-1}) + EMA(n)_{t-1}. \quad (2.1)$$

STT	Date	Price	EMA(12)	EMA(26)	MACD
1	24-Mar-10	22.27			
2	25-Mar-10	22.19			
3	26-Mar-10	22.08			
4	29-Mar-10	22.17			
5	30-Mar-10	22.18			
6	31-Mar-10	22.13			
7	1-Apr-10	22.23			
8	5-Apr-10	22.43			
9	6-Apr-10	22.24			
10	7-Apr-10	22.29			
11	8-Apr-10	22.15			
12	9-Apr-10	22.39	22.23		
13	12-Apr-10	22.38	22.26		
14	13-Apr-10	22.61	22.31		
15	14-Apr-10	23.36	22.47		
16	15-Apr-10	24.05	22.71		
17	16-Apr-10	23.75	22.87		
18	19-Apr-10	23.83	23.02		
19	20-Apr-10	23.95	23.16		
20	21-Apr-10	23.63	23.24		
21	22-Apr-10	23.82	23.33		
22	23-Apr-10	23.87	23.41		
23	26-Apr-10	23.65	23.45		
24	27-Apr-10	23.19	23.41		
25	28-Apr-10	23.10	23.36		
26	29-Apr-10	23.33	23.35	22.90	0.46
27	30-Apr-10	22.68	23.25	22.881045	0.37
28	3-May-10	23.10	23.23	22.897086	0.33
29	4-May-10	22.40	23.10	22.86045	0.24
30	5-May-10	22.17	22.96	22.809491	0.15

Hình 2.2: Bảng tính ví dụ tính MACD

Hình 2.2 minh họa cách tính MACD. Các bước tính như sau:

1. Giá trị cột EMA(12) đầu tiên nằm ở STT 12, giá có thể được tính bằng SMA(12). Từ STT 13, các giá trị EMA được tính theo công thức (2.1), ví dụ vào dòng có STT 12 giá trị EMA(12) là $\frac{2}{12+1} * (22.38 - 22.23) + 22.23 \approx 22.26$,
2. Giá trị cột EMA(26) tính tương tự như EMA(12),
3. MACD được tính từ STT 26, vào dòng có STT 26 giá trị MACD là $23.25 -$

22.881045 \approx 0.46.

Khi phân tích, bên cạnh đường MACD cơ bản còn có đường tín hiệu. Đường tín hiệu sẽ là đường trung bình động EMA 9 ngày của MACD. Nếu đường MACD giao với đường tín hiệu từ dưới lên sẽ báo hiệu giá sẽ tăng hơn mức hiện tại. Nếu đường MACD vượt đường tín hiệu từ trên xuống báo hiệu giá đang trên đà giảm.

2.2.3 Đường trung bình động giản đơn

Đường trung bình động giản đơn hay đường SMA (Simple Moving Average) là một trong 3 đường trung bình động (Moving Average) phổ biến, bên cạnh đường trung bình lũy thừa (EMA) và đường trung bình tỷ trọng tuyến tính (WMA). SMA được tính bằng trung bình cộng các mức giá đóng cửa trong một khoảng thời gian giao dịch nhất định

$$SMA_n = \frac{C_{n-13} + C_{n-12} + \dots + C_n}{n},$$

trong đó:

- C : Giá đóng cửa,
- $1, 2, 3, \dots, n$: Số thứ tự của phiên gần nhất tính từ 1. n chính là số phiên được tính cho SMA.

STT	Date	Price	10-day SMA	
1	24-Mar-10	22.27		
2	25-Mar-10	22.19		
3	26-Mar-10	22.08		
4	29-Mar-10	22.17		
5	30-Mar-10	22.18		
6	31-Mar-10	22.13		
7	1-Apr-10	22.23		
8	5-Apr-10	22.43		
9	6-Apr-10	22.24		
10	7-Apr-10	22.29	▼	22.22
11	8-Apr-10	22.15	▼	22.21
12	9-Apr-10	22.39	▼	22.23
13	12-Apr-10	22.38	▼	22.26
14	13-Apr-10	22.61	▼	22.31

Hình 2.3: Bảng tính ví dụ tính SMA(10)

Hình 2.3 minh họa cách tính SMA của 10 ngày. Giá trị SMA(10) của dòng có STT 12 là $\frac{22.27+22.19+\dots+22.61}{10} = 22.22$.

Đường trung bình động giản đơn có vai trò như đường hỗ trợ và kháng cự

trong phân tích kỹ thuật. Khi nến giá nằm trên đường trung bình động giản đơn, nó đóng vai trò là vùng hỗ trợ. Ngược lại khi SMA nằm trên nến giá, nó đóng vai trò là vùng kháng cự. Tùy thuộc vào độ lớn của đường trung bình động giản đơn (50, 100, 200) sẽ tương ứng với mức hỗ trợ hoặc kháng cự mạnh⁵.

2.2.4 Dải Bollinger

Dải Bollinger (Bollinger Bands) là một công cụ phân tích kỹ thuật được xác định bởi một tập hợp các đường xu hướng. Nó được vẽ như hai độ lệch chuẩn, cả tích cực và tiêu cực, nhằm so sánh khoảng cách so với đường trung bình động giản đơn (SMA) và có thể được điều chỉnh theo sở thích của người sử dụng.

Dải Bollinger được phát triển bởi nhà giao dịch kỹ thuật John Bollinger và được thiết kế để cung cấp cho các nhà đầu tư tự tin hơn khi xác định một tài sản bị bán quá mức hoặc mua quá mức⁶

$$BOLU = SMA(close, n) + m * \sigma[close, n],$$

$$BOLD = SMA(close, n) - m * \sigma[close, n],$$

trong đó:

- *BOLU*: đường Bollinger trên của dải,
- *BOLD*: đường Bollinger dưới của dải,
- *SMA*: đường trung bình động giản đơn,
- *n*: số lượng ngày nhất định,
- *m*: hệ số độ lệch chuẩn,
- *σ* : độ lệch chuẩn của của giá trong *n* ngày gần nhất.

⁵<https://www.fidelity.com/viewpoints/active-investor/moving-averages>

⁶<https://www.bollingerbands.com/bollinger-bands>

Bollinger Bands (20,2)						
	Date	Price	Middle Band 20-day SMA	20-day Standard Deviation	Upper Band 20-day SMA + STDEVx2	Lower Band 20-day SMA - STDEVx2
1	1-May-09	86.16				
2	4-May-09	89.09				
3	5-May-09	88.78				
4	6-May-09	90.32				
5	7-May-09	89.07				
6	8-May-09	91.15				
7	11-May-09	89.44				
8	12-May-09	89.18				
9	13-May-09	86.93				
10	14-May-09	87.68				
11	15-May-09	86.96				
12	18-May-09	89.43				
13	19-May-09	89.32				
14	20-May-09	88.72				
15	21-May-09	87.45				
16	22-May-09	87.26				
17	26-May-09	89.50				
18	27-May-09	87.90				
19	28-May-09	89.13				
20	29-May-09	90.70	88.71	1.29	91.29	86.12
21	1-Jun-09	92.90	89.05	1.45	91.95	86.14
22	2-Jun-09	92.98	89.24	1.69	92.61	85.87

Hình 2.4: Bảng tính cách tính dải Bollinger

Hình 2.4 minh họa cách tính dải Bollinger của 20 ngày và hệ số độ lệch chuẩn là hai. Các bước tính ở ngày có số 20 ở cột đầu tiên như sau:

1. Tính SMA(20), giá trị là 88.71;
2. Tính độ lệch chuẩn của 20 ngày gần nhất, giá trị là 1.29;
3. Giá trị BOLU bằng $88.71 + 2 * 1.29 = 91.29$;
4. Giá trị BOLD bằng $88.71 - 2 * 1.29 = 86.12$.

Dải Bollinger cung cấp cho các nhà giao dịch sự biến động dựa trên giá. Khi giá tiến gần hơn đến dải trên, điều đó cho thấy thị trường có thể bị mua quá mức. Ngược lại, thị trường có thể bị bán quá mức khi giá cuối cùng di chuyển đến gần dải dưới.

2.2.5 Chỉ số sức mạnh tương đối

Chỉ số sức mạnh tương đối hay RSI (Relative Strength Index) tính toán tỷ lệ giữa mức tăng giá và giảm giá trung bình trong một khoảng thời gian nhất định,

thể hiện tình trạng quá mua và quá bán của thị trường [7]

$$RSI = 100 - \left[\frac{100}{1 + RS} \right],$$

$$RS = \frac{Average\ gain}{Average\ loss},$$

trong đó *Average gain/loss* là phần trăm tăng/giảm trung bình trong một khoảng thời gian, thông thường đối với RSI là 14 ngày.

STT	Date	Close	Change	Gain	Loss	Avg Gain	Avg Loss	RS	14-day RSI
1	14-Dec-09	44.34							
2	15-Dec-09	44.09	-0.25		0.25				
3	16-Dec-09	44.15	0.06	0.06					
4	17-Dec-09	43.61	-0.54		0.54				
5	18-Dec-09	44.33	0.72	0.72					
6	21-Dec-09	44.83	0.50	0.50					
7	22-Dec-09	45.10	0.27	0.27					
8	23-Dec-09	45.42	0.33	0.33					
9	24-Dec-09	45.84	0.42	0.42					
10	28-Dec-09	46.08	0.24	0.24					
11	29-Dec-09	45.89	-0.19		0.19				
12	30-Dec-09	46.03	0.14	0.14					
13	31-Dec-09	45.61	-0.42		0.42				
14	4-Jan-10	46.28	0.67	0.67				RS	RSI
15	5-Jan-10	46.28				0.24	0.10	2.39	70.53
16	6-Jan-10	46.00	-0.28		0.28	0.22	0.11	1.97	66.32
17	7-Jan-10	46.03	0.03	0.03		0.21	0.10	1.99	66.55

Hình 2.5: Bảng tính chỉ số RSI

Hình 2.5 minh họa cách tính RSI của 14 ngày, các bước giá trị RSI ở dòng có STT 15 như sau:

1. Tính giá trị $Change_t = Close_t - Close_{t-1}$, ở ngày có STT 2 giá trị là $44.09 - 44.35 = -0.25$;
2. Nếu $Change_t > 0$, điền giá trị vào cột Gain, ngược lại điền giá trị vào cột Loss;
3. Tính $Average\ gain = \frac{0.06+0.72+...+0.67}{14} = 0.24$, $Average\ loss = \frac{0.25+0.54+...+0.42}{14} = 0.10$;
4. Giá trị $RS = \frac{0.24}{0.10} = 2.39$;
5. Giá trị RSI tương ứng là $100 - \frac{100}{1+2.39} = 70.53$.

Chỉ số RSI được hiển thị trực quan dưới dạng đồ thị đường nằm trong khoảng từ 0-100. Trong đó, nếu RSI lớn hơn 70, về lý thuyết có nghĩa là cổ phiếu đang bị mua quá mức, nhỏ hơn 30 có nghĩa là cổ phiếu đang bị bán quá mức. Ở giữa mức 30 và 70 được coi là vùng trung tính, với mức 50 được là dấu hiệu không có xu hướng [8].

2.2.6 Chỉ số kênh hàng hóa

Chỉ số kênh hàng hóa hay CCI (Commodity Channel Index) đánh giá xu hướng và sức mạnh của xu hướng giá, cho phép các nhà giao dịch xác định xem họ muốn tham gia hay thoát khỏi giao dịch, không tham gia giao dịch hay thêm vào một vị thế hiện có [9]

$$CCI = \frac{TP - SMA(TP, n)}{.015 * MD(n)},$$

trong đó:

- TP (Typical price): $TP = \frac{High + Close + Low}{3}$,
- $SMA(TP, n)$ là đường trung bình động giản đơn của đường TP trong n ngày gần nhất,
- MD (Mean deviation): $MD(n)_t = (\sum_{i=t-13}^t |SMA(TP, n)_i - TP_t|) / n$.

Microsoft						20-day	20-day	
Date	Open	High	Low	Close	Typical Price	SMA of TP	Mean Deviation	20-day CCI
1 24-Aug-10	23.94	24.20	23.85	23.89	23.98			
2 25-Aug-10	23.85	24.07	23.72	23.95	23.92			
3 26-Aug-10	23.94	24.04	23.64	23.67	23.79			
4 27-Aug-10	23.73	23.87	23.37	23.78	23.67			
5 30-Aug-10	23.60	23.67	23.46	23.50	23.54			
6 31-Aug-10	23.46	23.59	23.18	23.32	23.36			
7 1-Sep-10	23.53	23.80	23.40	23.75	23.65			
8 2-Sep-10	23.73	23.80	23.57	23.79	23.72			
9 3-Sep-10	24.09	24.30	24.05	24.14	24.16			
10 7-Sep-10	23.95	24.15	23.77	23.81	23.91			
11 8-Sep-10	23.92	24.05	23.60	23.78	23.81			
12 9-Sep-10	24.04	24.06	23.84	23.86	23.92			
13 10-Sep-10	23.83	23.88	23.64	23.70	23.74			
14 13-Sep-10	24.05	25.14	23.94	24.96	24.68			
15 14-Sep-10	24.89	25.20	24.74	24.88	24.94			
16 15-Sep-10	24.95	25.07	24.77	24.96	24.93			
17 16-Sep-10	24.91	25.22	24.90	25.18	25.10			
18 17-Sep-10	25.24	25.37	24.93	25.07	25.12			
19 20-Sep-10	25.13	25.36	24.96	25.27	25.20			
20 21-Sep-10	25.26	25.26	24.93	25.00	25.06	24.21	0.55	102.31
21 22-Sep-10	24.74	24.82	24.21	24.46	24.50	24.24	0.56	30.74

Hình 2.6: Bảng tính chỉ số CCI

Hình 2.6 minh họa cách tính chỉ số CCI của 20 ngày. Các bước tính toán ở ngày có giá trị 20 ở cột đầu tiên như sau:

1. Giá trị $TP = \frac{25.26 + 25.26 + 24.93}{3} = 25.06$;

2. Giá trị $SMA(TP, 20) = \frac{23.98 + \dots + 25.06}{20} = 24.21$;
3. Giá trị $MD(20) = \frac{\|24.21 - 23.98\| + \|24.21 - 23.92\| + \dots + \|24.21 - 25.06\|}{20} = 0.55$;
4. Giá trị CCI tương ứng là $\frac{25.06 - 24.21}{.015 * 0.55} = 102.31$.

Chỉ báo CCI là một đường trung bình động thường dao động quanh đường 0 và có giá trị từ -100 đến +100. CCI chủ yếu được sử dụng để phát hiện các xu hướng mới, theo dõi các mức mua quá mức và bán quá mức, đồng thời phát hiện điểm yếu trong các xu hướng khi chỉ báo phân kỳ với giá. Khi CCI di chuyển từ vùng âm hoặc gần bằng 0 lên trên 100, điều đó có thể cho thấy giá đang bắt đầu một xu hướng tăng mới. Khái niệm tương tự cũng áp dụng cho một xu hướng giảm khi chỉ báo chuyển từ giá trị dương hoặc gần bằng 0 xuống dưới -100, thì xu hướng giảm có thể đang bắt đầu.

2.2.7 Chỉ báo chuyển động định hướng trung bình

Chỉ báo chuyển động định hướng trung bình (Average Directional Movement Index - ADX) đo lường sức mạnh của xu hướng [7]. Xu hướng có thể tăng hoặc giảm, được thể hiện dựa trên hai chỉ báo đi kèm, đó là chỉ báo định hướng âm (Negative Directional Indicator - -DI) và chỉ báo định hướng dương (Positive Directional Indicator - +DI). Vì vậy, chỉ báo ADX được biểu diễn bởi ba đường khác nhau.

ADX thường biến động trong khoảng từ 0-100. Đường ADX thể hiện sức mạnh của xu hướng, trong khi đó hai đường còn lại thể hướng của xu hướng.

CHƯƠNG 2. KIẾN THỨC NỀN TẢNG

		High	Low	Close	TR	+DM 1	-DM 1	TR14	+DM14	-DM14	+DI14	-DI14	DI 14 Diff	DI 14 Sum	DX	ADX
3	13-Feb-09	30.45	29.96	30.10	0.48	0.17	0.00									
4	17-Feb-09	29.35	28.74	28.90	1.36	0.00	1.22									
5	18-Feb-09	29.35	28.56	28.92	0.79	0.00	0.19									
6	19-Feb-09	29.29	28.41	28.48	0.88	0.00	0.15									
7	20-Feb-09	28.83	28.08	28.56	0.75	0.00	0.33									
8	23-Feb-09	28.73	27.43	27.56	1.31	0.00	0.65									
9	24-Feb-09	28.67	27.66	28.47	1.11	0.00	0.00									
10	25-Feb-09	28.85	27.83	28.28	1.02	0.19	0.00									
11	26-Feb-09	28.64	27.40	27.49	1.24	0.00	0.44									
12	27-Feb-09	27.68	27.09	27.23	0.58	0.00	0.31									
13	2-Mar-09	27.21	26.18	26.35	1.05	0.00	0.91									
14	3-Mar-09	26.87	26.13	26.33	0.73	0.00	0.05									
15	4-Mar-09	27.41	26.63	27.03	1.08	0.54	0.00	13.33	0.90	4.32	6.75	32.42	25.67	39.17	65.54	
16	5-Mar-09	26.94	26.13	26.22	0.90	0.00	0.49	13.28	0.84	4.51	6.29	33.95	27.65	40.24	68.73	
17	6-Mar-09	26.52	25.43	26.01	1.09	0.00	0.70	13.42	0.78	4.89	5.78	36.43	30.65	42.21	72.60	
18	9-Mar-09	26.52	25.35	25.46	1.17	0.00	0.08	13.63	0.72	4.62	5.29	33.89	28.60	39.17	73.01	
19	10-Mar-09	27.09	25.88	27.03	1.63	0.57	0.00	14.29	1.24	4.29	8.70	30.02	21.32	38.71	55.06	
20	11-Mar-09	27.69	26.96	27.45	0.72	0.59	0.00	13.99	1.75	3.98	12.49	28.47	15.98	40.96	39.01	
21	12-Mar-09	28.45	27.14	28.36	1.31	0.76	0.00	14.30	2.38	3.70	16.68	25.87	9.19	42.55	21.60	
22	13-Mar-09	28.53	28.01	28.43	0.51	0.08	0.00	13.79	2.29	3.43	16.63	24.90	8.27	41.53	19.92	
23	16-Mar-09	28.67	27.88	27.95	0.78	0.14	0.00	13.59	2.27	3.19	16.69	23.47	6.78	40.16	16.87	
24	17-Mar-09	29.01	27.99	29.01	1.06	0.35	0.00	13.67	2.45	2.96	17.93	21.65	3.72	39.59	9.40	
25	18-Mar-09	29.87	28.76	29.38	1.11	0.86	0.00	13.80	3.14	2.75	22.73	19.92	2.81	42.64	6.59	
26	19-Mar-09	29.80	29.14	29.36	0.66	0.00	0.00	13.48	2.91	2.55	21.61	18.94	2.67	40.55	6.59	
27	20-Mar-09	29.75	28.71	28.91	1.04	0.00	0.43	13.56	2.71	2.80	19.95	20.64	0.68	40.59	1.69	
28	23-Mar-09	30.65	28.93	30.61	1.74	0.90	0.00	14.33	3.41	2.60	23.82	18.13	5.69	41.94	13.57	33.58
29	24-Mar-09	30.60	30.03	30.05	0.58	0.00	0.00	13.89	3.17	2.41	22.81	17.36	5.45	40.18	13.57	32.15

Hình 2.7: Bảng tính chỉ số ADX

Hình 2.7 minh họa cách tính chỉ số ADX của 14 ngày, các bước tính giá trị ADX như sau:

$$1. TR_t = \max(High_t - Low_t, \|High_t - Close_{t-1}\|, Low_t - Close_{t-1})$$

Ở dòng thứ hai, giá trị TR là $\max(30.28 - 29.32, \|30.28 - 29.87\|, \|29.32 - 29.87\| = 0.96)$;

$$2. +DM1_t = \begin{cases} \max(High_t - High_{t-1}, 0), & \text{nếu } High_t - High_{t-1} > Low_{t-1} - Low_t, \\ 0, & \text{ngược lại,} \end{cases}$$

ở dòng thứ hai vì $30.28 - 30.2 < 29.41 - 29.32$ nên giá trị sẽ là 0;

$$3. -DM1_t = \begin{cases} \max(Low_{t-1} - Low_t, 0), & \text{nếu } Low_{t-1} - Low_t > High_t - High_{t-1}, \\ 0, & \text{ngược lại,} \end{cases}$$

ở dòng thứ hai vì $29.41 - 29.32 > 30.28 - 30.2$ nên giá trị sẽ gần là 0.09;

4. Các giá trị $TR14$, $+DM14$, $-DM14$ tính theo công thức Wilder, ví dụ

$$TR14_t = \begin{cases} \sum_{i=0}^{13} TR_i, & t = 13, \\ TR14_{t-1} - \frac{TR14_{t-1}}{14} + TR_t, & t > 13, \end{cases}$$

ở dòng thứ 16, TR14 có giá trị là $13.33 - \frac{13.33}{14} + 0.9 \approx 13.28$;

5. Các giá trị $+DI14$, $-DI14$ được tính như sau

$$\begin{aligned} +DI14 &= 100 * \frac{+DM14}{TR14}, \\ -DI14 &= 100 * \frac{-DM14}{TR14}, \end{aligned}$$

ở ngày thứ 16 giá trị $+DI14$ là $100 * \frac{0.9}{13.33} \approx 6.75$;

6. Tính giá trị $DI14Diff$ như sau $DI14Diff = || +DI14 - -DI14 ||$, ở dòng thứ 16 giá trị này là $32.42 - 6.75 = 25.67$;

7. Tính giá trị $DI14Sum$ như sau $DI14Sum = +DI14 + -DI14$, ở dòng thứ 16 giá trị này là $32.42 + 6.75 = 39.17$;

8. Tính giá trị $DX = \frac{DI14Diff}{DI14Sum}$, ở dòng thứ 16 giá trị này là $100 * \frac{25.67}{39.17} \approx 65.54$;

9. Tính giá trị ADX

$$ADX_t = \begin{cases} \sum_{i=0}^{13} DX_i, & t = 13, \\ (ADX_{t-1} * 13) + DX_t / 14, & t > 13, \end{cases}$$

ở dòng 29 giá trị này là $((33.58 * 13) + 13.57) / 14 \approx 32.15$.

2.3 Học tăng cường

2.3.1 Giới thiệu

Định nghĩa. Học tăng cường khác với học giám sát ở chỗ không có sự giám sát, chỉ có phần thưởng được trả về. Tác nhân sẽ học dựa trên một quá trình thử và sai. Ngoài ra, phần thưởng trong học tăng cường nhận về bị trì hoãn chứ không phải ngay lập tức. Trong RL, thời gian đóng vai trò quan trọng. Tác nhân sẽ tương tác trong một quá trình tuần tự theo thời gian [10].

Một quá trình học tăng cường bao gồm một giá trị phần thưởng vô hướng (scalar reward) R_t mà tác nhân (agent) nhận về ở mỗi thời điểm t . Mục tiêu của agent là thu thập R_t để tối ưu tổng phần thưởng nhận về. Có thể nói mọi mục tiêu (goal) hay rộng hơn là mọi bài toán trong RL có thể được mô tả qua các tín hiệu phần thưởng và được giải quyết bằng cách tối ưu tổng phần thưởng tích lũy kỳ vọng. Đây được gọi là *giả thuyết phần thưởng (reward hypothesis)*.

Để đạt được mục tiêu, tác nhân cần phải luôn luôn chọn hành động (action) sao cho tối ưu tổng phần thưởng. Nhưng hành động có thể ảnh hưởng đến phần thưởng nhận về trong tương lai. Vì vậy, tác nhân không thể luôn chọn các lựa chọn

tham lam ở hiện tại mọi lúc. Do đó, tác nhân tốt hơn nên hy sinh các hành động đem lại phần thưởng tức thì để ưu tiên có được phần thưởng tốt hơn trong tương lai. Đây được gọi là ưu tiên khám phá hơn khai thác.

Các thành phần liên quan trong học tăng cường ở thời điểm t bao gồm:

- Quan sát (observation) O_t mà tác nhân quan sát được,
- Hành động A_t mà tác nhân sẽ đưa ra,
- Giá trị phần thưởng vô hướng R_t mà tác nhân nhận được tương ứng với hành động.

Tác nhân thực hiện hành động đối với môi trường (environment), do đó ảnh hưởng đến môi trường phản hồi lại một phần thưởng và một quan sát mà tác nhân sẽ nhận lại và đưa ra một hành động mới.

Lịch sử (history) là một chuỗi của quan sát, hành động và phần thưởng

$$H_t = A_1, O_1, R_1, \dots, A_t, O_t, R_t.$$

Do đó, lịch sử là tất cả các biến quan sát được cho đến thời điểm t . Các biến quan sát được tiếp theo trong tương lai phụ thuộc rất lớn vào lịch sử. Để giải quyết bài toán, cần phải có một ánh xạ từ lịch sử đến hành động tiếp theo. Lưu ý rằng phần thưởng phải luôn là giá trị vô hướng và thông thường phần thưởng được định nghĩa qua một công thức nội tại của môi trường.

Trạng thái (state) là hàm phụ thuộc vào lịch sử

$$S_t = f(H_t).$$

Trạng thái có thể được chia ra làm nhiều loại khác nhau. Đầu tiên là trạng thái môi trường (environment state) S_t^e , có thể là hiểu là biểu diễn thông tin nội tại của môi trường. Nó có thể là các giá trị số và nó sẽ quyết định quan sát tiếp theo mà môi trường sẽ phản hồi. Thông tin này thông thường sẽ không tiết lộ cho tác nhân. Vì vậy các giải thuật trong RL thông thường sẽ không truy cập đến thông tin này. Thông tin này không phải bao giờ cũng có nghĩa đối với tác nhân do nó có thể chứa các mẫu thông tin không liên quan và ảnh hưởng đến khả năng đưa ra hành động của tác nhân. Tiếp theo là trạng thái tác nhân (agent state) S_t^a , nó có thể ghi lại toàn bộ quan sát và/hoặc phần thưởng mà nó nhận được, thông thường nó chỉ ghi lại các giá trị gần nhất. Một phần trong việc giải một bài toán RL là tìm và xây dựng một hàm f ở đó trạng thái tác nhân S_t^a là hàm phụ thuộc vào lịch sử. Lưu ý rằng với bài toán RL có sự tham gia của nhiều tác nhân, trạng thái

của tác nhân này sẽ được xem là một phần của môi trường theo góc nhìn của các tác nhân còn lại.

Hình 2.8 sẽ minh họa về cách thiết lập một bài toán học tăng cường. Ví dụ như sử dụng học tăng cường trong việc chơi trò chơi. Như trong trò Mario, để giành được chiến thắng người chơi phải tiến xa nhất có thể. Hành động A_t sẽ là các bước di chuyển đi sang trái, phải hoặc nhảy lên, quan sát O_t là khung ảnh hiện tại người chơi đang nhìn thấy. Phần thưởng R_t được tự định nghĩa sao cho thắng được trò chơi, ở đây phần thưởng có thể được thiết lập như thời gian ít nhất, khoảng cách xa điểm khởi đầu nhất, ... Trạng thái S_t có thể chọn là quan sát mà người chơi quan sát được hoặc nhiều khung cảnh gần nhất quan sát được.



Hình 2.8: Trò chơi Mario⁷

Trạng thái Markov (Markov State): Trạng thái Markov có thuộc tính Markov. Thuộc tính Markov nói rằng tương lai không phụ thuộc vào quá khứ, nếu biết được thông tin hiện tại. Nó cho ta biết trạng thái hiện chứa đủ thông tin so với lịch sử, nghĩa là nó chứa toàn bộ thông tin có giá trị trong lịch sử. Một trạng thái có thuộc tính Markov khi và chỉ khi

$$\Pr[S_{t+1}|S_t] = \Pr[S_{t+1}|S_1, \dots, S_t].$$

Có thể hiểu xác suất xảy ra S_{t+1} nếu biết được S_t ngang với xác suất xảy ra S_{t+1} khi biết được lịch sử của trạng thái. Ví dụ thực trong mô hình dự đoán dự báo thời tiết có thuộc tính Markov, xác suất dự đoán ngày mai trời có mưa chỉ dựa vào thời tiết hôm nay. Chẳng hạn nếu như hôm nay trời nắng, thì khả năng

⁷Nguồn ảnh: <https://github.com/Kautenja/gym-super-mario-bros/tree/master>

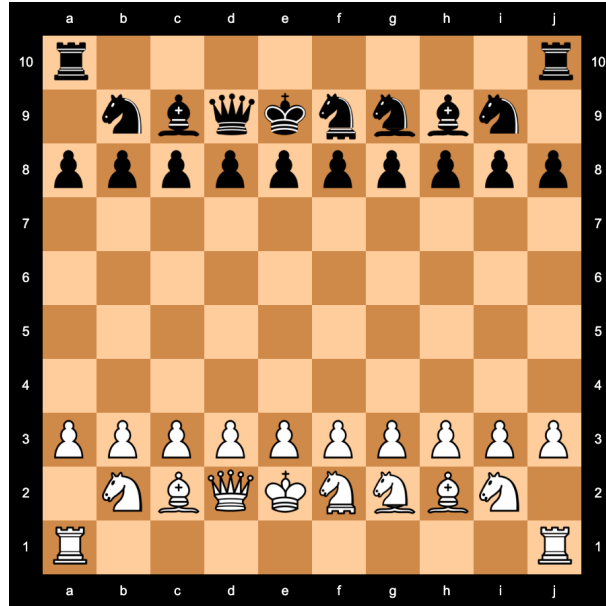
ngày mai mưa là rất thấp mặc cho ngày hôm trước có mưa to hay không.

Môi trường quan sát toàn phần (Fully Observable Environment).

Môi trường ở đó tác nhân quan sát trực tiếp trạng thái môi trường. Điều đó có nghĩa

$$O_t = S_t^a = S_t^e.$$

Hiểu theo khái niệm, điều này nói lên rằng môi trường không có trạng thái ẩn mà tác nhân không biết. Tác nhân biết mọi thứ về môi trường và một quan sát tiết lộ toàn bộ thông tin về môi trường. Môi trường này còn được gọi là quá trình quyết định Markov (Markov Decision Process - MDP). Trò chơi cờ vua là một môi trường quan sát toàn phần vì người chơi quan sát được toàn bộ bàn cờ cũng như các nước đi của đối thủ, lúc này bàn cờ vua có thể lấy làm quan sát đồng thời cũng là trạng thái của tác nhân lần môi trường.



Hình 2.9: Một thế cờ trong cờ vua⁸

Môi trường quan sát một phần (Partial Observability Environment).

Tác nhân chỉ có thể quan sát một phần trạng thái môi trường. Ví dụ như rô-bốt quan sát thông qua camera không được cho biết vị trí tuyệt đối trong không gian mà chỉ biết vị trí tương đối (hình ảnh hiện tại thu được) hay một bot giao dịch chỉ có thể biết giá hiện tại chứ không phải lịch sử giá. Trong trường hợp này, trạng thái tác nhân không tương đương với trạng thái môi trường

$$S_t^a \neq S_t^e.$$

⁸Nguồn ảnh: <https://www.chess.com>

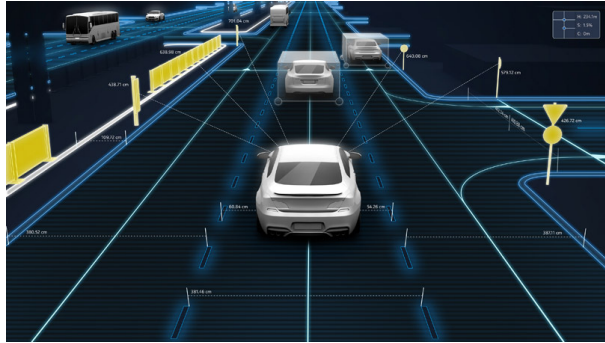
Môi trường này còn được gọi là quá trình quyết định Markov quan sát một phần (Partially Observable Markov Decision Process - POMDP). Trong trường hợp này, cần phải tìm những cách thay thế để biểu diễn trạng thái của tác nhân. Một cách là đặt trạng thái tác nhân bằng với lịch sử $S_t^a = H_t$. Một cách khác là dùng phân phối trạng thái có thể có của môi trường

$$S_t^a = (\Pr[S_t^e = s^1], \dots, \Pr[S_t^e = s^n]).$$

Ví dụ khi lái xe, khi người lái xe bấm còi với xe phía trước, người lái xe phía trước có thể chọn các cách phản ứng khác nhau dẫn đến trạng thái của người lái xe tiếp nhận khác nhau. Cuối cùng, trạng thái tác nhân có thể được biểu diễn thông qua RNN

$$S_t = \sigma(S_{t-1}W_s + O_tW_o).$$

Bài toán học xe tự hành minh họa cho thiết lập này, trạng thái hiện tại của tác nhân sẽ được biểu diễn với đầu vào là trạng thái lúc trước và quan sát cảnh vật hiện tại.



Hình 2.10: Xe tự hành đang quan sát cảnh vật xung quanh⁹

Tác nhân trong học tăng cường. Một tác nhân có thể có ít nhất một trong các thành phần sau:

- Chiến lược (policy): mô tả hành vi của tác nhân, cách mà tác nhân ra hành động khi biết trạng thái của nó,
- Hàm giá trị (value function): một hàm diễn tả liệu trạng thái hoặc/và hành động tốt đến đâu,
- Mô hình (model): biểu diễn thông tin của môi trường mà tác nhân có được.

Chiến lược. Chiến lược mô tả hành vi của tác nhân. Chiến lược sẽ ánh xạ từ trạng thái đến hành động, hoặc từ trạng thái đến phân phối của hành động.

⁹Nguồn ảnh: <https://www.pcmag.com>

Với chiến lược tất định (deterministic policy), chiến lược luôn trả về chung hành động đối với một trạng thái nhất định $a = \pi(s)$. Nhưng với chiến lược ngẫu nhiên (stochastic policy), nó sẽ trả về xác suất xảy ra hành động đối với một trạng thái $\pi(a|s) = \Pr[A = a|S = s]$. Ví dụ trong cờ vua, chiến lược chính là cách người chơi suy nghĩ để di chuyển các quân cờ nhằm thắng cuộc.

Hàm giá trị. Hàm này sẽ trả về một dự đoán về phần thưởng dự kiến trong tương lai. Khi có hai lựa chọn, hàm giá trị sẽ hướng dẫn trạng thái nào cần chuyển sang. Hàm giá trị v của chiến lược π sẽ nói lên phần thưởng dự kiến trong một khoảng thời gian trong tương lai

$$v_\pi(s) = \mathbb{E}[R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \dots | S_t = s].$$

Tham số $\gamma \in [0, 1]$ có thể được suy giảm theo thời gian. γ gần về 1 nói lên phần thưởng trong tương lai được xem trọng và ngược lại. Việc điều chỉnh tham số này cho ta kiểm soát mức độ quan trọng khi xử lý phần thưởng xa trong tương lai. Ví dụ với cờ vua, hàm giá trị là cách người chơi đưa ra các đánh giá về thế cờ như thế cờ bất lợi, có lợi hoặc chiếu hết đối thủ.

Mô hình. Mô hình sẽ dự đoán môi trường sẽ làm gì tiếp theo. Có hai loại là mô hình chuyển đổi (transition model) và mô hình phần thưởng (reward model). Mô hình chuyển đổi \mathcal{P} dự đoán trạng thái tiếp theo của môi trường khi biết trạng thái hiện tại

$$P_{ss'}^a = \Pr[S' = s' | S = s, A = a].$$

Mô hình phần thưởng \mathcal{R} dự đoán phần thưởng tiếp theo

$$R_s^a = \mathbb{E}[R | S = s, A = a].$$

Ví dụ với cờ vua, mô hình có thể là cách người chơi dự đoán nước cờ đối thủ sẽ đi và phần thưởng mà đối thủ đạt được khi ở thế cờ hiện tại.

Phân loại. Tác nhân trong học tăng cường có thể phân loại thành ba loại cơ bản:

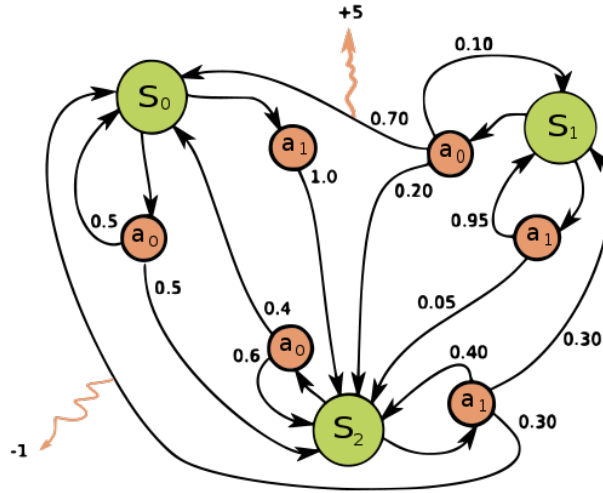
- Tác nhân học dựa trên giá trị (value based): các thuật toán không có chiến lược và chỉ có hàm giá trị,
- Tác nhân học dựa trên chiến lược (Policy based): các thuật toán không có hàm giá trị và chỉ có chiến lược,
- Actor critic: các thuật toán cả chiến lược và hàm giá trị.

Ngoài ra, tác nhân có thể được phân loại thành cách học *không có mô hình* (*model free*) và *dựa trên mô hình* (*model based*).

2.3.2 Quá trình quyết định Markov

Quá trình quyết định Markov (Markov Decision Process - MDP) dùng để mô tả môi trường trong RL. MDP rất quan trọng vì hầu hết môi trường đều có thể được biểu diễn dưới dạng MDP. Quá trình quyết định Markov mang thuộc tính Markov, vì vậy trạng thái của tương lai chỉ phụ thuộc vào hiện tại. Quá trình quyết định Markov là một bộ $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ [10]:

- \mathcal{S} là một tập hợp các trạng thái còn gọi là không gian trạng thái (state space),
- \mathcal{A} là tập hợp các hành động còn gọi là không gian hành động (action space),
- ma trận chuyển đổi trạng thái $\mathcal{P}_{ss'}^a = \Pr[S_{t+1} = s' | S_t = s, A_t = a]$,
- tín hiệu phần thưởng $\mathcal{R}_s^a = \mathbb{E}[R_{t+1} | S_t = s, A_t = a]$,
- hệ số khấu hao $\gamma \in [0, 1]$.



Hình 2.11: Ví dụ về biểu diễn MDP dưới dạng đồ thị¹⁰

Nếu một bài toán học tăng cường có thể chuyển về MDP, nó có thể biểu diễn dưới dạng một đồ thị như Hình 2.11. Có tổng cộng hai hành động a_0, a_1 . Các đỉnh S_0, S_1, S_2 biểu thị cho trạng thái. Ma trận chuyển đổi trạng thái $\mathcal{P}_{ss'}^a$ cho biết xác suất chuyển sang trạng thái s' khi đứng ở trạng thái s và hành động a . Ở Hình 2.11, xác suất để chuyển từ S_0 đến S_2 khi chọn hành động a_0 là 0.5. Tín hiệu phần

¹⁰Nguồn ảnh: https://en.wikipedia.org/wiki/Markov_decision_process

thường là một hàm phụ thuộc vào trạng thái và hành động và có thể được biểu diễn trên cạnh nối liền hai đỉnh, như từ S_1 đến S_0 khi chọn a_0 thì phần thưởng là $+5$.

Chiến lược π là một phân bố trên hành động, khi biết trước trạng thái

$$\pi(a|s) = \Pr[A_t = a | S_t = s].$$

Hàm giá trị của MDP. Có hai định nghĩa cần chú ý với hàm giá trị của MDP. Đầu tiên là hàm giá trị trạng thái $v_\pi(s)$, nó là lợi nhuận kỳ vọng (expected return) khi đi từ trạng thái s và tuân theo chiến lược π

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_\pi[G_t | S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) | S_t = s]. \end{aligned}$$

Thứ hai là hàm giá trị hành động $q_\pi(s, a)$ được hiểu là lợi nhuận kỳ vọng (expected return) khi đi từ trạng thái s , chọn hành động a và tuân theo chiến lược π

$$\begin{aligned} q_\pi(s, a) &= \mathbb{E}_\pi[G_t | S_t = s, A_t = a] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) | S_t = s, A_t = a]. \end{aligned}$$

Giá trị hành động, còn được gọi là q -value, chỉ tác nhân biết hành động cần chọn khi ở một trạng thái nhất định.

Hàm giá trị trạng thái có thể được hiểu là trung bình các q -value tương ứng với toàn bộ hành động có thể chọn khi ở trạng thái s

$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) q_\pi(s, a).$$

Tương tự, bằng cách dùng định nghĩa đệ quy, q -value có thể biểu diễn dùng giá trị trạng thái

$$q_\pi(s, a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a v_\pi(s').$$

Mục tiêu đạt được tổng phần thưởng kỳ vọng có thể đạt được thông qua việc tìm hàm giá trị trạng thái tối ưu

$$v_\star(s) = \max_{\pi} v_\pi(s).$$

hoặc hàm q -value tối ưu

$$q_{\star}(s, a) = \max_{\pi} q_{\pi}(s, a).$$

Trong các chiến lược tối ưu, các giá trị trạng thái $v_{\pi_{\star}}(s) = v_{\star}$ và q -value $q_{\pi_{\star}}(s, a) = q_{\star}(s, a)$ đều tối ưu. Một cách để tìm chiến lược tối ưu là dùng hàm giá trị hành động tối ưu

$$\pi_{\star}(a|s) = \begin{cases} 1, & \text{nếu } a = \operatorname{argmax}_{a \in \mathcal{A}} q_{\star}(s, a), \\ 0, & \text{trong các trường hợp còn lại.} \end{cases}$$

Chiến lược tối ưu này sẽ sinh ra phân bố xác suất luôn chọn hành động có q -value cao nhất và bỏ qua các hành động còn lại. Ví dụ với cờ vua, giả sử ở thế cờ hiện tại người chơi có thể đưa ra nhiều nước cờ khác nhau mà trong đó có một nước cờ có thể chiếu hết đối thủ, thì người chơi sẽ chọn nước cờ này và bỏ qua các nước cờ còn lại.

Công thức tối ưu Bellman (Bellman Optimality Equation). Công thức này định nghĩa cách tìm hàm giá trị tối ưu.

Hàm giá trị trạng thái tối ưu $v_{\star}(s)$ được tính bằng cách lấy giá trị của hàm giá trị hành động tối ưu q_{\star} của mỗi trạng thái, sau đó chọn giá trị hành động cao nhất.

$$v_{\star}(s) = \max_a q_{\star}(s, a).$$

Ngược lại, hàm giá trị hành động tối ưu q_{\star} có thể được tính bằng cách lấy trung bình các giá trị trạng thái của các trạng thái mà có thể đi đến qua hành động tối ưu khi tuân theo chiến lược π

$$q_{\star}(s, a) = \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_{\star}(s').$$

2.3.3 Xấp xỉ hàm giá trị

Các phương pháp tìm hàm giá trị lưu trữ lại trong bộ nhớ như Monte Carlo (MC) hay Temporal Difference (TD) sẽ tính ra $V(s)$ hoặc $Q(s, a)$. Có thể xem các phương pháp này tạo ra một bảng lưu trữ lại giá trị của hàm giá trị. Tuy nhiên, với các bài toán RL có không gian trạng thái lớn, việc lưu trữ sẽ gây tốn kém bộ nhớ và không hiệu quả [10].

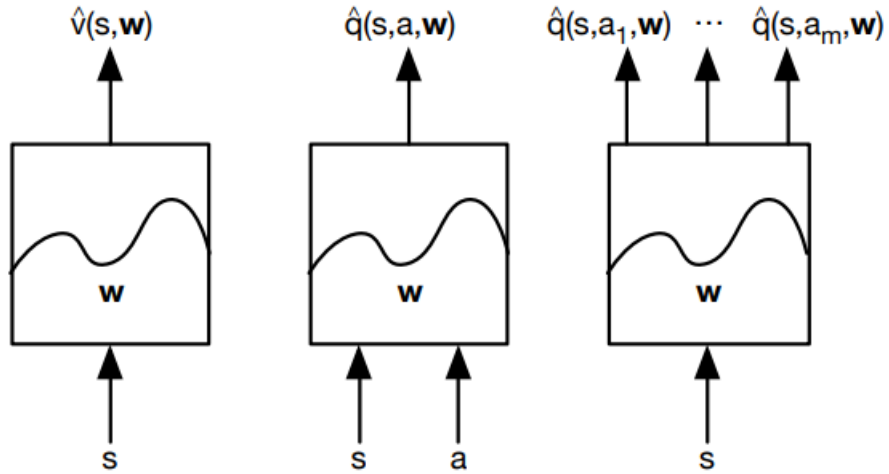
Ví dụ trong trò chơi cờ vua, nếu trạng thái của tác nhân chọn là vị trí các quân cờ thì không gian trạng thái là quá lớn. Một sự thật ít người biết đến là số

nước đi có thể của quân Mã là hơn 122 triệu nước¹¹, và khi xét đến toàn bộ vị trí mà các quân cờ có thể tạo ra thì việc lưu trữ một “bảng” giá trị trạng thái là bất khả thi.

Giải pháp khả thi hơn là tạo ra một hàm xấp xỉ các giá trị này. Để có thể xấp xỉ giá trị trạng thái $\hat{v}(s; \mathbf{w})$ hay q -value $\hat{q}(s, a; \mathbf{w})$, chúng sẽ được tham số hóa bằng các trọng số có thể huấn luyện w . Việc này sẽ tạo ra một giải pháp xấp xỉ gần đúng đáng tin cậy

$$\begin{aligned}\hat{v}(s; \mathbf{w}) &\approx v_{\pi}(s), \\ \hat{q}(s, a; \mathbf{w}) &\approx q_{\pi}(s, a).\end{aligned}$$

Để đạt được một hàm xấp xỉ tốt, các trọng số w cần được tối ưu có thể bằng mạng thần kinh nhân tạo hoặc phương pháp MC, TD. Thông thường có ba cách để tạo ra các hàm xấp xỉ này. $\hat{v}(s; \mathbf{w})$ có thể xấp xỉ từ đầu vào s . Và hàm xấp xỉ này sẽ nói lên mức độ tốt khi có đầu vào là trạng thái s . Đối với $\hat{q}(s, a; \mathbf{w})$, có hai lựa chọn. Lựa chọn đầu tiên là xấp xỉ q -value với đầu vào là cặp (trạng thái, hành động). Cách thứ hai là xấp xỉ chỉ với đầu vào chỉ bao gồm trạng thái, nhưng đầu ra bao gồm một vec-tơ $[\hat{q}(s, a_1; \mathbf{w}), \dots, \hat{q}(s, a_m; \mathbf{w})]^T$ mà mỗi q -value tương ứng với hành động.



Hình 2.12: Các lựa chọn khi tạo hàm xấp xỉ giá trị¹²

Hàm mục tiêu trong bài toán xấp xỉ. Điều đầu tiên để tạo ra hàm xấp xỉ giá trị trạng thái và giá trị hành động với mạng thần kinh nhân tạo là học cách

¹¹<https://vnexpress.net/22-phat-hien-thu-vi-ve-co-vua-2743691.html>

¹²Nguồn ảnh: <https://www.davidsilver.uk/wp-content/uploads/2020/03/FA.pdf>

áp dụng SGD (Stochastic Gradient Descent). Ở đây, $J(\mathbf{w})$ được định nghĩa là hàm mất mát của bài toán xấp xỉ có thể tính đạo hàm, thì đạo hàm của nó là:

$$\nabla_{\mathbf{w}} J(\mathbf{w}) = \begin{bmatrix} \frac{\partial J(\mathbf{w})}{\partial w_1} \\ \vdots \\ \frac{\partial J(\mathbf{w})}{\partial w_n} \end{bmatrix}.$$

Như vậy, trọng số w có thể được học qua mỗi bước lặp có công thức sau

$$\Delta \mathbf{w} = -\frac{1}{2} \alpha \nabla_{\mathbf{w}} J(\mathbf{w}),$$

trong đó α là tốc độ học. Nếu hàm mất mát dùng sai lệch bình phương (squared-error) thì trọng số w sẽ học theo công thức sau:

$$\begin{aligned} \nabla J(\mathbf{w}) &= -\frac{1}{2} \alpha \nabla_{\mathbf{w}} J(\mathbf{w}) \\ &= -\frac{1}{2} \alpha \nabla_{\mathbf{w}} \mathbb{E}_{\pi} [(v_{\pi}(S) - \hat{v}(S; \mathbf{w}))^2] \\ &= \alpha \mathbb{E}_{\pi} [(v_{\pi}(S) - \hat{v}(S; \mathbf{w})) \nabla_{\mathbf{w}} \hat{v}(S; \mathbf{w})]. \end{aligned}$$

Từ đó suy ra

$$\Delta \mathbf{w} = \alpha (v_{\pi}(S) - \hat{v}(S; \mathbf{w})) \nabla_{\mathbf{w}} \hat{v}(S; \mathbf{w}),$$

trong đó $v_{\pi}(S)$ chính là giá trị mà hàm xấp xỉ của giá trị trạng thái cần học.

MC và TD trong việc học hàm xấp xỉ. Hàm xấp xỉ $v_{\pi}(S)$ có thể được học bằng cách dùng SGD qua sai lệch giữa giá trị dự đoán $\hat{v}(S; \mathbf{w})$ và giá trị nhãn $v_{\pi}(S)$. Tuy nhiên, thực tế giá trị nhãn này không có sẵn nhưng nó có thể được tính thông qua các phương pháp MC hay TD. Với MC, giá trị lợi nhuận G_t có thể được dùng làm nhãn

$$\Delta \mathbf{w} = \alpha (G_t - \hat{v}(S; \mathbf{w})) \nabla_{\mathbf{w}} \hat{v}(S; \mathbf{w}),$$

trên một tập dữ liệu

$$(S_1, G_1), (S_2, G_2), \dots, (S_T, G_T).$$

Tương tự, nhãn cũng được tính khi dùng TD(0) và TD(λ)

$$\begin{aligned} \Delta \mathbf{w} &= \alpha (R_t + \gamma \hat{v}(S_{t+1}; \mathbf{w}) - \hat{v}(S; \mathbf{w})) \nabla_{\mathbf{w}} \hat{v}(S; \mathbf{w}), \\ \Delta \mathbf{w} &= \alpha (G_t^{\lambda} - \hat{v}(S; \mathbf{w})) \nabla_{\mathbf{w}} \hat{v}(S; \mathbf{w}). \end{aligned}$$

Sự khác biệt giữa MC và TD khi dùng để tính nhân là, MC có đặc điểm phương sai cao (high variance) trong khi TD có độ chệch cao (high bias). Lưu ý là tập dữ liệu dùng cho việc học nhân của hàm giá trị không có sẵn mà sẽ được thu thập từng chút một theo thời gian qua sự tương tác với môi trường.

Xấp xỉ q -value. Tương tự với việc tạo ra hàm xấp xỉ cho giá trị trạng thái, ý tưởng này có thể được áp dụng với $q_\pi(s, a)$ để tính xấp xỉ giá trị q -value. Với trường hợp này, đầu vào cho hàm xấp xỉ qua mạng thần kinh nhân tạo sẽ có sự thay đổi

$$\mathbf{x}(S, A) = \begin{bmatrix} x_1(S, A) \\ \vdots \\ x_n(S, A) \end{bmatrix}.$$

Ví dụ, để xấp xỉ q_π với hồi quy tuyến tính, công thức hàm xấp xỉ như sau:

$$\hat{q}(S, A; \mathbf{w}) = \mathbf{x}(S, A)^\top \mathbf{w} = \sum_{j=1}^n x_j(S, A) w_j.$$

Dùng lô (batch) trong tính hàm xấp xỉ. Ở trên hàm xấp xỉ được tính theo từng bước. Cách này không hiệu quả và có thể được cải thiện khi học theo lô tương tự như cách học mạng thần kinh nhân tạo. Cách này sẽ khiến việc học hàm xấp xỉ thành một bài toán học giám sát theo thời gian thực ở đó mục tiêu là cải thiện hàm xấp xỉ

$$\hat{v}(s; \mathbf{w}) \approx v_\pi(s).$$

Tập dữ liệu cho bài toán học giám sát này, còn được gọi là tập kinh nghiệm \mathcal{D} , sẽ bao gồm tập các cặp (trạng thái, giá trị trạng thái) và sẽ được giải khi tìm thấy giá trị nhỏ nhất của bình phương sai lệch nhằm tối ưu trọng số w

$$\min_{\mathbf{w}} LS(\mathbf{w}) = \min_{\mathbf{w}} \sum_{i=1}^T (v_t^\pi - \hat{v}(s_t; \mathbf{w}))^2.$$

Tập dữ liệu \mathcal{D} sẽ lưu trữ lịch sử tương tác giữa tác nhân và môi trường mà tác nhân thực hiện trong suốt một tập (episode). Khi có tập dữ liệu, có thể sử dụng SGD và cập nhật theo từng mẫu trong tập dữ liệu lưu trữ kinh nghiệm này $(s, v^\pi) \sim \mathcal{D}$ và cập nhật trọng số w

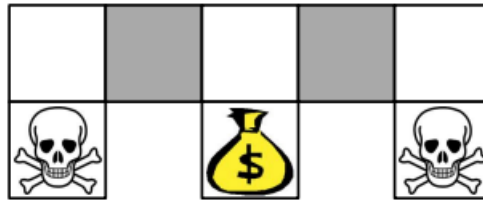
$$\Delta \mathbf{w} = \alpha (v^\pi - \hat{v}(s; \mathbf{w})) \nabla_{\mathbf{w}}(s; \mathbf{w}).$$

Một cách khác là dùng “minibatch gradient descent”. Thay vì chỉ cập nhật từng mẫu kinh nghiệm một, trọng số w sẽ được cập nhật theo một lô các cặp. Cách này còn được gọi là “experience replay”.

2.3.4 Tham số hóa chiến lược

Các thuật toán thuộc phương pháp tham số hóa chiến lược (Policy Gradient - PG) xây dựng chiến lược bằng cách thực hiện “Stochastic Gradient Descent” (SGD) trực tiếp. Nghĩa là chiến lược sẽ được tham số hóa [10], và mục tiêu là tìm ra các tham số tối ưu của chiến lược $\pi_{\theta}(a|s)$.

Phương pháp này rất hợp với các bài toán học tăng cường muốn xây dựng các chiến lược ngẫu nhiên như bài toán tìm kho báu mê cung.



Hình 2.13: Trò chơi tìm kho báu¹³

Ưu điểm của phương pháp này là không cần phải làm việc với hàm giá trị để tìm ra chiến lược, mà làm việc trực tiếp với chiến lược. Hơn nữa, phương pháp Policy Gradient hội tụ tốt hơn, hiệu quả trong các bài toán RL có không gian hành động có miền giá trị liên tục và có khả năng xây dựng nên các chiến lược ngẫu nhiên. Nhược điểm của nó là dễ hội tụ về các cực tiểu cục bộ (local minimum) thay vì điểm toàn cục (global minimum) và việc học các chiến lược này sẽ khó hơn và có đặc điểm phương sai cao.

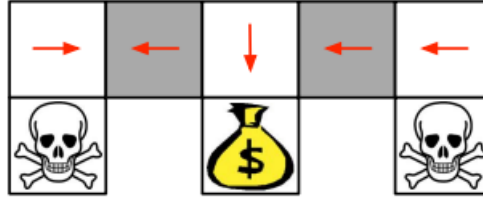
Một trong những đặc điểm cần chú ý là nó rất hiệu quả với cái bài toán có không gian lớn. Với phương pháp dựa trên giá trị, hàm giá trị thường phải làm việc với phép max như công thức dưới đây:

$$Q(s, a) \leftarrow Q(s, a) + \alpha(R_s^a + \gamma \max_{a'} Q(s', a') - Q(s, a)).$$

Nhưng các phép tính max rất tốn tài nguyên, do đó đây là ưu điểm là các phương pháp Policy Gradient lại làm tốt. Một điều cần chú ý là có thể xây dựng được các chiến lược ngẫu nhiên thay vì chiến lược tất định, các chiến lược tất định có sẽ hoạt động tốt hơn trong môi trường quyết định Markov một phần (POMDPs).

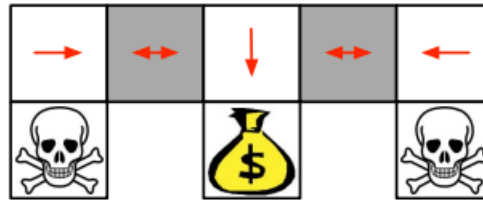
¹³Nguồn ảnh: <https://www.davidsilver.uk/wp-content/uploads/2020/03/pg.pdf>

Trong các môi trường POMDPs, các chiến lược tất định thường đưa ra các hành động dẫn đến các vòng lặp vô hạn, bởi vì nó sẽ đưa ra một hành động cụ thể trong một môi trường mà nó không biết được toàn bộ thông tin.



Hình 2.14: Vòng lặp vô hạn xảy ra khi vào ô vuông xám¹⁴

Ngược lại, xây dựng chiến lược dùng Policy Gradient có thể chậm hơn, nhưng có sự tin cậy cao do đầu ra là một giá trị xác suất.



Hình 2.15: Vòng lặp sẽ không xảy ra khi đi vào ô vuông xám¹⁵

Hàm mục tiêu. Mục tiêu của phương pháp này là đạt được một chiến lược có thể giải quyết được bài toán RL. Dựa vào đó hàm mục tiêu này có thể được kí hiệu là $J(\theta)$ với tham số θ đại diện cho trọng số cần tối ưu qua quá trình học. Một số ví dụ dưới đây có thể được dùng làm hàm mục tiêu của bài toán:

- tối ưu lợi nhuận kỳ vọng (expected return) của trạng thái khởi đầu (first state)

$$J_1(\theta) = V^{\pi_\theta}(s_1) = \mathbb{E}_{\pi_\theta}(v_1),$$

- tối ưu lợi nhuận kỳ vọng trung bình

$$J_{avV}(\theta) = \sum_s d^{\pi_\theta}(s) V^{\pi_\theta}(s),$$

¹⁴Nguồn ảnh: <https://www.davidsilver.uk/wp-content/uploads/2020/03/pg.pdf>

¹⁵Nguồn ảnh: <https://www.davidsilver.uk/wp-content/uploads/2020/03/pg.pdf>

- tối ưu phần thưởng kỳ vọng trung bình

$$J_{\text{avR}}(\theta) = \sum_s d^{\pi_\theta}(s) \sum_a \pi_\theta(s, a) \mathcal{R}_s^a,$$

ở đó, d^{π_θ} là phân phối xác suất xảy ra trạng thái bất kỳ trong không gian trạng thái. Có nhiều cách để tìm θ để tối ưu hàm mục tiêu $J(\theta)$ từ các phương pháp không dùng gradient tới các giải thuật tiến hóa hoặc có thể là phương pháp dùng gradient như SGD, Quasi-Newton.

Finite Differences. Phương pháp đầu tiên có thể sử dụng là Finite Differences. Phương pháp này ước tính đạo hàm một phần dùng công thức

$$\frac{\partial J(\theta)}{\partial \theta_k} \approx \frac{J(\theta + \varepsilon u_k) - J(\theta)}{\varepsilon}.$$

trên một khoảng nhỏ. Ở đây $u_k = (0, \dots, 1, \dots, 0)^\top$ là một “one-hot-encoded” vec-tơ thêm ε vào vào thành phần thứ k của tham số. Với tham số có n chiều, công thức này lặp lại n lần. Tuy nhiên, do hàm mục tiêu rất khó để tính, nên cách này không hiệu quả nhưng đây vẫn là cách có thể cân nhắc do sự đơn giản của nó.

Hàm score (score function). Phương pháp Finite Differences có thể sử dụng được với các hàm mục tiêu đơn giản nhưng khó tính toán với các trường hợp phức tạp, do đó hàm score sinh ra để giải quyết vấn đề. Giả sử hàm mục tiêu có thể tính được đạo hàm tại mọi điểm và có thể tính được đạo hàm của nó, đạo hàm của hàm mục tiêu sẽ biểu diễn theo hàm score. Đầu tiên, từ phép biến đổi

$$\begin{aligned} \nabla_\theta \pi_\theta(s, a) &= \pi_\theta(s, a) \frac{\nabla_\theta \pi_\theta(s, a)}{\pi_\theta(s, a)} \\ &= \pi_\theta(s, a) \nabla_\theta \log \pi_\theta(s, a). \end{aligned}$$

đúng vì $\nabla \log f(x) = \frac{1}{f(x)} f'(x)$ theo quy tắc đạo hàm của hàm hợp (chain rule). Biểu thức $\pi_\theta(s, a) \nabla_\theta \log \pi_\theta(s, a)$ được gọi là hàm score. Ưu điểm ở đây là tái sử dụng lại $\pi_\theta(s, a)$ thay vì $\nabla \pi_\theta(s, a)$ dẫn đến không còn bận tâm đến việc tính đạo hàm trực tiếp của chiến lược.

Chiến lược Gaussian (Gaussian Policy). Chiến lược Gaussian phù hợp cho các bài toán RL có không gian hành động liên tục do nó đưa ra hành động tuân theo phân phối chuẩn. Gọi $\varphi(s)$ là trường dữ liệu qua các phép biến đổi từ trạng thái, $\mu(s) = \varphi(s)^\top \theta$ là một phép biến đổi từ $\varphi(s)$, phương sai σ^2 có thể cố định hoặc tham số hóa. Xác suất xảy ra hành động trong một trạng thái bất kỳ

sẽ được quyết định theo phân phối chuẩn

$$a \sim \mathcal{N}(\mu(s), \sigma^2).$$

Lúc này, hàm score là

$$\nabla_{\theta} \log \pi_{\theta}(s, a) = \frac{(a - \mu(s))\varphi(s)}{\sigma^2},$$

trong đó a là hành động có giá trị thuộc miền liên tục tuân theo một phân phối chuẩn, a sẽ đo lường khoảng cách so với $\mu(s) = \varphi(s)^{\top} \theta$. Các bài toán lựa chọn chiến lược này thường là các bài toán không muốn giới hạn các không gian hành động của mình lại một số lượng hữu hạn do đặc tính tự nhiên của môi trường mà tác nhân đang tương tác. Có thể kể đến môi trường đường phố mà xe tự hành chạy vì môi trường này có các điều kiện vật lý và số lượng xe thay đổi theo thời gian.

Đạo hàm của chiến lược. Xét một MDP với một trạng thái và một tập hành động, hàm mục tiêu của chiến lược cho phần thưởng trung bình có thể được định nghĩa là

$$J(\theta) = \mathbb{E}_{\pi_{\theta}}[r] = \sum_{s \in \mathcal{S}} d(s) \sum_{a \in \mathcal{A}} \pi_{\theta}(s, a) \mathcal{R}_{s,a}.$$

và đạo hàm sẽ là

$$\begin{aligned} \nabla_{\theta} J(\theta) &= \sum_{s \in \mathcal{S}} d(s) \sum_{a \in \mathcal{A}} \nabla_{\theta} \pi_{\theta}(s, a) \mathcal{R}_{s,a} \\ &= \sum_{s \in \mathcal{S}} d(s) \sum_{a \in \mathcal{A}} \pi_{\theta}(s, a) \nabla_{\theta} \log \pi_{\theta}(s, a) \mathcal{R}_{s,a} \\ &= \mathbb{E}_{\pi_{\theta}}[\nabla_{\theta} \log \pi_{\theta}(s, a) \cdot r]. \end{aligned}$$

Tiếp theo phần thưởng r có thể thay bằng q -value $Q^{\pi_{\theta}}(s, a)$ để xem xét đến phần thưởng trong dài hạn

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\pi_{\theta}}[\nabla_{\theta} \log \pi_{\theta}(s, a) \cdot Q^{\pi_{\theta}}(s, a)].$$

Reinforce. Đây là thuật toán đơn giản nhất của phương pháp Policy Gradient. Nó dùng Monte Carlo để tính lợi nhuận G_t của một tập (episode) và cập nhật trọng số theo hướng của đạo hàm của chiến lược

$$\theta \leftarrow \theta + \alpha \nabla_{\theta} \log \pi_{\theta}(s_t, a_t) G_t.$$

Chú ý G_t ở đây là tổng phần thưởng bắt đầu thời điểm t cho đến thời điểm kết

thức T của một tập, G_t không giống nhau với mọi trạng thái. Các bước của giải thuật như sau:

1. Khởi tạo θ ngẫu nhiên,
2. Lưu giữ một tập kinh nghiệm của một tập tuân theo chiến lược hiện tại $\{s_1, a_1, r_2, \dots, s_{T-1}, a_{T-1}, r_T\} \sim \pi_\theta$,
3. Với mỗi bước thời gian $t \in [1, T - 1]$:
 - (a) Cập nhật trọng số θ

$$\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(s_t, a_t) G_t.$$

Lợi nhuận G_t đóng vai trò như biển báo chỉ dẫn về hướng và độ lớn cho việc cập nhật θ , G_t bắt đầu từ t và kết thúc ở T và có giá trị khác nhau ở mỗi bước thời gian trong một tập.

Actor Critic. Vấn đề với giải thuật Reinforce ở trên là nó rất chậm do phải tương tác đủ một tập và có đặc điểm phương sai cao. Actor critic ra đời nhằm giải quyết các hạn chế này, các thuật toán thuộc Actor Critic kết hợp khái niệm của Policy Gradient và lặp chính sách (Policy Iteration) và xấp xỉ giá trị (Value Approximation). Thay vì phải tính $Q^{\pi_\theta}(s, a)$ dựa trên G_t , một hàm xấp xỉ dùng mạng thần kinh nhân tạo có thể được sử dụng để tính $Q^{\pi_\theta}(s, a)$. Hàm xấp xỉ này được gọi là critic và nó làm nhiệm vụ đánh giá mức độ tốt các hành động mà actor, ở đây chính là chiến lược, đưa ra. Do đó sẽ có hai tham số cần học là w và θ . w là tham số của hàm xấp xỉ $Q_w(s, a)$, còn θ là tham số của chiến lược π đã được tham số hóa và có thể tính đạo hàm. Với sự tham gia của critic, các cập nhật trong việc học tham số sẽ có một số sự thay đổi sau

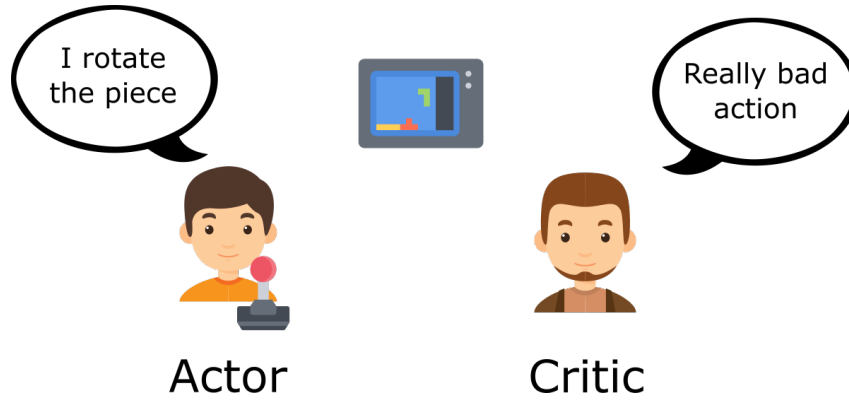
$$\begin{aligned} \nabla_\theta J(\theta) &\approx \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(s, a) Q_w(s, a)], \\ \Delta\theta &= \alpha \nabla_\theta \log \pi_\theta(s, a) Q_w(s, a). \end{aligned}$$

Giá trị $Q_w(s, a)$ có thể được học bằng cách dùng hồi quy tuyến tính hay mạng thần kinh nhân tạo cùng với Monte Carlo hoặc Temporal Difference đã được đề cập ở Tiểu mục 2.3.3. Ví dụ, nếu sử dụng TD(0) thì các bước của một giải thuật Actor critic tiến hành như sau:

1. Khởi tạo trạng thái s và tham số θ
2. Chọn hành động theo chiến lược hiện có $a \sim \pi_\theta$
3. Tại mỗi thời điểm t :

- (a) Có được phần thưởng $r = R_s^a$ và next state $s' \sim P_s^a$,
- (b) Chọn hành động $a' \sim \pi_\theta(s', a')$,
- (c) $\delta = r + \gamma Q_w(s', a') - Q_w(s, a)$,
- (d) $\theta = \theta + \alpha \nabla_\theta \log \pi_\theta(s, a) Q_w(s, a)$,
- (e) $w \leftarrow w + \beta \delta \varphi(s, a)$,
- (f) $a \leftarrow a', s \leftarrow s'$.

Một cách để tưởng tượng phương pháp Actor Critic trong thực tế là liên tưởng đến việc chơi trò chơi. Người chơi sẽ tiến hành các vòng chơi và nhận được các góp ý từ một người bạn. Người chơi đóng vai trò là actor, và người bạn là critic. Ban đầu, người chơi sẽ chơi tự do và cải thiện trình độ chơi dựa vào phản hồi của người bạn, người bạn sẽ đưa ra phản hồi dựa trên điểm số mà người chơi có. Điều này tạo thành một vòng lặp.



Hình 2.16: Liên tưởng phương pháp Actor Critic với việc chơi trò chơi¹⁶

Giảm phương sai. Vì phương pháp Actor Critic mang đặc điểm phương sai cao nên một cách để cải thiện là trừ đi một giá trị “baseline”. Thay vì định nghĩa hàm mục tiêu theo giá trị Q^{π_θ} , một hàm baseline $B(s)$ sẽ được đưa vào thay thế. Hàm này phụ thuộc vào trạng thái s và hàm mục tiêu sẽ được tối ưu theo $Q^{\pi_\theta} - B(s)$. Việc dùng hàm baseline phụ thuộc vào trạng thái sẽ không làm ảnh hưởng đến quá trình tối ưu (do nó không gây ảnh hưởng đến θ), nhưng sẽ làm giảm phương sai. Việc lựa chọn hàm baseline rất quan trọng, ở đây hàm giá trị trạng thái $V^{\pi_\theta}(s)$ hoàn toàn phù hợp với các tiêu chí đã kể trên. Như vậy, việc trừ hàm baseline khỏi hàm giá trị hành động sẽ có biểu diễn như sau

$$A^{\pi_\theta}(s, a) = Q^{\pi_\theta}(s, a) - V^{\pi_\theta}(s).$$

¹⁶Nguồn ảnh: <https://huggingface.co/blog/deep-rl-a2c>

Hàm này được gọi là hàm lợi thế (advantage function). Nó cho biết rằng mức độ tốt hơn khi chọn hành động a so với ở lại trạng thái s . Đạo hàm của hàm mục tiêu bây giờ sẽ được viết lại như sau

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\pi_{\theta}}[\nabla_{\theta} \log \pi_{\theta}(s, a) A^{\pi_{\theta}}(s, a)],$$

trong đó V sẽ được tính bằng việc tham số hóa qua một hàm xấp xỉ như mạng thần kinh nhân tạo. Tuy nhiên, khi xét đến định nghĩa của hàm giá trị hành động chính là bằng phần thưởng của hành động cộng với phần thưởng trung bình của toàn bộ các trạng thái tiếp theo có thể đi đến, công thức của hàm lợi thế sẽ thay đổi chỉ dựa vào hàm giá trị trạng thái và không còn cần tới q -value. Nếu xét đến hàm lợi thế A , TD error lúc này có dạng

$$\delta^{\pi_{\theta}} = r + \gamma V^{\pi_{\theta}}(s') - V^{\pi_{\theta}}(s),$$

và

$$\begin{aligned} \mathbb{E}[\delta^{\pi_{\theta}} | s, a] &= \mathbb{E}_{\pi_{\theta}}[r + \gamma V^{\pi_{\theta}}(s') | s, a] - V^{\pi_{\theta}}(s) \\ &= Q^{\pi_{\theta}}(s, a) - V^{\pi_{\theta}}(s) \\ &= A^{\pi_{\theta}}(s, a). \end{aligned}$$

Đạo hàm của chiến lược trở thành

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\pi_{\theta}}[\nabla_{\theta} \log \pi_{\theta}(s, a) \delta^{\pi_{\theta}}].$$

Điều này làm cho số tham số cần học nhỏ lại chỉ bao gồm tập trọng số v để xấp xỉ hàm giá trị trạng thái

$$V_v(s) \approx V_{\pi}(s).$$

PPO với hàm mục tiêu có lược bỏ (PPO-Clip). PPO (Proximal Policy Optimization) thuộc họ các giải thuật Policy Gradient, đã được đề cập ở Tiểu mục 2.3.4, nhằm giải quyết các bài toán học tăng cường không dựa trên mô hình (model-free). Thuật toán được phát triển bởi Schuman vào năm 2017 [11].

Trong phương pháp này, các tham số vẫn sẽ được tối ưu hóa bằng các giải thuật thuộc họ Gradient Descent với một chút thay đổi trong hàm mục tiêu. Đó là hàm mục tiêu sẽ có thêm một ràng buộc mềm nhằm khiến chiến lược sẽ không được điều chỉnh quá nhiều.

Trong quá trình thực hiện, hai chiến lược sẽ được duy trì. Đầu tiên là chiến

thuật hiện tại muốn điều chỉnh

$$\pi_{\theta}(a_t|s_t).$$

Thứ hai là chiến lược đã được sử dụng gần nhất để thu thập các mẫu

$$\pi_{\theta_{old}}(a_t|s_t).$$

Với ý tưởng của phương pháp Importance Sampling, chiến lược mới sẽ được đánh giá với các mẫu được thu thập từ một chiến lược cũ hơn. Điều này sẽ giúp cho việc cải thiện hiệu quả hoạt động

$$\max_{\theta} \hat{E}_t \left[\frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \hat{A}_t \right].$$

Nhưng khi điều chỉnh chiến lược hiện tại, độ chênh lệch giữa chiến lược hiện tại và chiến lược cũ ngày càng lớn. Phương sai của phép tính sẽ tăng lên. Vì vậy, giả sử cứ bốn vòng lặp, chiến lược cũ sẽ đồng bộ với chiến lược hiện tại

$$\pi_{\theta_{old}}(a_t|s_t) \leftarrow \pi_{\theta}(a_t|s_t).$$

Trong PPO, tỷ lệ giữa chiến lược mới và chiến lược cũ tính như sau

$$r_t(\theta) = \frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$$

Và một hàm mục tiêu mới được xây dựng nhằm cắt bớt hàm lợi thế đã được tính nếu chiến lược mới chênh lệch nhiều với chiến lược cũ. Hàm mục tiêu mới là:

$$\mathcal{L}_{\theta_{old}}^{\text{CLIP}}(\theta) = E_{\tau \sim \theta_k} \left[\sum_{t=0}^T \left[\min \left(r_t(\theta) \hat{A}_t^{\theta_{old}}, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t^{\theta_{old}} \right) \right] \right]$$

Nếu tỷ lệ xác suất giữa chiến lược mới và chiến lược cũ nằm ngoài phạm vi $1 - \epsilon$ và $1 + \epsilon$, hàm lợi thế sẽ bị cắt bớt. Giá trị ϵ thường được đặt là 0,2 cho các thí nghiệm trong bài báo về PPO. Điều này làm giảm sự thay đổi chiến lược quá nhiều. Dưới đây là thuật toán của giải thuật PPO với hàm mục tiêu có lược bỏ:

1. Khởi tạo tham số θ , giá trị ϵ ,
2. Ở mỗi tập (episode) $k = 0, 1, 2, \dots$:
 - (a) Thu thập tập kinh nghiệm \mathcal{D}_k dựa trên chiến lược $\pi_k = \pi(\theta_{old})$,
 - (b) Tính giá trị hàm lợi thế $\hat{A}_t^{\pi_{old}}$,

(c) Thực hiện K bước các thuật toán “gradient descent” với hàm mục tiêu

$$\mathcal{L}_{\theta_{old}}^{\text{CLIP}}(\theta) = E_{\tau \sim \theta_{old}} \left[\sum_{t=0}^T \left[\min \left(r_t(\theta) \hat{A}_t^{\theta_{old}}, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t^{\theta_{old}} \right) \right] \right],$$

(d) Cập nhật tham số θ_{old} theo θ

$$\pi_{\theta_{old}}(a_t|s_t) \leftarrow \pi_{\theta}(a_t|s_t).$$

Ví dụ sau đây trình bày các bước tính giải thuật PPO

1. Khởi tạo tham số θ là tham số của hàm tuyến tính có đầu ra là giá trị trung bình của hành động với đầu vào là trạng thái của tác nhân $\bar{a} = \theta s$ với $s \in \mathbb{R}$, giá trị $\epsilon = 0.2$ và $K = 3$

2. Ở tập (episode) $k = 1$:

(a) Tập kinh nghiệm $\mathcal{D}_1 = \{(1, 0.7, 0.6, 4, 3)\}$ chứa các bộ $(s, a, p(s|a), \hat{v}(s), r(s, a))$ được thu thập theo $\pi_{\theta_{old}}$,

(b) Giá trị $\hat{A}^{\theta_{old}} = G(s) - \hat{v}(s) = 3 - 4 = -1$,

(c) Thực hiện vòng lặp sau ba lần:

i. Tính lại $p(s, a)$ theo π_{θ} , ví dụ ở đây giá trị là 0.8. Từ đó suy ra $r(\theta) = \frac{0.8}{0.6} = \frac{4}{3}$,

ii. Vì hàm mất mát cần tìm giá trị nhỏ nhất nên cần đặt dấu - phía trước

$$\mathcal{L} = -\min\left(\frac{4}{3} * (-1), \text{clip}\left(\frac{4}{3}, 1 - 0.2, 1 + 0.2\right) * (-1)\right) = \frac{4}{3},$$

iii. Cập nhật θ theo giá trị hàm mất mát trên,

(d) Cập nhật θ_{old} theo θ như sau

$$\theta_{old} = \theta$$

.

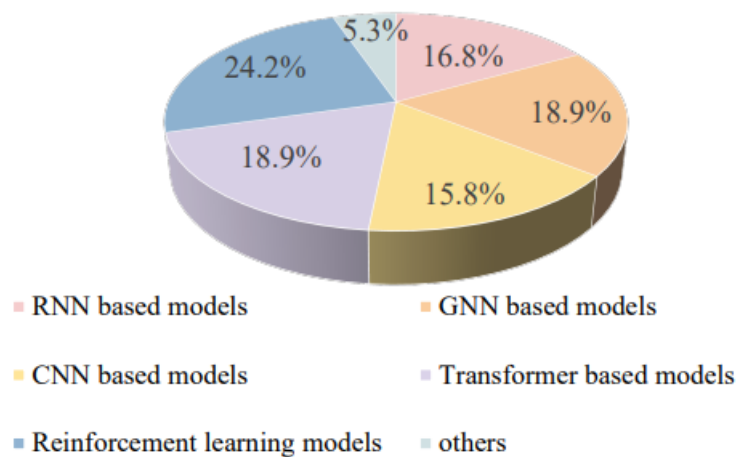
CHƯƠNG 3

CÁC CÔNG TRÌNH LIÊN QUAN

3.1 Tổng quan các quá trình nghiên cứu trên thế giới về đề tài

Dự đoán thị trường chứng khoán đã và đang luôn là mong muốn được khao khát từ giới đầu tư. Tính đến nay, đã có rất nhiều nghiên cứu được thực hiện về đề tài này. Trong các nghiên cứu trước đây về dự đoán chứng khoán, các phương pháp truyền thống như SVM (Support Vector Machine), hồi quy và KNN (K-Nearest Neighbors) đã được sử dụng rộng rãi. Do khả năng xấp xỉ phi tuyến mạnh, SVM có thể được áp dụng cả trong bài toán phân loại (Support Vector Classifier - SVC) và hồi quy (Support Vector Regression - SVR) [12]. Trong [13], các tác giả đã tìm hiểu làm thế nào để dự đoán chỉ số cổ phiếu bằng cách sử dụng SVR để tìm hiểu mối quan hệ giữa các chỉ số kỹ thuật và giá. Tuy nhiên, với sự tiến bộ của mạng thần kinh, đã có một sự thay đổi trong việc sử dụng các mạng này trong nghiên cứu dự đoán chứng khoán. Jasemi, Kimiagari và Memariani đã sử dụng MLP (Multilayer Perceptron) [14] để tìm các mẫu ẩn (hidden patterns) trong biểu đồ nến Nhật Bản với trọng tâm là khám phá các tín hiệu đảo chiều về giá, được biểu thị trong phân tích nến bằng một số mẫu trên biểu đồ, chẳng hạn như sao mai, búa ngược, harami, nhấn chìm, và một số mẫu khác. Các tín hiệu đảo chiều này thể hiện các điểm mua hoặc bán theo lý thuyết phân tích nến. Ballings [15] tập trung vào các mô hình phân loại cho các phương pháp tập hợp (ensemble methods) và các mô hình phân loại đơn lẻ trong dự đoán giá cổ phiếu .

Khi học sâu bắt đầu phát triển, nhiều công trình nghiên cứu đã chuyển từ các phương pháp truyền thống sang các cách tiếp cận cao cấp hơn như Recurrent Neural Network (RNN), Long Short-term Memory (LSTM), Graph Neural Network (GNN) và Convolutional Neural Network (CNN) [16]. Gần đây, các nhà nghiên cứu đã bắt đầu khám phá và ứng dụng các mô hình mới hơn dựa trên Transformer hay học tăng cường trong việc nghiên cứu bài toán dự đoán chứng khoán. Hình 3.1 từ [16] cho thấy được một cái nhìn tổng quan về các mô hình được sử dụng gần đây trong chủ đề nghiên cứu. Có thể thấy rằng ngoài RNN, mô hình thường được sử dụng trong bài toán về chuỗi thời gian, chiếm 18.9% còn có mô hình thuộc họ Transformer hay sử dụng các thuật toán của RL lần lượt chiếm 18.9% và 24.2%.



Hình 3.1: Phân bố các bài báo theo các mô hình [16]

Dự đoán thị trường chứng khoán là một chủ đề rộng, ngày nay nó có thể được chia thành bốn bài toán con được các nhà nghiên cứu tập trung vào nhiều nhất. Bao gồm:

- **Dự đoán giá.** Mục tiêu của bài toán này là dự đoán giá tương lai của cổ phiếu và tài sản khác được giao dịch trên các sàn.
- **Dự đoán xu hướng giá.** Khác với bài toán trên là dự báo một giá cụ thể trong tương lai, bài toán này không cần dự đoán chính xác mà chỉ cần đưa ra dự đoán về xu hướng giá có thể trong tương lai. Xu hướng giá có thể chia làm hai loại là lên giá hoặc xuống giá, một số bài báo thêm vào xu hướng thứ ba là không biến động.
- **Quản lý danh mục.** Bài toán này liên quan đến việc lựa chọn và điều chỉnh danh mục cổ phiếu một cách có tính toán tùy theo điều kiện của thị trường nhằm mục đích sau cùng là đạt được các lợi ích tài chính. Mục tiêu của nó là

xây dựng một chiến lược phân bổ thông minh danh mục đầu tư cùng lúc giảm thiểu rủi ro.

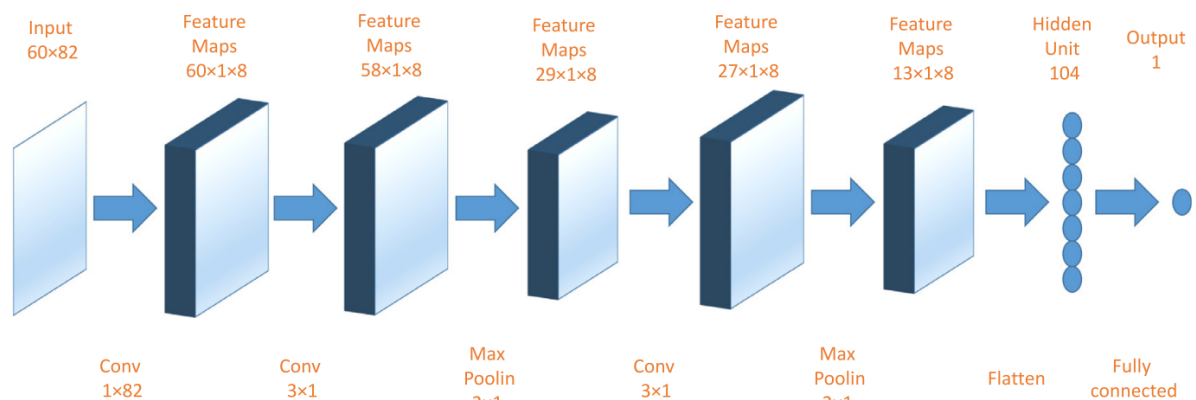
- **Các chiến lược giao dịch.** Chiến lược giao dịch là một tập các quy tắc được xây dựng từ trước nhằm đưa ra các quyết định giao dịch bao gồm mua, giữ hoặc bán cổ phiếu. Chiến lược đầu tư phụ thuộc vào rất nhiều yếu tố từ sở thích đầu tư (giá trị hoặc tăng trưởng), vốn hóa thị trường, phân tích cơ bản, phân tích kỹ thuật ...

Có thể thấy, đề tài này đã và đang được nghiên cứu rất nhiều ở các nơi trên thế giới và vẫn chưa có dấu hiệu dừng lại. Trong phần tiếp theo, sinh viên sẽ trình bày qua một số công trình mà sinh viên chọn là tiêu biểu và có liên quan đến luận văn, để ta có thể thấy các hướng tiếp cận xử lý bài toán đến thời điểm hiện tại.

3.2 Hướng tiếp cận dựa trên CNN

Các mô hình thuộc họ CNN được sử dụng rộng rãi trong các bài toán về thị giác máy tính (Computer Vision - CV) và xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP). Các mô hình tích chập trong các bài toán này thường dùng các bộ lọc hai chiều do dữ liệu đầu vào là hình ảnh. Tuy nhiên, trong chủ đề dự đoán chứng khoán, CNN được xử lý với dữ liệu là các chuỗi thời gian, tức là dữ liệu một chiều. Do sự khác biệt này, CNN được sử dụng thường dùng các bộ lọc một chiều có kích thước phù hợp.

Các bộ lọc được sử dụng nhằm thu được các đầu ra hữu ích, do đó mô hình CNN được các tác giả sử dụng để bắt được các tín hiệu có ích ẩn trong biến động về giá cả. Hoseinzad [17] đề xuất một mô hình CNN sử dụng các bộ lọc một chiều ở các lớp khác nhau nhằm trích xuất đầu ra hữu ích trong việc phân tích biến động giá cả. Hình 3.2 minh họa về kiến trúc của mô hình được các tác giả lựa chọn.



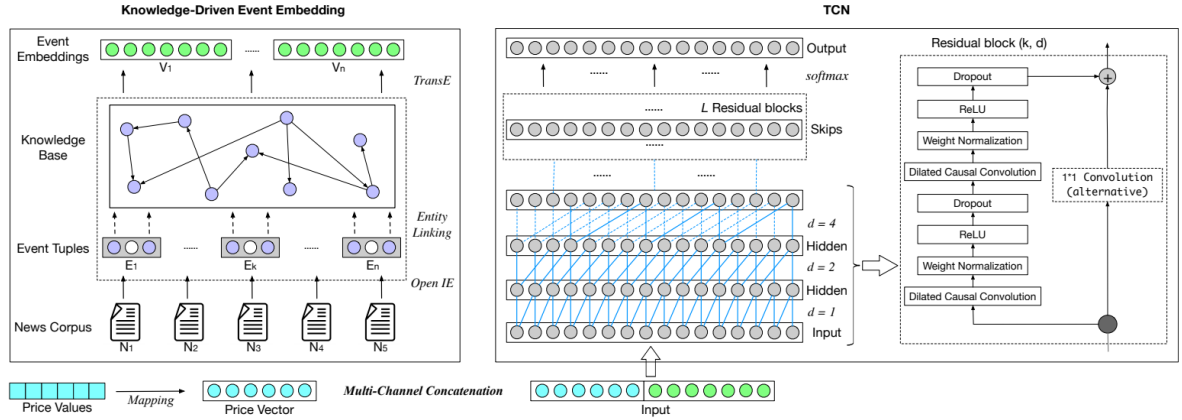
Hình 3.2: Trực quan hóa của mô hình 2D-CNNpred (nguồn: [17])

Có thể thấy được điều đó qua hai lớp tích chập đầu tiên của mô hình. Lớp tích chập thứ nhất dùng bộ lọc có kích thước là 1×82 với mục tiêu tạo ra đặc trưng mới có độ dài hơn. Trong trường hợp các đặc trưng mang thông tin thừa, nó cũng có thể loại bỏ các đặc trưng đó bằng cách chọn các giá trị 0 tương ứng với vị trí trên bộ lọc. Ở lớp tích chập thứ hai, với mục tiêu trích xuất các trường dữ liệu mới dựa trên liên kết về thời gian có thể có, các tác giả dùng một lớp tích chập khác có kích thước 3×1 . Mô hình trên được họ gọi là 2D-CNNpred, do dữ liệu đầu vào chỉ có hai chiều. Ngoài ra, các tác giả cũng đề cập đến mô hình 3D-CNNpred. Mô hình này giả sử có mối liên hệ giữa các thị trường với nhau, do đó để đưa ra dự đoán giá của một thị trường, các thông tin của các thị trường khác cần phải được xem xét. Sự khác biệt này dẫn đến một chút thay đổi trong biểu diễn dữ liệu và kiến trúc mô hình. Biểu diễn dữ liệu bây giờ là một tensor 3 chiều, kiến trúc mô hình CNN cũng sẽ thay đổi tương ứng.

Market-Model	Technical	CNN-cor	PCA+ANN	2D-CNNpred	3D-CNNpred
S&P 500	0.5627	0.5723	0.5165	0.5408	0.5532
DJI	0.5518	0.5253	0.5392	0.5562	0.5612
NASDAQ	0.5487	0.5498	0.5312	0.5521	0.5576
NYSE	0.5251	0.5376	0.5306	0.5472	0.5592
RUSSELL	0.5665	0.5602	0.5438	0.5463	0.5787

Hình 3.3: Kết quả độ đo Macro Average F-measure của mô hình CNNpred (nguồn: [17])

Sử dụng các thông tin về các mối liên kết của các ngành nghề trong nền kinh tế, một số mô hình tích hợp các đồ thị thông tin (knowledge graph) và CNN để cải thiện hiệu suất. Một mô hình theo cách này là Knowledge-Driven Temporal Convolutional Network (KDTCN) [18]. Mô hình này sử dụng Open IE [19] để trích xuất các sự kiện liên quan đến đồ thị thông tin và đưa ra các dự đoán về cổ phiếu. Một vấn đề thông thường với bộ lọc một chiều là rò rỉ dữ liệu (data leakage), vì thông tin từ thời điểm $t-1$ và $t+1$ có thể ảnh hưởng đến dữ liệu ở thời điểm t . Để giải quyết vấn đề, mô hình KDTCN sử dụng *causal convolution*, nó chỉ dùng thông tin từ các bước thời gian hiện tại và quá khứ trong các lớp trước. Mô hình này đã được kiểm chứng sự hiệu quả trong việc giải thích các thay đổi giá đột ngột bằng cách trích xuất các đặc trưng quan trọng trong dữ liệu lịch sử giá [18].



Hình 3.4: Kiến trúc mô hình KDTCN (nguồn: [18])

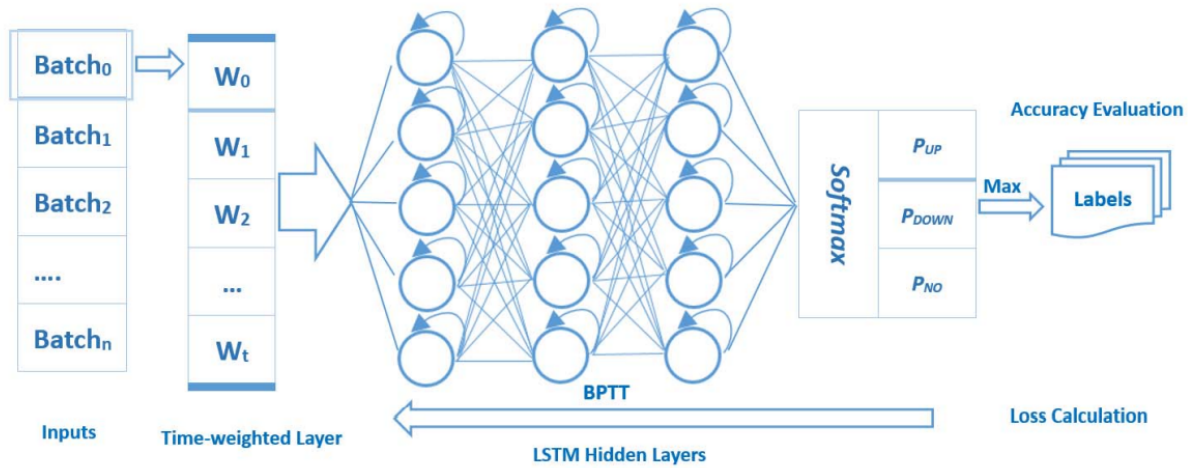
Việc tích hợp CNN và LSTM có thể nâng cao khả năng dự đoán chuỗi thời gian hơn nữa. Lu [20] đã giới thiệu mô hình CNN-LSTM để dự báo giá đóng cửa cổ phiếu hàng ngày, trong đó thành phần CNN trích xuất các đặc điểm từ chuỗi thời gian dữ liệu lịch sử 10 ngày và thành phần LSTM đưa ra dự đoán giá. Trong nghiên cứu tiếp theo, Lu [21] đề xuất mô hình CNN-BiLSTM-AM kết hợp cơ chế tập trung để nắm bắt các biến động cổ phiếu trong lịch sử có ảnh hưởng trên dữ liệu lịch sử giá và cải thiện hiệu suất của mô hình dựa trên CNN. Wang [22] ứng dụng mô hình CNN-BiLSTM để dự đoán giá đóng cửa cổ phiếu và cải thiện hiệu suất của mô hình bằng cách thêm hàm kích hoạt tanh vào cổng đầu ra của Bi-LSTM. Ngoài ra, việc sử dụng GRU gần đây cũng đã được chứng minh là có hiệu quả trong các công trình nghiên cứu. Zhou, Zhou và Wang [23] đã đề xuất một mô hình dự đoán thị trường chứng khoán tổng hợp bao gồm CNN và GRU hai chiều (Bidirectional GRU). CNN chịu trách nhiệm xử lý và trích xuất các đặc trưng mới, trong khi GRU chịu trách nhiệm xử lý dữ liệu chuỗi thời gian. Họ sử dụng giá đóng cửa của thị trường chứng khoán làm đầu ra của mô hình và tất cả các dữ liệu khác làm đầu vào và thu được kết quả tốt hơn các mô hình cơ bản khác.

3.3 Hướng tiếp cận dựa trên RNN

RNN [24] là mô hình học sâu có khả năng xử lý hiệu quả bài toán có dữ liệu tuần tự. Do dữ liệu lịch sử giá thường được biểu diễn là các chuỗi thời gian nên RNN chính là lựa chọn lý tưởng cho việc đưa ra dự đoán dựa vào thông tin, lịch sử giá. Không chỉ RNN, các mô hình được phát triển sau này như LSTM (Long Short Term Memory) [25] hay GRU (Gate Recurrent Unit) đều được ưa chuộng và sử dụng rộng rãi do chúng có thể khắc phục nhược điểm của RNN trong việc ghi

nhớ các thông tin dài hạn.

Zhao [26] đề xuất một mô hình LSTM có trọng số theo thời gian (Time-Weighted LSTM). Thứ nhất, thay vì xem dữ liệu là độc lập, họ sử dụng hàm trọng số thời gian để gán trọng số một cách cẩn thận cho dữ liệu theo độ gần về thời gian của chúng với dữ liệu được dự đoán. Tiếp đến, các định nghĩa về xu hướng chứng khoán được đưa ra chính thức bằng cách tham khảo các lý thuyết tài chính và thực tiễn. Cuối cùng, LSTM đã được tùy chỉnh để khám phá các phụ thuộc cơ bản về thời gian trong dữ liệu. Các thử nghiệm về các hàm trọng số cho thời gian khác nhau cho thấy mối quan hệ giữa tầm quan trọng của dữ liệu và chuỗi thời gian của là không cố định. Thay vào đó, nó nằm trong hàm tuyến tính và bậc hai, gần như là một hàm tựa tuyến tính. Được trang bị với hàm trọng số này, LSTM đã vượt trội hơn các mô hình khác và có thể tổng quát hóa cho các chỉ số chứng khoán khác.

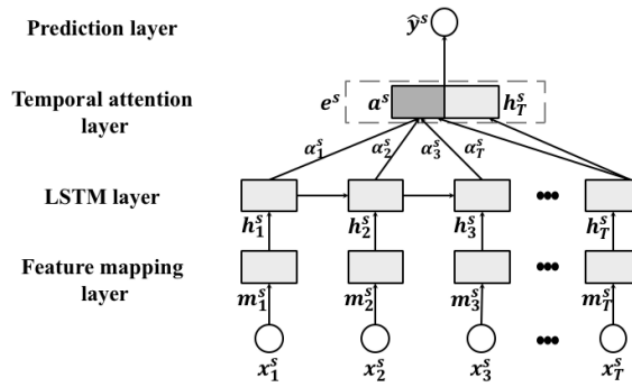


Hình 3.5: Kiến trúc mô hình Time-Weighted LSTM (nguồn: [26])

Trên thị trường chứng khoán, số lượng dữ liệu được thu thập khi tính theo ngày rất ít, điều này có thể dẫn đến vấn đề khớp quá mức (overfitting). Nguyen và Yoon [27] đề xuất một khung mới, được đặt tên là chuyển tiếp sâu với thông tin liên quan chứng khoán (deep transfer with related stock information - DTRSI), tận dụng mạng thần kinh sâu và học chuyển tiếp. Đầu tiên, một mô hình cơ sở sử dụng các ô LSTM được huấn luyện trước dựa trên một lượng lớn dữ liệu thu được từ các cổ phiếu khác nhau để tối ưu hóa các tham số huấn luyện ban đầu. Tiếp đến, mô hình cơ sở được tinh chỉnh bằng cách sử dụng một lượng nhỏ dữ liệu từ cổ phiếu mục tiêu và các đặc trưng khác nhau (được xây dựng dựa trên mối quan hệ giữa các cổ phiếu) để nâng cao hiệu suất. Kết quả thử nghiệm chứng minh tính hiệu quả của việc học chuyển tiếp và sử dụng thông tin mối quan hệ chứng khoán

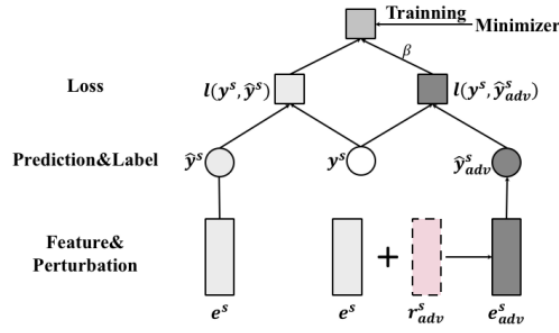
trong việc giúp cải thiện hiệu suất của mô hình và phương pháp đề xuất cho thấy hiệu suất vượt trội (so với các mô hình dự đoán khác) về độ chính xác.

Huấn luyện đối nghịch có thể mô phỏng biến động giá cổ phiếu bằng cách đưa ra các nhiễu loạn, điều này có thể nâng cao tính chính xác của dự đoán biến động cổ phiếu. Feng [28] đề xuất một phương pháp kết hợp Attention LSTM và đào tạo đối nghịch để dự đoán diễn biến thị trường chứng khoán. Tính hợp lý của việc đào tạo đối nghịch ở đây là các đặc điểm đầu vào để dự đoán cổ phiếu thường dựa trên giá cổ phiếu, về bản chất là một biến ngẫu nhiên và liên tục thay đổi theo thời gian. Do đó, việc đào tạo thông thường với các đặc trưng tĩnh (ví dụ: giá đóng cửa) có thể dễ dàng khớp dữ liệu quá mức, không đủ để có được mô hình đáng tin cậy. Để giải quyết vấn đề này, họ đề xuất thêm các nhiễu loạn để mô phỏng tính ngẫu nhiên của biến giá và huấn luyện mô hình để hoạt động tốt dưới các nhiễu loạn nhỏ nhưng có chủ ý. Tuy nhiên, dữ liệu chứng khoán thường được đưa vào mô hình một cách tuần tự, việc thêm các thay đổi vào các trường dữ liệu trên toàn bộ miền thời gian khá tốn công sức. Ngoài ra, việc thêm như vậy có thể tạo ra các tương tác không mong có thể dẫn đến kết quả không thể kiểm soát. Để giải quyết mối lo ngại này, các thay đổi này sẽ được thêm vào các trường đầu ra ở các tầng sâu của mô hình. Vì hầu hết các mô hình học sâu đều biểu diễn trừu tượng dữ liệu ở các tầng sâu nên kích thước của chúng thường nhỏ hơn nhiều so với kích thước đầu vào. Trong bài báo này, họ sẽ sử dụng phương pháp này trên mô hình Attention LSTM.



Hình 3.6: Kiến trúc của mô hình ALSTM nhằm học đầu vào cho việc huấn luyện đối nghịch [28]

Dữ liệu ở tầng ẩn cuối cùng của ALSTM sẽ được thêm vào các thay đổi mà nó sẽ được tối ưu động để trở thành đầu vào cho việc huấn luyện đối nghịch.



Hình 3.7: Kiến trúc của mô hình sử dụng huấn luyện đối nghịch (nguồn: [28])

Các thử nghiệm mở rộng trên hai tập dữ liệu chứng khoán được giao dịch trong thực tế cho thấy rằng phương pháp vượt trội hơn giải pháp của Xu và Chen [29] với mức cải thiện tương đối trung bình là 3,11% về độ chính xác.

Method	ACL18		KDD17	
	Acc	MCC	Acc	MCC
MOM	47.01±—	-0.0640±—	49.75±—	-0.0129±—
MR	46.21±—	-0.0782±—	48.46±—	-0.0366±—
LSTM	53.18±5e-1	0.0674±5e-3	51.62±4e-1	0.0183±6e-3
ALSTM	54.90±7e-1	0.1043±7e-3	51.94±7e-1	0.0261±1e-2
StockNet	54.96±—	0.0165±—	51.93±4e-1	0.0335±5e-3
Adv-ALSTM	57.20±—	0.1483±—	53.05±—	0.0523±—
RI	4.02%	42.19%	2.14%	56.12%

Hình 3.8: Kết quả trên thang đo độ chính xác và MCC của mô hình Adv-ALSTM và một số mô hình khác (nguồn: [28])

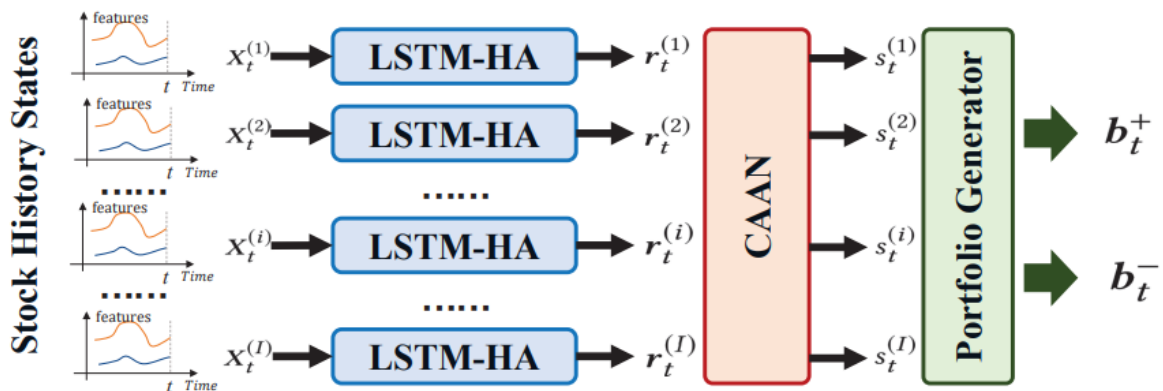
3.4 Hướng tiếp cận dựa trên học tăng cường

Mục đích sau cùng của bất kỳ bài toán dự đoán chứng khoán nào đều là các lợi ích tài chính. Với các bài toán dự đoán giá hay xu hướng giá, các nhà nghiên cứu sẽ dựa vào đó để lập ra các quy tắc giao dịch để thu về lợi nhuận. Sau này, thay vì thiết lập các quy tắc dựa vào các mô hình dự đoán, họ quyết định xây dựng mô hình tự động tạo ra các quy tắc phù hợp. Học tăng cường có thể làm được điều đó và nó được sử dụng rất nhiều trong các bài toán về quản lý danh mục đầu tư hay xây dựng chiến lược giao dịch tự động.

Các giải thuật thuộc không có mô hình (model-free) trong học tăng cường là một hướng phát triển rất tốt trong những thập kỷ gần đây, nơi tác tử tương tác trực tiếp với môi trường. Các thuật toán như tham số hóa chiến lược, Q-learning hay các dạng lai được sử dụng rộng rãi trên thị trường tài chính. huật

toán REINFORCE đã thu hút được sự chú ý trong lĩnh vực giao dịch tài chính nhờ khả năng được tối ưu hóa thông qua tăng dần độ dốc (gradient descent). Liang [30] đã đánh giá tính hiệu quả của ba thuật toán RL khác nhau, bao gồm DDPG, PPO và REINFORCE, trong môi trường huấn luyện đối nghịch và nhận thấy rằng REINFORCE hoạt động tốt nhất. Những thử nghiệm này được thực hiện trên thị trường chứng khoán Trung Quốc và các tác giả cho rằng phương pháp dựa trên tham số hóa chiến lược đặc biệt phù hợp với các kịch bản tài chính. Ngoài ra, việc kết hợp thông tin lịch sử vào trạng thái bằng cách sử dụng tham số hóa chiến lược với mạng hồi quy cũng là một cách tiếp cận đầy hứa hẹn. Jiang, Xu và Liang [31] đã phát triển khung model-free để quản lý danh mục đầu tư kết hợp CNN, RNN và LSTM và được xây dựng trên Deterministic Policy Gradient (DPG). Khung này đã được thử nghiệm trên thị trường tiền điện tử và được chứng minh là hoạt động tốt hơn các phương pháp khác.

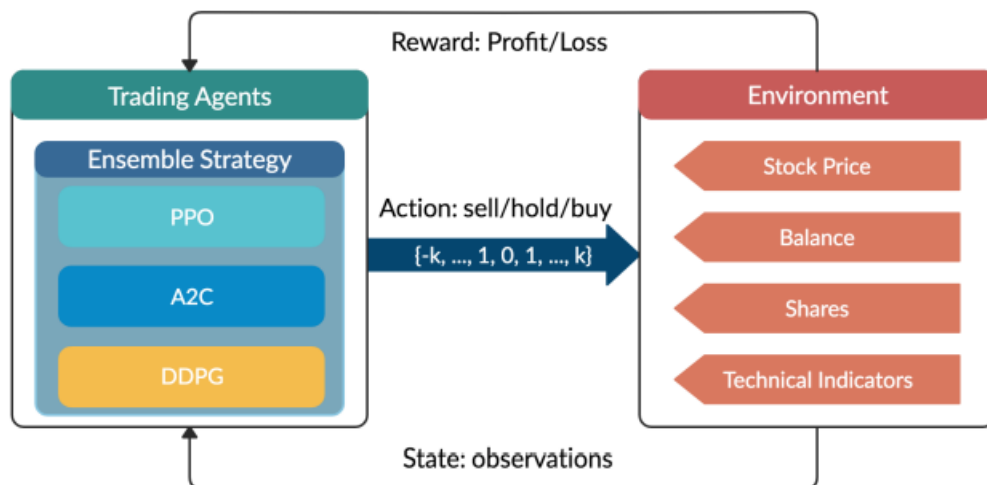
Một số nghiên cứu đã tìm cách cải thiện hiệu suất trong giao dịch định lượng bằng cách sử dụng thông tin về mối quan hệ giữa các cổ phiếu khác nhau thay vì chỉ tập trung vào việc cải thiện thuật toán RL. Wang [32] đã phát triển phương pháp AlphaStock, sử dụng hướng tiếp cận tham số hóa chiến lược theo tỉ lệ Sharpe để giải quyết các thách thức trong quản lý danh mục đầu tư như cân bằng lợi nhuận và rủi ro nhằm tránh tổn thất nghiêm trọng. Phương pháp này tổng hợp thông tin về mối quan hệ giữa các tài sản tài chính trong danh mục. Thuật toán đã được thử nghiệm ở cả thị trường chứng khoán Mỹ và Trung Quốc và kết quả cho thấy tính hiệu quả, mạnh mẽ và khả năng tổng quát hóa tốt. Nó có xu hướng lựa chọn những cổ phiếu có xu hướng tăng giá và độ biến động thấp.



Hình 3.9: Bộ khung của mô hình AlphaStock (nguồn: [32])

Ngoài các nghiên cứu về chủ đề quản lý danh mục, học tăng cường cũng được sử dụng trong việc học các chiến lược giao dịch tự động. Yang [33] đề xuất một

chiến lược kết hợp sử dụng các giải thuật học tăng cường sâu để học một chiến lược giao dịch chứng khoán bằng cách tối đa hóa lợi nhuận đầu tư. Họ huấn luyện một tác tử học tăng cường sâu và có được chiến lược giao dịch kết hợp bằng cách sử dụng ba thuật toán actor-critic: Proximal Policy Optimization (PPO), Advantage Actor Critic (A2C), và Deep Deterministic Policy Gradient (DDPG). Chiến lược kết hợp kế thừa và tích hợp các thuộc tính tốt nhất của ba thuật toán, từ đó điều chỉnh mạnh mẽ theo các tình huống thị trường khác nhau.



Hình 3.10: Tổng quan mô hình giao dịch chứng khoán bằng học tăng cường (nguồn: [33])

Kết quả kiểm tra thuật toán được ghi lại trên 30 cổ phiếu Dow Jones. Hiệu suất của tác tử với các thuật toán học tăng cường khác nhau được đánh giá và so sánh với cả chỉ số Trung bình Công nghiệp Dow Jones và chiến lược phân bổ danh mục đầu tư có phương sai tối thiểu truyền thống (traditional min-variance portfolio allocation strategy). Chiến lược kết hợp được chứng minh là vượt trội hơn ba thuật toán riêng lẻ và hai mô hình khác về mặt lợi nhuận điều chỉnh theo rủi ro (risk-adjusted return) được đo bằng tỷ lệ Sharpe.

(2016/01/04-2020/05/08)	Ensemble (Ours)	PPO	A2C	DDPG	Min-Variance	DJIA
Cumulative Return	70.4%	83.0%	60.0%	54.8%	31.7%	38.6%
Annual Return	13.0%	15.0%	11.4%	10.5%	6.5%	7.8%
Annual Volatility	9.7%	13.6%	10.4%	12.3%	17.8%	20.1%
Sharpe Ratio	1.30	1.10	1.12	0.87	0.45	0.47
Max Drawdown	-9.7%	-23.7%	-10.2%	-14.8%	-34.3%	-37.1%

Hình 3.11: Kết quả so sánh hiệu suất của mô hình [33]

CHƯƠNG 4

PHƯƠNG PHÁP ĐỀ XUẤT

4.1 Phát biểu bài toán

Thị trường chứng khoán mang đặc điểm của sự vô định và luôn thay đổi không ngừng. Giá cả trên thị trường chứng khoán biến động rất lớn, bị ảnh hưởng bởi rất nhiều yếu tố khác nhau: từ kết quả tài chính của các doanh nghiệp, chính sách kinh tế quốc gia, tác động của các cổ đông lớn hay các nhận định của các chuyên gia tài chính lên công chúng và rất nhiều yếu tố khác.

Ở luận văn này, trong bốn bài toán được tập trung nhiều nhất đã được đề cập ở Tiểu mục 3.1, sinh viên quyết định chọn tập trung vào bài toán dự đoán xu hướng giá cổ phiếu. Tuy nhiên, khác với các bài báo trước, bài toán chủ yếu được nghiên cứu bằng cách sử dụng các mô hình học sâu và kiến thức về học giám sát. Ở luận văn này, sinh viên sẽ mô tả và nghiên cứu bài toán dưới góc nhìn của một bài toán học tăng cường. Qua luận văn này, sinh viên thực hiện mong muốn có thể đóng góp và mang lại một hướng giải quyết khác biệt so với các công trình được nghiên cứu trước đó.

Ở luận văn này, sinh viên thực hiện sẽ định nghĩa đầu vào và đầu ra như sau:

- **Đầu vào:** Lịch sử giá của 30 ngày trong quá khứ của ngày cần dự đoán từ tập dữ liệu được thu thập từ 04-06-2018 đến 24-02-2023 và trải qua các bước xử lý cần thiết trước khi đưa qua mô hình.
- **Đầu ra:** Mô hình sẽ phân tích các mối liên hệ có trong đầu vào và đưa ra dự

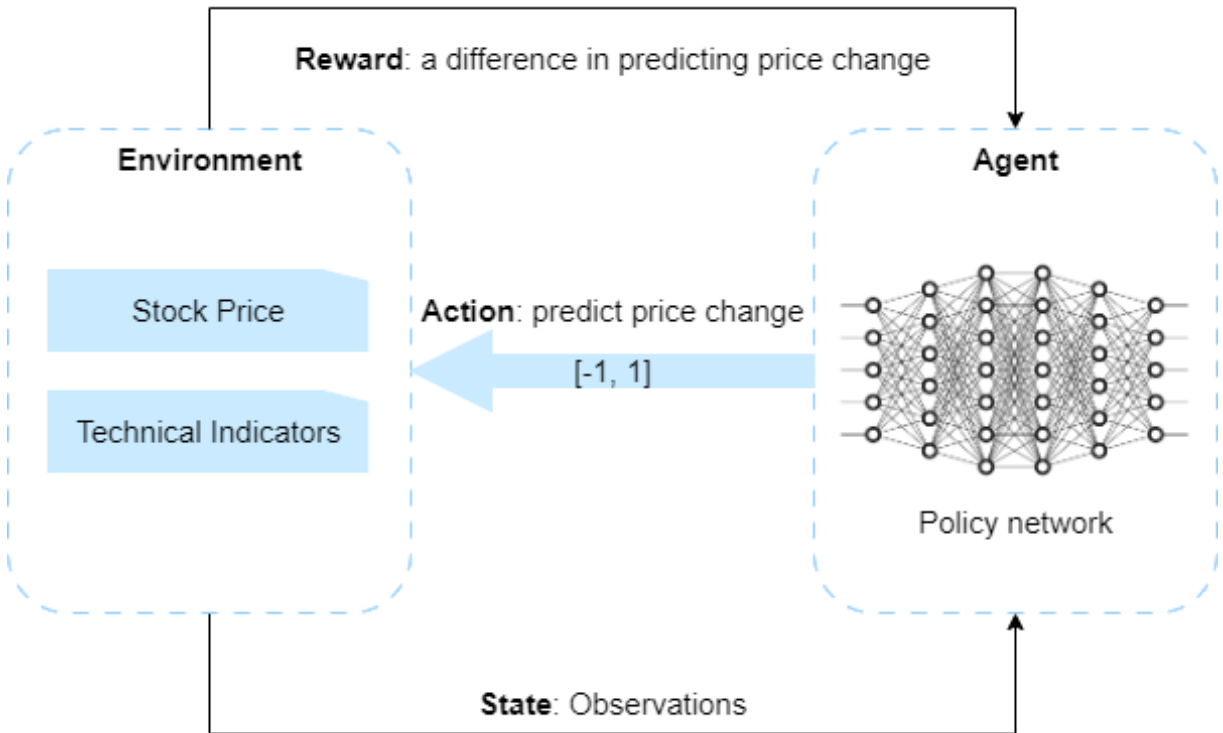
đoán liệu giá đóng cửa của cổ phiếu trong ngày mai là tăng hay giảm?

4.2 Kiến trúc mô hình đề xuất

Việc học một mô hình dự đoán chứng khoán nhằm thu về lợi ích tài chính sử dụng học tăng cường có thể biểu diễn là một quá trình ra quyết định, biểu diễn bởi một MDP đã được đề cập ở Tiểu mục 2.3.2. Ở MDP này:

- Trạng thái s_t có thể chứa các thông tin về lịch sử giá, phân tích cơ bản, phân tích kỹ thuật ở thời điểm t ,
- Hành động a_t ở thời điểm t có thể là một vec-tơ rời rạc hay liên tục mang ý nghĩa mức độ tự tin trong việc đoán xu hướng giá,
- Phần thưởng $r(s_t, a_t, s_{t+1})$ biểu thị cho mức độ đúng đắn khi thực hiện hành động a_t ở trạng thái s_t và sẽ đến được trạng thái s_{t+1} ,
- Chiến lược $\pi(a_t|s_t)$ sẽ được huấn luyện để sinh ra phân bố xác suất sẽ chọn hành động a_t như thế nào khi ở trạng thái s_t nhằm tối ưu lợi nhuận kỳ vọng thu được.

Sinh viên thực hiện sẽ trình bày cách định nghĩa chi tiết về cách định nghĩa không gian trạng thái, hành động, hàm phần thưởng ... trong Mục 4.4. Dưới đây là tổng quan mô hình dự đoán chứng khoán khi dùng học tăng cường:



Hình 4.1: Tổng quan của mô hình dự đoán chứng khoán sử dụng học tăng cường

4.3 Chuẩn bị dữ liệu

4.3.1 Thu thập dữ liệu

Thách thức trong việc thu thập dữ liệu: Chất lượng của dữ liệu có tác động rất lớn đến chất lượng của các mô hình học máy, đặc biệt là mô hình học sâu. Do đó, việc tìm kiếm một tập dữ liệu tốt cho luận văn là một trong những nhiệm vụ quan trọng nhất đã khiến sinh viên thực hiện phải đắn đo trong một thời gian dài. Có nhiều cách tiếp cận khác nhau trong việc tìm kiếm một tập dữ liệu tài chính. Cách tiếp cận đầu tiên là sử dụng các tập dữ liệu đã có sẵn, được mở trên các website trực tuyến dành cho các cuộc thi về Học máy như Kaggle. Đây là cách đầu tiên mà sinh viên thực hiện đã xem xét do sự thuận tiện của nó. Ưu điểm của cách này là dễ dàng tiếp cận, dữ liệu đã được xử lý và đã đảm bảo được độ tin cậy. Khuyết điểm là tập dữ liệu thường có kích cỡ nhỏ về thời gian thu thập cũng như số lượng cổ phiếu được cố định sẵn, định dạng của dữ liệu sẽ không được đảm bảo phù hợp để sử dụng trong luận văn. Cách tiếp cận thứ hai cũng đã được xem xét là mua dữ liệu từ các sàn giao dịch chứng khoán và các tổ chức cung cấp dữ liệu tài chính. Cách tiếp cận này rất tuyệt vời vì các sàn chứng khoán và các tổ chức thường có cơ sở hạ tầng đủ mạnh để thu thập và ghi lại dữ liệu tần suất cao có thể tính theo phút, bao gồm mọi giao dịch được ghi lại trên thị trường chứng khoán và còn chưa xét tính chính xác, độ tin cậy cao khi mua trực tiếp từ các nguồn cung cấp uy tín. Tuy nhiên, vì vẫn còn là sinh viên đại học, việc mua dữ liệu quá tốn kém và không khả thi về mặt kinh tế. Có thể nói rằng việc tìm kiếm dữ liệu là một thách thức rất lớn và đòi hỏi nhiều công sức. Cách tiếp cận sau cùng, cũng là cách sinh viên thực hiện sẽ chọn, là tự thu thập dữ liệu từ các trang web chuyên đưa tin tức về tài chính và được truy cập nhiều từ người dùng qua công cụ tìm kiếm trên các trình duyệt web. Ưu điểm của cách này là vẫn đảm bảo được độ tin cậy do lấy từ các trang web uy tín, việc tự do lựa chọn được khoảng thời gian, số lượng cổ phiếu cần thu thập. Nhưng nó vẫn tồn tại những khuyết điểm, vì các trang web được truy cập tự do trên Internet một cách miễn phí nên dữ liệu thu thập sẽ không ở mức chi tiết theo phút mà chỉ được cung cấp theo ngày.

Khi chúng tôi quyết định là sẽ tự thu thập dữ liệu, đầu tiên chúng tôi sẽ cố gắng thu thập dữ liệu từ các nguồn uy tín nhất có thể kể tới như Yahoo Finance¹, Google Finance². Tuy nhiên, các nguồn này lại không cung cấp các dữ liệu cổ phiếu của thị trường chứng khoán Việt Nam. Vì vậy chúng tôi quyết định sẽ thu thập

¹<https://finance.yahoo.com/>

²<https://www.google.com/finance/>

dữ liệu từ các nguồn cung cấp dữ liệu tài chính miễn phí, ở đây chúng tôi chọn fingroup³. Một trong những lý do chúng tôi chọn nguồn này ngoài vì miễn phí chính là dữ liệu cổ phiếu đã được điều chỉnh khi trải qua đợt chia cổ tức cho các nhà đầu tư của các công ty. Ở luận văn này, chúng tôi sẽ thu thập 30 cổ phiếu từ chỉ số VN30 trong khoảng thời gian từ 04-06-2018 đến 24-02-2023.

³<https://fiingroup.vn/>

CHƯƠNG 4. PHƯƠNG PHÁP ĐỀ XUẤT

Bảng 4.1: Danh sách 30 cổ phiếu của chỉ số VN30 đầu năm 2023

Mã chứng khoán	Tên công ty
ACB	Ngân hàng thương mại cổ phần Á Châu
BCM	Tổng Công ty Dầu tư và phát triển Công nghiệp
BID	Ngân hàng Thương mại cổ phần Dầu tư và Phát triển Việt Nam
BVH	Tập đoàn Bảo Việt
CTG	Ngân hàng thương mại cổ phần công thương Việt Nam
FPT	Công ty cổ phần FPT
GAS	Tổng công ty khí Việt Nam
GVR	Tập đoàn công nghiệp Cao su Việt Nam
HDB	Ngân hàng Thương mại Cổ phần Phát triển Thành phố Hồ Chí Minh
HPG	Công ty Cổ phần Tập đoàn Hòa Phát
MBB	Ngân hàng Thương mại Cổ phần Quân đội
MSN	Công ty cổ phần tập đoàn Masan
MWG	Công ty cổ phần đầu tư Thế giới di động
NVL	Công ty cổ phần Đầu tư Địa ốc No Va
PDR	Công ty cổ phần Phát triển Bất động sản Phát Đạt
PLX	Tập đoàn Xăng dầu Việt Nam
POW	Tổng công ty Điện lực Dầu khí Việt Nam
SAB	Tổng công ty cổ phần Bia - Rượu - Nước giải khát Sài Gòn
SSI	Công ty cổ phần chứng khoán SSI
STB	Ngân hàng Thương mại Cổ phần Sài Gòn Thương Tín
TCB	Ngân hàng Thương mại Cổ phần Kỹ Thương Việt Nam - Techcombank
TPB	Ngân hàng Thương mại cổ phần Tiên Phong
VCB	Ngân hàng Thương mại Cổ phần Ngoại thương Việt Nam
VHM	Công ty cổ phần Vinhomes
VIB	Ngân hàng Thương mại Cổ phần Quốc tế Việt Nam
VIC	Tập đoàn VINGROUP
VJC	Công ty cổ phần Hàng không VIETJET
VNM	Công ty cổ phần Sữa Việt Nam
VPB	Ngân hàng Thương mại Cổ phần Việt Nam thịnh vượng
VRE	Công ty cổ phần Vincom Retail

Để chuẩn bị tập dữ liệu bao gồm 30 cổ phiếu của VN30, chúng tôi sẽ sử dụng một số công cụ hỗ trợ và cuối cùng sẽ lưu trữ lại dưới các tập tin có đuôi .csv. Đầu tiên, chúng tôi sẽ dùng thư viện *requests* của Python để lấy nội dung HTML chứa tên 30 mã chứng khoán của VN30, sau đó sẽ tách tên 30 mã chứng khoán ra

bằng thư viện *BeautifulSoup*. Sau đó, chúng tôi sẽ dùng chúng để làm tham số cho một mã nguồn mà chúng tôi đã tự tạo ra bằng *Scrappy* nhằm thu thập dữ liệu lịch sử của cổ phiếu. Dữ liệu sẽ được chia thành 30 tập tin đuôi .csv do có 30 cổ phiếu, các đặc trưng của dữ liệu khi thu thập trực tiếp được mô tả như sau:

- **date**: ngày giao dịch,
- **open**: giá mở cửa,
- **high**: giá cao nhất,
- **low**: giá thấp nhất,
- **close**: giá đóng cửa,
- **cumulativeRateAdjusted**: tỉ lệ giá khi điều chỉnh giữa giá ngày đó và giá hiện tại,
- **volume**: tổng khối lượng giao dịch khớp lệnh,
- **totalMatchValue**: Tổng giá trị giao dịch,

Bảng 4.2: Một vài ngày dữ liệu giao dịch của cổ phiếu ACB

date	open	high	low	close
03-12-2018	12295.12	12870.82	12295.12	12870.82
04-12-2018	12870.82	12994.18	12665.21	12788.57
03-12-2018	12706.33	12829.69	12500.73	12747.45
06-12-2018	12665.21	12870.82	12582.97	12665.21
cumulativeRateAdjusted		volume	totalMatchValue	
0.411208		4691062.0	1.440243e+11	
0.411208		2802120.0	8.684115e+10	
0.411208		3447060.0	1.060999e+11	
0.411208		3345040.0	1.036216e+11	

4.3.2 Xử lý dữ liệu

Dữ liệu mà chúng tôi thu thập trực tiếp ở Tiểu mục 4.3.1 là dữ liệu thô. Chúng vẫn chưa phù hợp để có thể dùng làm đầu vào cho mô hình học tăng cường mà chúng tôi sẽ sử dụng. Đầu tiên, tám đặc trưng của dữ liệu sẽ không được sử dụng hết, chúng tôi chỉ sử dụng năm đặc là **open**, **high**, **low**, **close**, **volume** do các đặc trưng này đã đại diện đầy đủ cho dữ liệu giao dịch. Mặc dù vậy, các đặc trưng dữ liệu trên vẫn chưa có được định dạng cần thiết, dẫn đến chúng tôi sẽ thực hiện một số thay đổi sau trên toàn bộ dữ liệu của 30 cổ phiếu:

- Loại bỏ các ngày giao dịch mà có một vài giá trị rỗng,
- Loại bỏ những ngày mà cổ phiếu không được giao dịch.

Tới đây, dữ liệu đã có các định dạng cần thiết. Tuy nhiên, để tạo ra một mô hình học tăng cường với hiệu suất cao, ngoài dữ liệu giao dịch về giá và khối lượng chúng tôi sẽ thêm vào đó các chỉ báo kỹ thuật. Chúng tôi sẽ cung cấp các thông tin cần thiết, các mối quan hệ ẩn trong các thông tin về giá và khối lượng, góp phần làm giàu cho dữ liệu. Chúng tôi sẽ thêm vào một số đặc trưng sau, dựa trên các chỉ báo kỹ thuật mà chúng tôi đã trình bày ở Mục 2.2:

Bảng 4.3: Danh sách các đặc trưng dữ liệu sẽ được thêm vào đầu vào của mô hình

Đặc trưng	Mô tả
macd	Chỉ báo MACD
boll_ub	Cận trên của chỉ báo Dải Bollinger
boll_lb	Cận dưới của chỉ báo Dải Bollinger
rsi_30	Chỉ báo RSI của 30 ngày gần nhất
cci_30	Chỉ báo CCI của 30 ngày gần nhất
adx_30	Chỉ báo ADX của 30 ngày gần nhất
close_sma_30	Chỉ báo SMA của 30 ngày gần nhất
close_sma_60	Chỉ báo SMA của 60 ngày gần nhất

4.3.3 Chuẩn hóa dữ liệu

Vì các cổ phiếu khác nhau có các mức giá khác nhau, nên chúng tôi sẽ thực hiện chuẩn hóa dữ liệu để tạo ra một mô hình có khả năng tổng quát hóa. Mục đích của chúng tôi là tạo ra các mô hình có thể được sử dụng để dự đoán xu hướng giá trên các cổ phiếu thuộc các ngành khác nhau và các quãng thời gian không chỉ trên tập dữ liệu hiện có mà còn trong tương lai. Dữ liệu sẽ được đưa vào mô hình Học tăng cường sẽ bao gồm 13 đặc trưng, năm trong số đó là các giá trị OHLVC (*open, high, low, close, volume*) và tám đặc trưng còn lại là chỉ báo kỹ thuật ở Bảng 4.3. Các đặc trưng dữ liệu này có các khoảng giá trị khác nhau do đó chúng tôi sẽ thực hiện các kỹ thuật chuẩn hóa khác nhau. Đầu tiên với các đặc trưng với giá trị có chặn trên và chặn dưới, chúng đã được đề cập ở Mục 2.2, chúng bao gồm: *rsi_30, cci_30, adx_30*, chúng sẽ được chuẩn hóa bằng *min-max normalization* nhằm thu giảm giá trị lại trong khoảng $[0, 1]$

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}.$$

Trong các đặc trưng còn lại, ngoài đặc trưng *volume* thì các đặc trưng còn lại đều chứa các thông tin về giá trong các phiên giao dịch tính theo ngày. Chúng tôi sẽ chuẩn hóa các đặc trưng dữ liệu về giá sau: *open*, *high*, *low*, *macd*, *boll_ub*, *boll_lb*, *close_sma_30*, *close_sma_60* theo đặc trưng *close*. Cách chuẩn hóa này sẽ giúp mô hình học tăng cường khám phá các mối quan hệ giữa chúng và giá đóng cửa. Chúng tôi tạm gọi nó bằng *close_normalization*, công thức của nó là

$$x' = \frac{x}{\text{close}}. \quad (4.1)$$

Với hai đặc trưng còn lại là *close*, *volume*, chúng sẽ được so sánh với giá trị của phiên gần nhất trước đó theo công thức sau

$$x'_t = \frac{x_t}{x_{t-1}}. \quad (4.2)$$

Việc chuẩn hóa các đặc trưng dữ liệu theo công thức (4.1) và (4.2) có ưu điểm là không phụ thuộc vào độ dài của dữ liệu do chúng so sánh với các giá trị gần nhất và giúp mô hình có khả năng học tổng quát hơn.

4.3.4 Dán nhãn dữ liệu

Dán nhãn là một phần quan trọng trong việc xây dựng một mô hình học máy đáng tin cậy. Việc dán nhãn dữ liệu sai sẽ khiến kết quả của mô hình trở nên vô nghĩa. Mục tiêu của mô hình là dự đoán liệu xu hướng giá có tăng lên hay không? Ở luận văn này chúng tôi quyết định sẽ đoán xu hướng giá ngày hôm sau có tăng lên so với ngày hôm trước, tức là khoảng thời gian $t+1$. Trên cơ sở này, chúng tôi sẽ dùng công thức đơn giản sau để xác định xu hướng giá

$$\text{label}_t = \begin{cases} 1, & \text{nếu } \text{close}_{t+1} > \text{close}_t \\ 0, & \text{trong các trường hợp còn lại.} \end{cases}$$

4.4 Hiện thực mô hình

4.4.1 Không gian trạng thái

Khi một ngày giao dịch kết thúc, dữ liệu về giá, khối lượng, ... của ngày đó sẽ được phản hồi lại cho tác nhân để dự đoán xu hướng ngày tiếp theo. Chúng tôi sẽ sử dụng 13 trường dữ liệu ở Tiểu mục 4.3.2 đã được chuẩn hóa qua Tiểu mục 4.3.3 để làm quan sát o cho môi trường dự đoán xu hướng giá trong mô hình học tăng cường.

open	high	low	close	volume	macd	boll_ub	boll_lb	rsi_30	cci_30	dx_30	close_sma	close_sma_60
1	1.007806	0.997658	1.002347	1.621862	0.015708	1.011909	0.967794	0.504058	1.262253	0.164232	0.940333	0.95173
1	1.00541	0.987635	1.010148	0.894689	0.01736	1.005303	0.969659	0.513447	1.262169	0.146149	0.934029	0.941023
0.992349	1	0.987758	1.010046	0.912567	0.019188	1.00254	0.967774	0.522792	1.275154	0.135206	0.927672	0.930821
1.014752	1.014752	0.994565	0.985463	1.017576	0.019664	1.017277	0.985207	0.508039	1.136093	0.149212	0.944513	0.942767

Hình 4.2: Một vài quan sát của môi trường trả về cho tác nhân

Tuy nhiên, nếu chỉ sử dụng o làm trạng thái s cho tác nhân nhận định xu hướng giá thì mô hình sẽ không có hiệu suất tốt do dữ liệu chưa đủ cơ sở để đưa ra dự đoán. Ở luận văn này, sinh viên thực hiện quyết định sẽ dùng 30 ngày gần nhất của quan sát o làm trạng thái cho tác nhân s . Vì vậy, trạng thái của tác nhân sẽ là một vec-tơ có 390 điểm với các thành phần được liệt kê dưới đây:

- 30 ngày gần nhất của **open**,
- 30 ngày gần nhất của **high**,
- 30 ngày gần nhất của **low**,
- 30 ngày gần nhất của **close**,
- 30 ngày gần nhất của **volume**,
- 30 ngày gần nhất của **macd**,
- 30 ngày gần nhất của **boll**,
- 30 ngày gần nhất của **rsi**,
- 30 ngày gần nhất của **cci**,
- 30 ngày gần nhất của **dx**,
- 30 ngày gần nhất của **sma**,
- 30 ngày gần nhất của **sma_60**.

4.4.2 Không gian hành động

Hành động a sẽ có giá trị nằm trong đoạn $[-1, 1]$. Giá trị hành động sẽ chọn theo một phân phối chuẩn với giá trị trung bình và phương sai được học từ chiến lược. Chi tiết sẽ được đề cập ở Tiểu mục 4.4.4. Giá trị hành động đại diện cho mức độ tự tin suy đoán xu hướng giá tăng hay giảm. Nếu giá tăng càng cao, hành động được đưa ra sẽ có giá trị gần về 1. Ngược lại, khi giá giảm quá sâu, giá trị hành động sẽ đi về gần -1. Nếu giá thay đổi không quá lớn, giá trị hành động sẽ xấp xỉ bằng 0. Để thuận tiện trong việc xác định liệu giá trị hành động gần về hai đầu mút tương ứng với việc giá tăng giảm bao nhiêu %, chúng tôi sẽ dùng biên độ dao động giá là 7% được quy định cho sàn HOSE. Dựa trên cơ sở đó, nếu giá của ngày

hôm sau có độ tăng hoặc độ giảm về mức 7% thì giá trị hành động đi về hai chặn 1 và -1.

Tuy nhiên, mục đích của chúng tôi là dự đoán xu hướng giá nên chúng tôi sẽ đưa ra dự đoán của mình từ giá trị hành động được đưa ra theo công thức sau

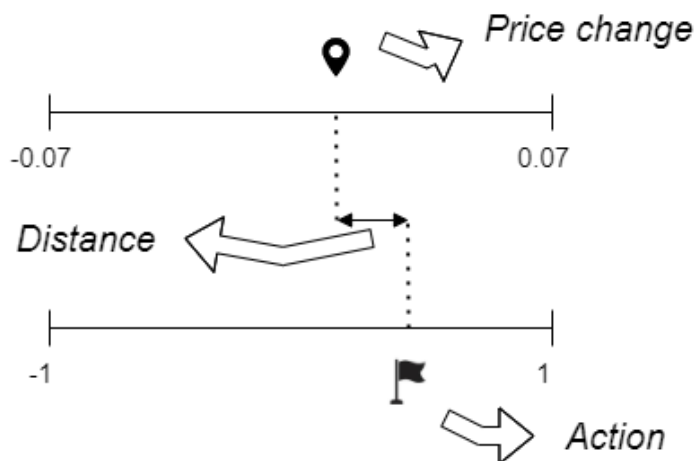
$$prediction_t = \begin{cases} 1, & \text{nếu } a_t > 0, \\ 0, & \text{trong các trường hợp còn lại.} \end{cases}$$

Việc đưa ra dự đoán sẽ có ích trong việc xác định hiệu quả của mô hình qua các độ đo được trình bày trong các chương sau.

4.4.3 Hàm phần thưởng

Việc thiết kế hàm phần thưởng rất quan trọng trong học tăng cường. Hàm phần thưởng đóng vai trò như một cơ chế khuyến khích, nó nói lại cho tác nhân biết hành động đúng nếu là một khoản thưởng hoặc sai nếu là một khoản phạt. Chúng tôi sẽ định nghĩa hàm phần thưởng với một số giả định riêng của bài toán mà chúng tôi đang thực hiện, chúng bao gồm:

- Không đề cập đến chi phí giao dịch và ảnh hưởng của nó trong thiết kế,
- Khi thực hiện các dự đoán lên giá hoặc xuống giá của cổ phiếu, chúng tôi sẽ xem như chúng tôi đã sở hữu các cổ phiếu đó.



Hình 4.3: Ý tưởng của hàm phần thưởng

Nhìn vào Hình 4.3, ở đó có hai trục lần lượt nói về hai thứ là mức thay đổi giá của ngày cần dự đoán so với ngày hiện tại và giá trị hành động mà tác nhân đưa ra. Ở trục thứ nhất, trục này có hai chặn là -0.07 và 0.07, đây chính là biên độ dao động của các cổ phiếu giao dịch trong sàn HOSE. Giá trị hành động của trục

thứ hai bị giới hạn bởi -1 và 1, đã được đề cập ở Tiểu mục 4.4.2. Chúng tôi thiết kế hàm phần thưởng với ý tưởng là đo lường khoảng cách giữa thay đổi giá của ngày cần dự đoán với ngày hiện tại, ở đây là ngày hôm sau với giá trị hành động mà tác nhân sẽ thực hiện với giá trị hành động. Mục tiêu của hàm phần thưởng là giảm thiểu khoảng cách này. Khi khoảng cách nhỏ, phần thưởng sẽ lớn và ngược lại. Dựa trên ý định đó, hàm phần thưởng sẽ có công thức như sau

$$r(s_t, a_t, s_{t+1}) = \left(\text{reward_shifting} - \frac{|\text{price_change}_t * 100 - a_t * 7|}{14} \right) * \text{reward_scaling}$$

$$\text{price_change}_t = \frac{\text{close}_{t+1}}{\text{close}_t} - 1.$$

Hàm phần thưởng được thiết kế với tập giá trị liên tục hay “đặc” so với “thưa” theo kỹ thuật “Reward shaping” [34] sẽ giúp cho mô hình hội tụ tốt hơn. Ta xét biểu thức ở tử số trước $\frac{|\text{price_change}_t * 100 - a_t * 7|}{14}$. Do khoảng giá trị của hành động và thay đổi giá nằm ở các khoảng khác nhau nên đầu tiên chúng tôi sẽ điều chỉnh hai khoảng này cho chúng cùng bằng nhau, điều này thực hiện bằng cách lấy 100 lần thay đổi giá và bảy lần giá trị hành động do chúng có khoảng giá trị lần lượt là $[-0.07, 0.07]$ và $[-1, 1]$. Như vậy khoảng giá trị sẽ được nâng lên là $[-7, 7]$. Vì chúng tôi muốn tìm khoảng cách giữa chúng nên chúng tôi chọn phép trị tuyệt đối của hiệu hai giá trị này. Khoảng cách này sẽ có khoảng giá trị từ $[0, 14]$, việc chia mẫu là 14 sẽ thu giảm nó về $[0, 1]$. Tuy nhiên theo như ý tưởng đã đề cập ở trên chúng tôi muốn khoảng cách càng nhỏ, mức thưởng càng cao nên chúng tôi lấy giá trị âm của biểu thức, như vậy khoảng giá trị của nó sẽ là $[-1, 0]$.

Đến đây, nếu thay đổi giá và hành động bằng nhau thì mức thưởng tối đa mà tác nhân nhận được sẽ là 0. Do chúng tôi muốn thưởng thêm cho các trường hợp đó, chúng tôi sẽ thêm vào một hằng số **reward_shifting**. Như vậy, mức thưởng tối đa sẽ nâng lên tương ứng. Cuối cùng, chúng tôi sẽ thêm vào một hằng số khác là **reward_scaling** để thay đổi tập giá trị lên nhiều lần. Qua quá trình thử và sai chúng tôi quyết định chọn hai hằng số này lần lượt là 0.2 và 20.

4.4.4 Chiến lược

Ở Tiểu mục này, sinh viên thực hiện sẽ trình bày cách cấu trúc mô hình của chiến lược, nó mô tả hành vi của tác nhân khi nhận được quan sát sau khi hết một ngày giao dịch. Trong nghiên cứu này, sinh viên thực hiện sẽ xây dựng mô hình

của chiến lược theo phương pháp Actor Critic đã được đề cập ở Tiểu mục 2.3.4.

Bảng 4.4: Cấu trúc của actor

	Lớp	Kích thước đầu ra	Lượng tham số
1	Input	(Nx30x13)	0
2	Flatten	(Nx390)	0
3	Linear(390, 64) Tanh	(Nx64)	25024
4	Dropout(0.2)	(Nx64)	0
5	Linear(64, 64) Tanh	(Nx64)	4160
6	Dropout(0.2)	(Nx64)	0
7	Linear(64, 1) Tanh	(Nx1)	65
Tổng			29249

Bảng 4.5: Cấu trúc của critic

	Lớp	Kích thước đầu ra	Lượng tham số
1	Input	(Nx30x13)	0
2	Flatten	(Nx390)	0
3	Linear(390, 64) Tanh	(Nx64)	25024
4	Dropout(0.2)	(Nx64)	0
5	Linear(64, 64) Tanh	(Nx64)	4160
6	Dropout(0.2)	(Nx64)	0
7	Linear(64, 1)	(Nx1)	65
Tổng			29249

Chiến lược là cách mà tác nhân hành động tùy theo phản hồi từ môi trường. Vì mục tiêu của bài toán là dự đoán xu hướng giá cổ phiếu nên dữ liệu đầu vào chứa nhiều thông tin nhất có thể. Dữ liệu đầu vào cho actor và critic chính là trạng thái s đã được mô tả ở Tiểu mục 4.4.1.

Việc lựa chọn các kiểu thiết kế mô hình xấp xỉ của actor và critic thông qua quá trình thử nghiệm. Actor và critic có thể học chung một mô hình và chia sẻ trọng số ở các lớp dưới hoặc học trọng số riêng rẽ. Sinh viên đã thử nghiệm cả hai thiết kế và việc thiết kế tách ra actor và critic theo cách sau đem lại hiệu quả tốt hơn. Actor được thiết kế để học theo chiến lược Gaussian, do sinh viên muốn đầu ra của actor là giá trị trung bình có một miền giá trị liên tục trải dài từ âm tới dương do giá cổ phiếu có thể tăng hoặc giảm. Như đã thấy đầu ra của actor sẽ có giá trị nằm trong khoảng $[-1, 1]$ vì đã qua hàm kích hoạt Tanh của lớp cuối cùng

trong cấu trúc actor ở Bảng 4.4. Giá trị phương sai lại có hai kiểu thiết kế bao gồm thêm một đầu ra trong cấu trúc actor hoặc được thiết lập như một tham số. Sinh viên chọn cách sau và thiết kế dựa theo ý tưởng suy giảm tỉ lệ học α và thêm vào bốn tham số

Bảng 4.6: Các tham số được thêm vào nhằm điều chỉnh giá trị phương sai

Tham số	Mô tả
action_std	Phương sai ban đầu của hành động
action_std_decay_rate	Độ suy giảm của phương sai của hành động
min_action_std	Giá trị nhỏ nhất của phương sai của hành động
action_std_decay_freq	Số tập (episode) mà sẽ xảy ra một lần suy giảm phương sai của hành động

Hàm mục tiêu. Hàm mục tiêu để học chiến lược bao gồm hai hàm mục tiêu của actor và critic

$$\mathcal{L} = \mathcal{L}_{actor} + \mathcal{L}_{critic}.$$

Trong đó, \mathcal{L}_{critic} là hàm mục tiêu để học critic nhằm xấp xỉ giá trị trạng thái $\hat{v}(s)$. Việc học hàm này giống với một bài toán học giám sát, nó sẽ có đầu vào là tập kinh nghiệm \mathcal{D} với kích thước là batch_size và nhãn là lợi nhuận G . Do đó, sinh viên thực hiện chọn MSE (Mean Square Error) để học hàm này

$$\mathcal{L}_{critic} = \sum_{t=0}^T \frac{1}{2} (G(s_t) - \hat{v}_{\pi}(s_t))^2.$$

Trong khi đó, hàm mục tiêu của actor sẽ giúp chiến lược có thể đưa ra giá trị trung bình của hành động và từ đó chọn ra giá trị hành động thực sự theo phân phối Gaussian. Sinh viên quyết định học actor theo thuật toán PPO và hàm mục tiêu này có dạng

$$\mathcal{L}_{actor} = \sum_{t=0}^T \left[\min \left(r_t(\theta) \hat{A}_t^{\theta_{old}}, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t^{\theta_{old}} \right) \right].$$

Trong đó, T chính là bước thời gian của mẫu kinh nghiệm cuối cùng trong \mathcal{D} ,

$$r_t(\theta) = \frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)},$$

$$\hat{A}_t^{\theta_{old}} = G(s_t) - \hat{v}_{\pi_{old}}(s_t).$$

CHƯƠNG 5

THỰC NGHIỆM VÀ KẾT QUẢ

5.1 Bộ dữ liệu

Bộ dữ liệu được sử dụng cho thực nghiệm là dữ liệu cổ phiếu VN30. Dữ liệu này do sinh viên thực hiện tự chuẩn bị, chi tiết các bước xử lý đã được trình bày ở Mục 4.3. Ở luận văn này, sinh viên thực hiện sử dụng cách chia thường thấy trong các bài toán về học máy. Đó là chia dữ liệu làm hai phần dành cho hai tập huấn luyện và kiểm thử. Chi tiết độ dài của dữ liệu của mỗi tập và phân phối dữ liệu của tập kiểm thử sẽ được trình bày ở hai bảng sau.

Bảng 5.1: Bộ dữ liệu được sử dụng

Bộ dữ liệu	Khoảng thời gian	Độ dài	Mô tả
VN30	04/06/2018 - 03/05/2022	874 ngày	Huấn luyện
VN30	04/05/2022 - 24/02/2023	205 ngày	Kiểm thử

Bảng 5.2: Phân bố nhãn trên tập kiểm thử của VN30

Mã chứng khoán	% Nhãn 0	% Nhãn 1
ACB	50.98	49.02
BCM	59.80	40.20
BID	50.49	49.51
BVH	51.47	48.53
CTG	50.98	49.02
FPT	51.47	48.53
GAS	51.47	48.53
GVR	52.94	47.06
HDB	50.49	49.51
HPG	58.82	41.18
MBB	52.45	47.55
MSN	52.94	47.06
MWG	53.43	46.57
NVL	62.74	37.26
PDR	62.25	37.75
PLX	53.43	46.57
POW	50.98	49.02
SAB	52.94	47.06
SSI	54.90	45.10
STB	52.45	47.55
TCB	52.94	47.06
TPB	55.88	44.12
VCB	52.45	47.55
VHM	59.31	40.69
VIB	56.86	43.14
VIC	57.35	42.65
VJC	55.39	44.61
VNM	53.43	46.57
VPB	55.88	44.12
VRE	54.90	45.10

5.2 Các độ đo được sử dụng

Trong phần này, sinh viên thực hiện sẽ trình bày các độ đo được sử dụng để đánh giá độ hiệu quả của mô hình dự đoán giá cổ phiếu. Vì bài toán này có thể

được xếp vào loại bài toán phân loại do kết quả đầu ra bao gồm xu hướng lên hoặc xuống nên sinh viên quyết định sẽ sử dụng độ chính xác và F1-Score do đây là các độ đo phổ biến nhất với bài toán phân loại.

Độ chính xác (Accuracy): độ đo của bài toán phân loại đơn giản nhất, tính toán bằng cách lấy số dự đoán đúng chia cho toàn bộ các dự đoán. Nhược điểm của cách đánh giá này là chỉ cho biết được bao nhiêu phần trăm lượng dữ liệu được phân loại đúng mà không chỉ ra được cụ thể mỗi loại được phân loại như thế nào, lớp nào được phân loại đúng nhiều nhất hay dữ liệu của lớp nào thường bị phân loại nhầm nhất vào các lớp khác.

F1-Score: Độ chính xác chỉ tốt đối với các tập dữ liệu có nhãn được đánh mà kết quả cân bằng. Với các trường hợp bộ dữ liệu có nhãn lệch hẳn về một bên ví dụ như trong một bộ phân loại chó mèo có 100 ảnh với 90 ảnh chó và 10 ảnh mèo, nếu ta tạo ra một bộ phân loại chỉ ra kết quả chó ta sẽ được một mô hình có độ chính xác 90%. Tuy nhiên, bộ phân loại như vậy là không tốt do nó không phân loại được ảnh mèo. Do đó, độ đo F1-Score sẽ giải quyết vấn đề đó.

Các công thức dùng để tính các độ đo được trình bày dưới đây:

$$\begin{aligned} Accuracy &= \frac{TP + TN}{TP + FP + TN + FN}, \\ Precision &= \frac{TP}{TP + FP}, \\ Recall &= \frac{TP}{TP + FN}, \\ F1-Score &= 2 * \frac{Precision * Recall}{Precision + Recall}, \end{aligned}$$

trong đó:

- TP : Nhãn được dự đoán là đúng và nhãn thực sự là đúng,
- TN : Nhãn được dự đoán là sai và nhãn thực sự là sai,
- FP : Nhãn được dự đoán là đúng và nhãn thực sự là sai,
- FN : Nhãn được dự đoán là sai và nhãn thực sự là đúng.

Ở đây, sinh viên thực hiện sẽ sử dụng một phiên bản khác của F1-Score là *Weighted F1-Score* do F1-Score nguyên bản chỉ nói lên các số liệu về một nhãn khác với bài toán dự đoán xu hướng giá tập trung vào cả sự lên và xuống giá của cổ phiếu

$$Weighted\ F1-Score = \sum_{i=1}^N w_i * F1-Score_i,$$

trong đó:

- N : số lớp trong bài toán phân loại có nhiều lớp,
- w : trọng số tương ứng gắn với từng lớp trên toàn thể tập dữ liệu.

5.3 Thí nghiệm đánh giá

Ở mục này, sinh viên thực hiện sẽ tiến hành thí nghiệm đo độ chính xác và F1-Score với tập kiểm thử của VN30. Thí nghiệm sẽ được lặp lại 30 lần tương ứng với 30 mã chứng khoán của VN30. Mỗi thí nghiệm sẽ bắt đầu bằng việc nạp và xử lý dữ liệu lịch sử của cổ phiếu được chọn, sau đó tạo ra một môi trường mô phỏng giao dịch. Ở đó, tác nhân sẽ thu được thông tin lịch sử giá và dựa vào đó đưa ra hành động. Môi trường nhận giá trị hành động và trả về phần thưởng dựa trên thay đổi giá của ngày tương ứng với giá trị hành động, đồng thời trả về nhân và dự đoán của ngày đó. Khi tác nhân dự đoán hết toàn bộ khoảng thời gian trong tập kiểm thử của cổ phiếu đó, thí nghiệm sẽ dừng lại.

Sinh viên thực hiện sẽ sử dụng thư viện học sâu Pytorch và Gymnasium của OpenAI và sử dụng GPU của máy tính cá nhân để huấn luyện trên nền tảng tính toán CUDA. Môi trường và thư viện sử dụng trong phần hiện thực được trình bày ở bảng sau.

Bảng 5.3: Môi trường và thư viện sử dụng

GPU	CUDA	CuDNN	Pytorch	Gymnasium
NVIDIA Geforce GTX 1050	12.0	8.7	1.13.1	0.27.1

Ngoài ra, toàn bộ 30 thí nghiệm đều sẽ sử dụng chung 30 bộ siêu tham số. Các giá trị siêu tham số này đều được chọn qua quá trình thử và sai nhằm đạt được kết quả gần tối ưu. Chi tiết thiết lập của các siêu tham số được trình bày sau đây.

Bảng 5.4: Thiết lập siêu tham số cho từng thí nghiệm

Siêu tham số	Mô tả	Giá trị
episode	Một vòng ở đó mô hình chạy qua toàn bộ khoảng thời gian của tập huấn luyện	500
gamma	Hệ số γ trong RL	0.01
batch_size	Kích cỡ một lô	300
entropy_coeff	Hệ số này dùng để kiểm soát giá trị $\mathcal{L}_{entropy}$	0.01
lr_actor	Tốc độ học của actor	0.0003
lr_critic	Tốc độ học của critic	0.001
K_epochs	Tham số K trong thuật toán PPO	20
action_std	Phương sai của hành động	0.6
action_std_decay_rate	Độ suy giảm của phương sai của hành động	0.05
min_action_std	Giá trị nhỏ nhất của phương sai của hành động	0.1
action_std_decay_freq	Số epoch mà sẽ xảy ra sự suy giảm phương sai của hành động	30

Bảng 5.5 mô tả lại kết quả thí nghiệm bao gồm độ chính xác và F1-Score của 30 mã chứng khoán trong giai đoạn huấn luyện.

Bảng 5.5: Kết quả độ đo trong thí nghiệm của 30 mã chứng khoán trên trong giai đoạn huấn luyện

Mã chứng khoán	Độ chính xác	F1-Score
ACB	61.56	61.62
BCM	74.14	74.34
BID	63.50	63.45
BVH	67.62	67.67
CTG	63.50	63.23
FPT	61.67	61.60
GAS	62.58	62.58
GVR	68.76	68.76
HDB	61.67	61.74
HPG	62.58	62.49
MBB	63.04	63.06
MSN	65.57	65.83
MWG	65.44	65.29
NVL	62.58	62.60
PDR	59.38	59.53
PLX	63.61	63.66
POW	67.20	67.54
SAB	68.19	68.29
SSI	63.16	62.95
STB	64.19	64.00
TCB	62.81	62.84
TPB	66.02	66.01
VCB	64.99	64.97
VHM	66.36	66.42
VIB	68.53	68.49
VIC	63.73	63.72
VJC	59.61	59.66
VNM	60.98	61.06
VPB	62.36	61.98
VRE	64.19	64.16
Trung bình	64.32	64.31

Bảng 5.6 mô tả lại kết quả thí nghiệm bao gồm độ chính xác và F1-Score của 30 mã chứng khoán trong giai đoạn kiểm thử. Nhìn vào kết quả, ta có thể rút ra

vài nhận xét sau đây:

- Xét trên từng mã chứng khoán, ta có thể thấy được rằng độ chính xác và F1-Score không có sự chênh lệch quá lớn. Điều đó nói lên rằng mô hình đã học được cách ra quyết định cân bằng, có thể đưa ra dự đoán cả xu hướng lên và xuống giá, không lệch hẳn về một bên nào.
- Xét trên tổng thể toàn bộ VN30, các độ đo trả về kết quả có giá trị nằm trong khoảng 56-62% với đa phần các mã chứng khoán cho ra giá trị nằm tập trung ở mức 57-59%. Sự khác biệt này xảy ra do phân phối nhãn của từng mã là khác nhau và việc chọn một chung bộ siêu tham số đã dẫn đến kết quả này.

Bảng 5.6: Kết quả độ đo trong thí nghiệm của 30 mã chứng khoán trong giai đoạn kiểm thử

Mã chứng khoán	Độ chính xác	F1-Score
ACB	57.56	57.00
BCM	60.10	60.10
BID	61.46	61.42
BVH	60.50	60.40
CTG	59.51	59.55
FPT	61.96	61.93
GAS	55.20	55.00
GVR	56.10	56.10
HDB	62.44	61.25
HPG	58.04	58.00
MBB	56.60	55.70
MSN	58.54	56.51
MWG	61.47	61.78
NVL	60.49	56.40
PDR	55.20	55.17
PLX	55.61	55.60
POW	56.10	56.00
SAB	57.10	57.00
SSI	62.00	61.12
STB	60.00	59.10
TCB	58.05	57.08
TPB	57.57	56.81
VCB	56.00	55.91
VHM	61.47	61.32
VIB	58.97	58.16
VIC	62.93	63.13
VJC	59.03	57.90
VNM	59.03	58.25
VPB	61.46	61.6
VRE	60.97	60.75
Trung bình	58.20	58.34

Ngoài ra sinh viên thực hiện cũng tiến hành đo kết quả của mô hình trên với các mô hình khác. Chúng tôi đã chạy lại các mô hình nghiên cứu trước theo hướng

học giám sát với dữ liệu cổ phiếu trong VN30. Có thể thấy mô hình của chúng tôi có kết quả tốt hơn rất nhiều với trung bình 58.20%. Điều này chứng minh phương pháp đề xuất của sinh viên đạt được hiệu suất mô hình tốt, có thể mở rộng để triển khai thành một bot giao dịch có thể hoạt động thực tế trên thị trường giao dịch hàng ngày.

Bảng 5.7: Kết quả độ đo so với các mô hình khác

Mã chứng khoán	Độ chính xác	F1-Score
Mô hình của chúng tôi	58.20%	58.34%
Mô hình dựa trên CNN [17]	51.37%	49.44%
Mô hình dựa trên LSTM [28]	50.4%	37.91%
Mô hình dựa trên CNN-LSTM [20]	53.29%	49.93%

5.4 Thí nghiệm mô phỏng dự đoán xu hướng giá

Sinh viên đã xây dựng một trò chơi mô phỏng dự đoán xu hướng giá và tiến hành ghi lại kết quả. Trò chơi bắt đầu với điểm (score) là 0. Với mỗi lần dự đoán đúng điểm sẽ tăng 1, và trừ 1 khi dự đoán sai. Hình ảnh trực quan sẽ hiển thị số điểm hiện tại đang có được cũng như biểu đồ nến và khối lượng giao dịch tương ứng. Hình 5.1 mô tả kết quả mô phỏng dự đoán đối với cổ phiếu VRE trong khoảng thời gian nằm trong tập kiểm thử đã đề cập ở Tiểu mục 5.1. Trò chơi kết thúc với số điểm là 25, trên khoảng thời gian là 205 ngày, như vậy độ chính xác khoảng 56%.



Hình 5.1: Trò chơi mô phỏng dự đoán xu hướng giá

CHƯƠNG 6

TỔNG KẾT

Trong chương này, sinh viên thực hiện sẽ tiến hành tổng kết các kết quả đã đạt được, chỉ ra các hạn chế và đề xuất hướng phát triển trong tương lai.

6.1 Các kết quả đạt được

Như vậy, trong luận văn này, sinh viên thực hiện đã tiến hành tìm hiểu những kiến thức nền tảng (Chương 2) và những công trình nghiên cứu liên quan (Chương 3) về đề tài dự đoán xu hướng giá cổ phiếu. Từ đó sinh viên thực hiện đã đề xuất và hiện thực mô hình sử dụng học tăng cường khác với các công trình nghiên cứu trước chỉ tập trung vào học giám sát, cũng như lựa chọn cách xử lý dữ liệu, kiến trúc mô hình và đặc tả hàm phần thưởng phù hợp với dữ liệu của đề tài. Kết quả của đề tài là mô hình đã có thể tự đưa ra các dự đoán xu hướng tăng giảm giá của cổ phiếu. Các độ đo của mô hình với các cổ phiếu của chỉ số VN30 ở Mục 5.3 có thể đạt kết quả vượt mức 60% về độ chính xác ở một vài cổ phiếu, khá tốt so với các kết quả đạt được ở các công trình nghiên cứu gần đây.

Bên cạnh đó, quá trình thực hiện luận văn đã giúp sinh viên thực hiện được hướng dẫn và rèn luyện về kỹ năng quản lý thời gian, kỹ năng xử lý vấn đề, phương pháp luận trong việc làm luận văn. Những khó khăn trong quá trình thực hiện đề tài đã giúp sinh viên tăng sự tự tin khi đối diện thử thách, để có thể phân tích trở ngại tìm hướng giải quyết. Những hành trang về kỹ năng và kinh nghiệm này sẽ rất quý giá với sinh viên trong con đường học tập và làm việc lâu dài.

6.2 Hạn chế và hướng phát triển

6.2.1 Hạn chế của mô hình

Bên cạnh những kết quả tốt đã nêu trên, mô hình dự đoán xu hướng giá không thể tránh khỏi tồn tại một số hạn chế sau:

- Dữ liệu sử dụng cho đề tài là dữ liệu giá cổ phiếu theo ngày. Việc sử dụng dữ liệu này làm cho bộ dữ liệu có kích thước nhỏ khiến cho mô hình không học được nhiều mẫu nhất có thể tương ứng với từng cổ phiếu. Kết quả đầu ra của mô hình vẫn chưa phù hợp với thực tế khi giao dịch nếu tính đến khoảng thời gian khớp lệnh và tiền từ cổ phiếu chảy về tài khoản nhà đầu tư,
- Sinh viên đã không xem xét đến chi phí giao dịch và các yếu tố rủi ro vô hình khi thiết kế mô hình hay sử dụng các độ đo về lợi nhuận khi xét đến kết quả của mô hình,
- Mô hình vẫn chưa đạt được khả năng tổng quát. Thí nghiệm được lặp lại riêng rẽ cho từng cổ phiếu và đạt được giá trị khác nhau khi sử dụng chung một thiết lập tham số. Mô hình muốn triển khai trên thực tế cần phải tổng quát hóa, ở đây là tạo ra một mô hình sử dụng chung cho toàn bộ VN30.

6.2.2 Hướng phát triển trong tương lai

Những hạn chế đã nêu trên cần được nghiên cứu sâu hơn và sẽ được dùng làm các hướng phát triển trong tương lai:

- Để mô hình có thể triển khai trên thực tế, dữ liệu dùng cho việc huấn luyện cần được gia tăng. Nếu sinh viên muốn tạo ra một bot giao dịch tự động thì dữ liệu yêu cầu phải theo thời gian thực, có thể tính theo phút,
- Ngoài dữ liệu về giá ra, xu hướng giá của cổ phiếu chịu ảnh hưởng bởi rất nhiều yếu tố từ chính trị, kinh tế, tin tức trên mạng xã hội cần được lưu ý khi mở rộng nghiên cứu sâu hơn,
- Sinh viên cần thêm vào các yếu tố rủi ro và lợi nhuận khi thiết kế mô hình và các yếu tố riêng của thị trường chứng khoán khi hiện thực bot giao dịch,
- Nhằm đạt được khả năng tổng quát hóa của mô hình, một hướng nghiên cứu tiềm năng có thể sử dụng là *meta-learning* có thể tạo ra một mô hình học tăng cường phản ứng với các dữ liệu thuộc các cổ phiếu khác nhau ở các khoảng thời gian khác nhau.

TÀI LIỆU THAM KHẢO

- [1] Benjamin Graham and David L. Dodd. *Security Analysis: Principles and Technique*. McGraw Hill, 1934.
- [2] David L. Dodd Benjamin Graham and Warren Buffett. *Security Analysis: Principles and Technique*. McGraw Hill, 2020.
- [3] Eugene Fama. “The behavior of stock-market prices”. In: *The journal of Business* 38 (1965), pp. 34–105.
- [4] Weiwei Jiang. ““Applications of deep learning in stock market prediction: Recent progress”. In: *Expert Systems with Applications* 184.115537 (Dec. 2021).
- [5] Alan Northcott. *The Complete Guide to Using Candlestick Charting: How to Earn High Rates of Return - Safely*. Atlantic Publishing Group, 2009, pp. 15–17.
- [6] Gerald Appel. *Technical Analysis Power Tools for Active Investors*. Financial Times Prentice Hall, 2005, p. 166.
- [7] J. Welles Wilder Jr. *New Concepts in Technical Trading Systems*. Trend Research, 1978.
- [8] Constance M. Brown. *Technical Analysis for the Trading Professional*. McGraw Hill Professional, 2012.
- [9] B Schlossberg. *Technical Analysis of the Currency Market: Classic Techniques for Profiting from Market Swings and Trader Sentiment*. Wiley Trading. Wiley, 2006, p. 91.
- [10] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

- [11] John Schulman et al. “Proximal policy optimization algorithms”. In: *arXiv preprint arXiv:1707.06347* (2017).
- [12] Guo-qiang Xie. “The optimization of share price prediction model based on support vector machine”. In: *2011 International Conference on Control, Automation and Systems Engineering (CASE)*. IEEE. 2011, pp. 1–4.
- [13] Jianxue Chen. “SVM application of financial time series forecasting using empirical technical indicators”. In: *2010 International Conference on Information, Networking and Automation (ICINA)*. Vol. 1. IEEE. 2010, pp. V1–77.
- [14] Milad Jasemi, Ali M Kimiagari, and Azizollah Memariani. “A modern neural network model to do stock market timing on the basis of the ancient investment technique of Japanese Candlestick”. In: *Expert Systems with Applications* 38.4 (2011), pp. 3884–3890.
- [15] Michel Ballings et al. “Evaluating multiple classifiers for stock price direction prediction”. In: *Expert systems with Applications* 42.20 (2015), pp. 7046–7056.
- [16] Jinan Zou. “Stock Market Prediction via Deep Learning Techniques: A Survey”. In: *arXiv preprint arXiv:2212.12717* (2022).
- [17] Ehsan Hoseinzade and Saman Haratizadeh. “CNNpred: CNN-based stock market prediction using a diverse set of variables”. In: *Expert Systems with Applications* 129 (2019), pp. 273–285.
- [18] Shumin Deng et al. “Knowledge-driven stock trend prediction and explanation via temporal convolutional network”. In: *Companion Proceedings of The 2019 World Wide Web Conference*. 2019, pp. 678–685.
- [19] Oren Etzioni et al. “Open information extraction from the web”. In: *Communications of the ACM* 51.12 (2008), pp. 68–74.
- [20] Wenjie Lu et al. “A CNN-LSTM-based model to forecast stock prices”. In: *Complexity* 2020 (2020), pp. 1–10.
- [21] Wenjie Lu et al. “A CNN-BiLSTM-AM method for stock price prediction”. In: *Neural Computing and Applications* 33 (2021), pp. 4741–4753.
- [22] Haiyao Wang et al. “A stock closing price prediction model based on CNN-BiSLSTM”. In: *Complexity* 2021 (2021), pp. 1–12.
- [23] Qihang Zhou, Changjun Zhou, and Xiao Wang. “Stock prediction based on bidirectional gated recurrent unit with convolutional neural network and feature selection”. In: *Plos one* 17.2 (2022), e0262501.

- [24] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. “Learning representations by back-propagating errors”. In: *nature* 323.6088 (1986), pp. 533–536.
- [25] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.
- [26] Zhiyong Zhao et al. “Time-weighted LSTM model with redefined labeling for stock trend prediction”. In: *2017 IEEE 29th international conference on tools with artificial intelligence (ICTAI)*. IEEE. 2017, pp. 1210–1217.
- [27] Thi-Thu Nguyen and Seokhoon Yoon. “A novel approach to short-term stock price movement prediction using transfer learning”. In: *Applied Sciences* 9.22 (2019), p. 4745.
- [28] Fuli Feng et al. “Enhancing stock movement prediction with adversarial training”. In: *arXiv preprint arXiv:1810.09936* (2018).
- [29] Yumo Xu and Shay B Cohen. “Stock movement prediction from tweets and historical prices”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018, pp. 1970–1979.
- [30] Zhipeng Liang et al. “Adversarial deep reinforcement learning in portfolio management”. In: *arXiv preprint arXiv:1808.09940* (2018).
- [31] Zhengyao Jiang, Dixing Xu, and Jinjun Liang. “A deep reinforcement learning framework for the financial portfolio management problem”. In: *arXiv preprint arXiv:1706.10059* (2017).
- [32] Jingyuan Wang et al. “Alphastock: A buying-winners-and-selling-losers investment strategy using interpretable deep reinforcement attention networks”. In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019, pp. 1900–1908.
- [33] Hongyang Yang et al. “Deep reinforcement learning for automated stock trading: An ensemble strategy”. In: *Proceedings of the first ACM international conference on AI in finance*. 2020, pp. 1–8.
- [34] Adam Daniel Laud. *Theory and application of reward shaping in reinforcement learning*. University of Illinois at Urbana-Champaign, 2004.