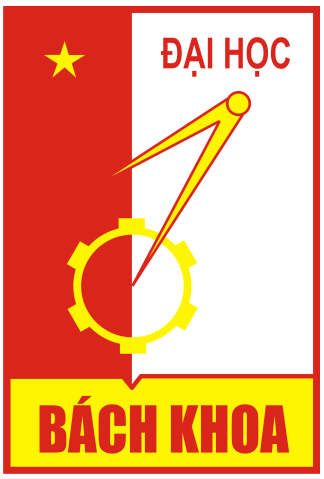
TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN TOÁN ỨNG DỤNG VÀ TIN HỌC



Dự đoán giá chứng khoán sử dụng mô hình LSTM - PCA.

ĐỒ ÁN 2

Chuyên ngành: Hệ thống thông tin quản lý

Giảng viên hướng dẫn:	**TS. Tạ Thị Thanh Mai**
Sinh viên thực hiện:	**Đoàn Lê Tường Vy**
Lớp:	**Hệ thống thông tin 01, K64.**

HÀ NỘI – 2023

NHẬN XÉT CỦA GIẢNG VIÊN

1. Mục đích và nội dung của đồ án

...

2. Kết quả đạt được

(a) ...

3. Ý thức làm việc của sinh viên

(a) ...

Hà Nội, ngày ... tháng ... năm 2023

Giảng viên

Lời cảm ơn

Trong suốt quá trình thực hiện Đồ án, tôi đã nhận được sự quan tâm, giúp đỡ và tạo điều kiện của các thầy cô giáo trong Viện Toán ứng dụng và Tin học, nhất là các các giảng viên trong bộ môn Toán ứng dụng.

Đặc biệt, tôi xin gửi lời biết ơn sâu sắc tới TS. Tạ Thị Thanh Mai, người đã tận tình hướng dẫn, giúp đỡ tôi trong suốt thời gian qua có thể hoàn thành Đồ án này.

Tôi xin chân thành cảm ơn!

Hà Nội, tháng 1 năm 2023

Sinh viên

Đoàn Lê Tường Vy

Mục lục

Bảng ký hiệu và chữ viết tắt

LSTM	Long-short term memory
PCA	Principal Components Analysis
RNN	Recurrent Neural Network
MAE	Mean absolute error
MAPE	Mean absolute percentage error

Mở đầu

Đối với các nhà đầu tư, chứng khoán đáp ứng các mong muốn đầu tư và nhu cầu đầu tư khác nhau, mở rộng phạm vi lựa chọn đầu tư, mở rộng các kênh đầu tư, đáp ứng ở mức độ nào đó khả năng nhà đầu tư có được thu nhập tương ứng, và ở mức độ nào đó nâng cao tính linh hoạt và tính thanh khoản của vốn [1]. Nếu nhìn về phía doanh nghiệp, chứng khoán có thể đóng vai trò quan trọng trong việc quản lý và phát triển doanh nghiệp, có lợi cho việc thiết lập và hoàn thiện cơ chế quản lý doanh nghiệp tự phát triển. Đối với đất nước, chứng khoán cũng là một công cụ tuyệt vời để ngăn chặn sự lạm phát.

Chứng khoán có ba đặc điểm chính: (1) không thể hoàn trả: một khi đã bán, chứng khoán không thể được trả lại cho công ty và không thể yêu cầu hoàn lại tiền mà chỉ có thể được bán cho bên thứ ba thông qua thị trường thứ cấp; (2) thu nhập không ổn định: lãi lỗ của chứng khoán phụ thuộc vào hoạt động của công ty và thị trường chứng khoán, cả hai đều không chắc chắn và có thể thay đổi, vì vậy nhà đầu tư cần chấp nhận nhiều rủi ro; và (3) mang tính đầu cơ: thị trường chứng khoán thường xuyên biến động và giá cả thị trường không ổn định, làm tăng tính rủi ro đối với các nhà đầu tư [2].

Có nhiều nguyên nhân ảnh hưởng đến sự biến động của giá chứng khoán và chính những yếu tố thường xuyên thay đổi này đã gây ra sự biến động của thị trường chứng khoán [3]. Rủi ro khách quan của thị trường chứng khoán có thể mang lại lợi nhuận cho nhà đầu tư, đồng thời có thể gây thiệt hại về kinh tế và cũng có thể ảnh hưởng tiêu cực đến điều kiện hoạt động của công, thậm chí mang lại ảnh hưởng tới công cuộc xây dựng kinh tế đất nước.

Những vấn đề này là không thể tránh khỏi, vì vậy dự đoán xu hướng chứng khoán đã trở thành một vấn đề được các bên quan tâm rất nhiều. Nghiên cứu dự đoán xu hướng chứng khoán cũng đã trở thành một hướng nghiên cứu ứng dụng của Big Data trong lĩnh vực tài chính và nhiều học giả đã áp dụng các phương pháp Neural Network để dự đoán xu hướng chứng khoán, trở thành một trong những vấn đề nghiên cứu phổ biến trong lĩnh vực học thuật hiện nay [4]. Trong những năm gần đây, với sự phát triển nhanh chóng của công nghệ máy tính, Neural Network đã trở thành đối tượng nghiên cứu chính và lĩnh vực ứng dụng ngày càng được mở rộng, bao gồm cả lĩnh vực tài chính. Hiện nay, thị trường tài chính chiếm vị trí then chốt trong toàn bộ hệ thống kinh tế của đất nước, chứng khoán là một bộ phận cấu thành quan trọng của thị trường tài chính nên việc mua chứng khoán đã trở thành một phương thức quản lý tài chính phổ biến hiện nay. Theo đó là sự cần thiết của một phần mềm phân tích chứng khoán với nhiều tính năng khác nhau có thể đáp ứng nhu cầu thực của nhà đầu tư. Đồ án này thiết kế và triển khai hệ thống dự đoán xu hướng chứng khoán bằng cách sử dụng dữ liệu lịch sử giao dịch chứng khoán để khuyến nghị chứng khoán và dịch vụ dự đoán xu hướng chứng khoán cho nhà đầu tư nhằm giảm hoặc tránh rủi ro đầu tư, từ đó mang lại cho nhà đầu tư lợi nhuận kinh tế tương đối ổn định. Do sự đổi mới liên tục của các kỹ thuật học máy, ngày càng có nhiều nhà nghiên cứu chuyển sang sử dụng các kỹ thuật học máy để phân tích dữ liệu chứng khoán và tạo ra các mô hình phương pháp hiệu quả để dự đoán sự biến động của chứng khoán trong tương lai. Các mô hình dự đoán thị trường chứng khoán được xây dựng bằng cách học hỏi từ dữ liệu giá lịch sử để dự đoán giá trong tương lai [5]. Theo nhiều nhà nghiên cứu, dữ liệu lịch sử giá và các chỉ số khác của chứng khoán tiết lộ mối tương quan với các mô hình biến động giá chứng khoán, từ đó có thể dự đoán được giá chứng khoán trong tương lai dựa trên những dữ liệu lịch sử một cách tương đối khách quan và chính xác. Các thuật toán học máy phổ biến như hồi quy logistic, regression và support vector machine (SVM) đã được sử dụng cho kết quả tốt trong dự báo. Với sự phát triển của công nghệ Neural

Network, việc xây dựng mạng lưới để mô tả giá chứng khoán và dự đoán chuyển động của chứng khoán đã nhận được rất nhiều sự chú ý và một số học giả đã tiến hành nghiên cứu chuyên sâu về lĩnh vực này. Để nâng cao độ chính xác của dự đoán xu hướng chứng khoán, nhiều thuật toán cải tiến và chiến lược tối ưu hóa đã lần lượt xuất hiện và được áp dụng thành công trong thực tế.

Trong đề tài này, tôi đã khảo sát phương pháp học máy tiên tiến Long short-term memory (LSTM), kết hợp với thuật toán Phân tích thành phần chính (Principal Components Analysis - PCA) để phân tích dữ liệu lịch sử giá và các chỉ số kỹ thuật trong lĩnh vực chứng khoán, từ đó dự báo xu hướng giá của chứng khoán trong tương lai. Mặc dù kết quả của mô hình xây dựng chưa đạt mức cao nhưng có độ tin cậy đáng kể, cung cấp thêm thông tin hữu ích cho các quyết định mua hoặc bán chứng khoán của các tổ chức và nhà đầu tư, đồng thời định hướng cho các nghiên cứu tiếp theo của tôi.

Chương 1

Bài toán dự đoán xu hướng chứng khoán

1.1 Chứng khoán và thị trường chứng khoán

1.1.1 Tổng quan thị trường chứng khoán

Thị trường chứng khoán là một bộ phận quan trọng của Thị trường vốn, hoạt động của nó nhằm huy động những nguồn vốn tiết kiệm nhỏ trong xã hội tập trung thành nguồn vốn lớn tài trợ cho doanh nghiệp, các tổ chức kinh tế và Chính phủ để phát triển sản xuất, tăng trưởng kinh tế hay cho các dự án đầu tư. Thị trường chứng khoán là nơi diễn ra các hoạt động giao dịch mua bán các loại chứng khoán. Việc mua bán được tiến hành ở hai thị trường sơ cấp và thứ cấp, do vậy thị trường chứng khoán là nơi chứng khoán được phát hành và trao đổi. Thị trường chứng khoán cung cấp nơi cho các hoạt động giao dịch mua bán chứng khoán bao gồm:

- Thị trường sơ cấp: người mua mua được chứng khoán lần đầu từ những người phát hành.
- Thị trường thứ cấp: nơi diễn ra sự mua đi bán lại các chứng khoán đã được phát hành ở Thị trường sơ cấp.

Hàng hóa giao dịch trên Thị trường chứng khoán bao gồm: các chứng khoán,

trái phiếu và một số công cụ tài chính khác có thời hạn trên 1 năm.

1.1.2 Các chỉ số cơ bản của một mã chứng khoán trên sàn giao dịch

Với mỗi mã chứng khoán khi thực hiện giao dịch trên thị trường chứng khoán niêm yết có các chỉ số cơ bản thể hiện trên bảng giá, ý nghĩa của các cột trên các bảng giá theo từng sàn như sau:

- Cột giá trần (High: HI): Là mức giá cao nhất mà nhà đầu tư có thể đặt lệnh mua, lệnh bán chứng khoán.
- Cột giá sàn (Low: LO): Là mức giá thấp nhất mà nhà đầu tư có thể đặt lệnh mua, bán chứng khoán.
- Cột giá mở cửa (Open: OP): Là mức giá thực hiện đầu tiên trong ngày giao dịch.
- Cột giá đóng cửa (Close: CL): Là mức giá thực hiện cuối cùng trong ngày giao dịch.
- Cột giá đóng cửa điều chỉnh (Adj Close: AD): giá được điều chỉnh lại để xác định giá trị chính xác của chứng khoán, trong trường hợp công ty có các hoạt động làm thay đổi số lượng chứng khoán.

Khi kết thúc phiên giao dịch, Bảng điện tử sẽ hiện thị các thông tin về khối lượng giao dịch trong ngày (Volume: VO).

1.1.3 Các chỉ số kỹ thuật dùng trong dự báo chứng khoán

Trong phân tích kỹ thuật của chứng khoán có nhiều chỉ số khác nhau được sử dụng.

Các chỉ số kỹ thuật có thể được thông qua để dự đoán giá chứng khoán của công ty [9]. Các chỉ số kỹ thuật là các tính toán toán học trên cơ sở dữ liệu giao

dịch cơ bản của chứng khoán. Chúng đã được chứng minh là có nhiều thông tin tiềm ẩn và phong phú về thị trường chứng khoán.

Phân tích kĩ thuật là một bộ môn giao dịch được sử dụng để đánh giá các khoản đầu tư và xác định cơ hội giao dịch bằng việc phân tích xu hướng của những số liệu thống kê được từ hoạt động giao dịch, như là biến động của giá và khối lượng.

Không giống như các nhà phân tích cơ bản, là những người xác định giá trị nội tại của một chứng khoán dựa trên dữ liệu tài chính và kinh tế, các nhà phân tích kĩ thuật tập trung vào mẫu hình của biến động giá, chỉ báo giao dịch và nhiều công cụ phân tích đồ thị khác để đánh giá điểm mạnh và điểm yếu của một chứng khoán.

Trong Đồ án này, tôi sẽ sử dụng thư viện có sẵn `https://github.com/bukosabino/ta` để tính toán 86 chỉ báo kỹ thuật thị trường chứng khoán và đưa vào đầu vào của mô hình dự đoán. Các chỉ báo kỹ thuật sẽ được mô tả ngắn gọn dưới đây.

Chỉ báo Volume

Volume thể số lượng giao dịch được thực hiện trong một khoảng thời gian cụ thể. Bằng cách đo lường khối lượng giao dịch, nhà đầu tư có thể dự đoán liệu một xu hướng có khả năng kéo dài hay không.

Ví dụ: Khối lượng lớn trong xu hướng tăng cho thấy nhu cầu giao dịch cao và do đó, giá sẽ tăng. Tương tự, khối lượng lớn trong xu hướng giá giảm cho thấy nguồn cung cao và mã chứng khoán vẫn có khả năng giảm. Một vài chỉ báo Volume phổ biến:

- Chỉ báo dòng tiền (Money Flow Index - MFI): chỉ số kỹ thuật dùng để đánh giá cường độ của dòng tiền bằng cách so sánh giá tăng tích cực và tiêu cực trong một khoảng thời gian nhất định, có tính đến khối lượng giao dịch.

- Chỉ số tích lũy/ phân phối (Accumulation/Distribution Index - ADI): là một chỉ số tích lũy sử dụng khối lượng và giá cả để đánh giá xem một cặp

tiền đang được tích lũy hay phân phối. Nếu giá tăng nhưng chỉ báo giảm, điều này cho thấy khối lượng mua hoặc tích lũy có thể không đủ để hỗ trợ cho việc tăng giá và có thể sẽ giảm giá.

- Chỉ báo cân bằng khối lượng (On-Balance Volume - OBV): là chỉ báo giúp đo lường sức mua và sức bán trên thị trường, dựa trên cả khối lượng giao dịch và chuyển động của giá. Nếu động lực của xu hướng hiện tại mạnh, thị trường sẽ tiếp diễn xu hướng cũ. Ngược lại, nếu động lực yếu thì khả năng sẽ đảo chiều sang xu hướng mới.

- Chaikin Money Flow (CMF): là chỉ báo đo lường khối lượng dòng tiền trong một chu kỳ nhất định. Chỉ báo CMF dao động lên xuống quanh mức 0 và là công cụ để nhà đầu tư đánh giá áp lực mua/bán dựa trên sự thay đổi của dòng tiền.

- Chỉ số Force (Force Index - FI): Chỉ số này đo lường sức mua và bán trong thị trường xu hướng dựa trên giá cả, hướng đi thị trường và khối lượng giao dịch. Chỉ số có giá trị dương cao có nghĩa là có một xu hướng tăng mạnh và giá trị âm thấp cho thấy một xu hướng giảm mạnh.

- Ease of Movement (EoM, EMV): là chỉ báo dạng biểu đồ đường, di chuyển lên xuống so mới mức 0. Khi đường này nằm phía trên mức 0 (tức là có giá trị dương), là tín hiệu cho thấy thị trường đang ở trong xu hướng tăng. Và xu hướng giảm khi đường này cắt xuống bên dưới mức 0.

Chỉ báo Volatility

Chỉ báo Volatility trong giao dịch xác định mức độ biến động của giá theo thời gian. Biến động cao cho thấy sự thay đổi giá nhanh chóng và không thể đoán trước. Các chỉ số biến động đo phạm vi giá của một tài sản và giúp nắm bắt những thời điểm biến động cao.

Nhiều nhà giao dịch ủng hộ thị trường có tính biến động cao vì chúng mang lại nhiều cơ hội giao dịch cùng lợi nhuận nhanh hơn và cao hơn. Một vài chỉ báo Volatility phổ biến:

- Khoảng dao động trung bình thực tế (Average True Range - ATR): ATR đo lường sự biến động giá. Nó giúp các nhà giao dịch dự đoán giá có thể di chuyển bao xa trong tương lai.

- Bollinger Bands (BB): là một công cụ phân tích kỹ thuật xác định bởi đường trung bình đơn giản (Simple Moving Average - SMA) ở dải giữa, dải trên và dải dưới. Dải Bollinger Bands sẽ tự điều chỉnh mở rộng trong các giai đoạn thị trường biến động và thu hẹp trong các giai đoạn thị trường ít biến động hơn.

- Keltner Channel (KC): là chỉ báo xu hướng được thiết lập để tìm điểm đảo chiều dựa trên mức độ dao động của giá với hai đường bao, tương tự như dải Bollinger. KC được xác định bởi đường trung bình động hàm mũ (Exponential Moving Average - EMA) ở dải giữa, dải trên và dải dưới.

Chỉ báo Trend

Các chỉ số này mô tả hướng chuyển động của giá trong một khoảng thời gian dài. Ví dụ: khi giá liên tục tăng, đó là xu hướng tăng và khi giá giảm, đó là xu hướng giảm.

Các chỉ báo xu hướng có thể giúp xác định hướng đi của thị trường. Một vài chỉ báo Trend phổ biến:

- Đường trung bình đơn giản (Simple Moving Average - SMA): là số trung bình cộng của giá đóng cửa thị trường trong khoảng thời gian nhất định sau khi đã loại bỏ các yếu tố bất thường. Do đó, đường SMA được đánh giá mang lại hiệu quả cao cho các nhà đầu tư khi có nhu cầu xác định sự biến đổi về giá.

- Đường trung bình lũy thừa (Exponential Moving Average - EMA): được tính bằng công thức hàm mũ, trong đó đặt nặng các biến động giá gần nhất. Do đó, EMA khá nhạy cảm với các biến động ngắn hạn, nhận biết các tín hiệu thất thường tốt hơn đường SMA giúp nhà đầu tư phản ứng nhanh hơn trước các biến động giá ngắn hạn.

- Đường trung bình tỷ trọng tuyến tính (Weighted Moving Average - WMA): WMA chú trọng hơn vào các tham số có tần suất xuất hiện cao nhất. Tức là sẽ đặt nặng các bước giá có khối lượng giao dịch lớn, quan tâm đến yếu tố chất lượng của dòng tiền.

- Đường chuyển động trung bình phân kỳ hội tụ (Moving Average Convergence Divergence - MACD): Chỉ báo MACD được thiết kế để xác định rõ độ mạnh yếu, hướng, động lượng và thời gian của một xu hướng giá tăng hay giảm. MACD là một trong những chỉ báo có thể xác định chính xác giá trị mà nó tạo ra thông qua 2 yếu tố chính là hội tụ, phân kỳ, đây cũng là hai yếu tố mà các nhà đầu tư rất quan tâm.

Chỉ báo Momentum

Động lượng trong giao dịch đề cập đến tốc độ thay đổi giá. Các chỉ báo Momentum đo tốc độ này, giúp nhà đầu tư có thể thấy sự thay đổi xu hướng sắp tới. Một vài chỉ báo Momentum phổ biến:

- Chỉ số sức mạnh tương đối (Relative Strength Index - RSI): 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.

- Chỉ số sức mạnh thực sự (True strength index - TSI): là chỉ báo xung lượng dao động trong khoảng -100 và +100 và có giá trị cơ sở là 0. Chỉ báo có thể hữu ích để xác định tình trạng quá mua hoặc quá bán, giao cắt với đường trung tâm, giao cắt với đường phân kì và đường tín hiệu.

1.2 Bài toán dự đoán xu hướng chứng khoán

Phân tích chứng khoán đã trở thành công việc hết sức quan trọng mà các nhà đầu tư chứng khoán cần thực hiện để có được quyết định đầu tư thích hợp. Từ kết quả phân tích, người đầu tư chứng khoán sẽ quyết định khi nào thì mua vào, khi nào thì bán ra, khi nào thì giữ lại và nên đầu tư vào doanh nghiệp nào.

Có hai phương pháp phân tích được sử dụng phổ biến hiện nay ở hầu hết các thị trường chứng khoán trên thế giới, đó là phân tích cơ bản và phân tích kỹ thuật.

- Phân tích cơ bản (Fundamental Analysis): là phương pháp phân tích chứng khoán dựa trên các nhân tố mang tính chất nền tảng có tác động hoặc là nguyên nhân dẫn tới sự thay đổi giá của chứng khoán. Các nhân tố cần chú trọng trong phân tích cơ bản là: hoạt động kinh doanh của công ty, khả năng lợi nhuận (hiện tại và ước đoán), kết quả sản xuất kinh doanh, chất lượng quản lý, sức ép cạnh tranh, chính sách giá cả, vị thế,...
- Phân tích kỹ thuật (Technical Analysis): sử dụng các mô hình toán học (đồ thị, biến đổi miền, xác suất thống kê, dãy đại số,...) dựa trên dữ liệu thu thập về thị trường trong quá khứ và hiện tại để chỉ ra trạng thái của thị trường trong một thời điểm xác định, thông thường là nhận định xu hướng thị trường.

Các kỹ thuật phân tích cho thấy rằng giá trong lịch sử và các chỉ số khác có thể tiết lộ mối tương quan và các mô hình biến động giá chứng khoán, do đó có thể dự đoán giá chứng khoán trong tương lai. Trong những năm gần đây, nhiều kỹ thuật tiên tiến như giải thuật di truyền (GA), máy véctơ hỗ trợ (SVM), mạng nơron nhân tạo (ANN) đã hỗ trợ tốt việc phân tích. Theo Mackinlay, giá chứng khoán có mối tương quan nhạy cảm với tin tức và các sự kiện mang lại thông tin cho thị trường chứng khoán (kể cả các tin tức thời sự, kinh tế, chính trị, thời tiết...đều ảnh hưởng tới thị trường chứng khoán). Với hướng tiếp cận này, sử dụng các kỹ thuật học máy và khai phá dữ liệu để tìm ra mối tương quan giữa giá trong quá khứ và xu hướng giá trong tương lai.

Thông tin của các giao dịch chứng khoán được lưu trữ lại dưới dạng dữ liệu chuỗi thời gian (time series). Dữ liệu thời gian hay chuỗi thời gian là một chuỗi các giá trị của một đại lượng nào đó được ghi nhận theo thời gian. Các mô hình học máy như Mạng thần kinh tái phát (RNN) hoặc Mạng bộ nhớ dài-ngắn hạn (LSTM) là những mô hình phổ biến được áp dụng để dự đoán dữ liệu chuỗi thời

gian như dự báo thời tiết, kết quả bầu cử, giá nhà và tất nhiên là giá chứng khoán.

Đây là kho dữ liệu khổng lồ để chúng ta có thể khai phá và dự báo xu thế của thị trường chứng khoán. Các dữ liệu lịch sử giao dịch này thường được lưu trữ bao gồm các thông tin về thời gian (Date), giá mở cửa (Open), giá đóng cửa (Close), giá cao nhất trong ngày (High), giá thấp nhất trong ngày (Low) và khối lượng giao dịch (Volume).

Ý tưởng giải quyết bài toán dự báo giá chứng khoán là sử dụng những dữ liệu trong quá khứ, đồng thời xác định tham số nào ảnh hưởng nhiều nhất đến giá "hiện tại" hoặc "tiếp theo". Trong Đồ án này, tôi sẽ sử dụng mô hình máy học LSTM kết hợp với phương pháp Phân tích thành phần chính (PCA) để dự báo.

1.3 Tình hình nghiên cứu

Dự báo thị trường chứng khoán từ lâu đã thu hút nhiều nghiên cứu từ các nhà kinh tế học và các nhà khoa học máy tính. Gần đây, dự báo thị trường chứng khoán sử dụng các mô hình máy học là một lĩnh vực mới nổi và đã thu hút một số nghiên cứu trên thế giới nói chung và Việt Nam nói riêng. Hiện nay, nhiều nghiên cứu nước ngoài đã sử dụng phương pháp dự báo chuỗi thời gian bằng cách kết hợp các mô hình học máy truyền thống và học sâu.

Liu và các cộng sự (2017) đã sử dụng mạng RNN để dự báo sự biến động của chứng khoán. Kết quả thu được mô hình mạng RNN tốt hơn mô hình mạng nhiều tầng truyền thẳng (MLP) và mô hình máy véc tơ hỗ trợ (SVM). Bên cạnh đó, Gao, Chai và Liu (2017) đã sử dụng bộ dữ liệu về lịch sử giao dịch chứng khoán Chỉ số cổ phiếu 500 của Standard & Poor (S&P500) trong 20 ngày và dự báo thị trường chứng khoán bằng bốn phương pháp khác nhau: trung bình trượt (MA), trung bình trượt hàm mũ (EMA), SVM và LSTM. Kết quả cho thấy LSTM có độ chính xác cao nhất.

Khi dự báo doanh số bán của các công ty bất động sản, Soy Temür, Akgün, và Temür (2019) đã sử dụng các mô hình ARIMA, LSTM và ARIMA-LSTM

cùng với bộ số liệu gồm 124 tháng, từ tháng 01 năm 2008 đến tháng 04 năm 2018 về tổng doanh thu bán nhà ở Thổ Nhĩ Kỳ. Kết quả là mô hình kết hợp ARIMA-LSTM có phần trăm sai số tuyệt đối trung bình (MAPE) và sai số bình phương trung bình (MSE) là thấp nhất. Hyeong Kyu Choi (2018) cũng sử dụng sự kết hợp này trong việc dự báo hệ số tương quan về giá của hai chứng khoán riêng biệt. Kết quả là mô hình kết hợp ARIMA-LSTM có khả năng dự báo vượt trội so với các mô hình truyền thống.

Tsai và Hsiao đã đề xuất một giải pháp kết hợp của các phương pháp Lựa chọn đặc trưng (Feature Selection) để dự đoán dữ liệu chứng khoán từ năm 2000 đến năm 2007 lấy từ cơ sở dữ liệu Tạp chí Kinh tế Đài Loan (TEJ) làm nguồn dữ liệu. Phương pháp mà họ sử là Sliding window kết hợp với mạng MLP. Họ cũng áp dụng phương pháp phân tích thành phần chính (PCA) để giảm kích thước mẫu, thuật toán di truyền (GA) và cây phân loại-hồi quy (CART) để lựa chọn các đặc trưng quan trọng. Họ không chỉ dựa vào các chỉ số kỹ thuật mà cũng dựa vào bao gồm cả các chỉ số cơ bản và kinh tế vĩ mô.

Tại Việt Nam, những nghiên cứu tương tự còn khá mới mẻ. Phạm Nguyễn Hoàng Phúc và Trương Tấn Phát (2020) đã sử dụng LSTM dự báo biến động của thị trường giao dịch tài chính.

Chương 2

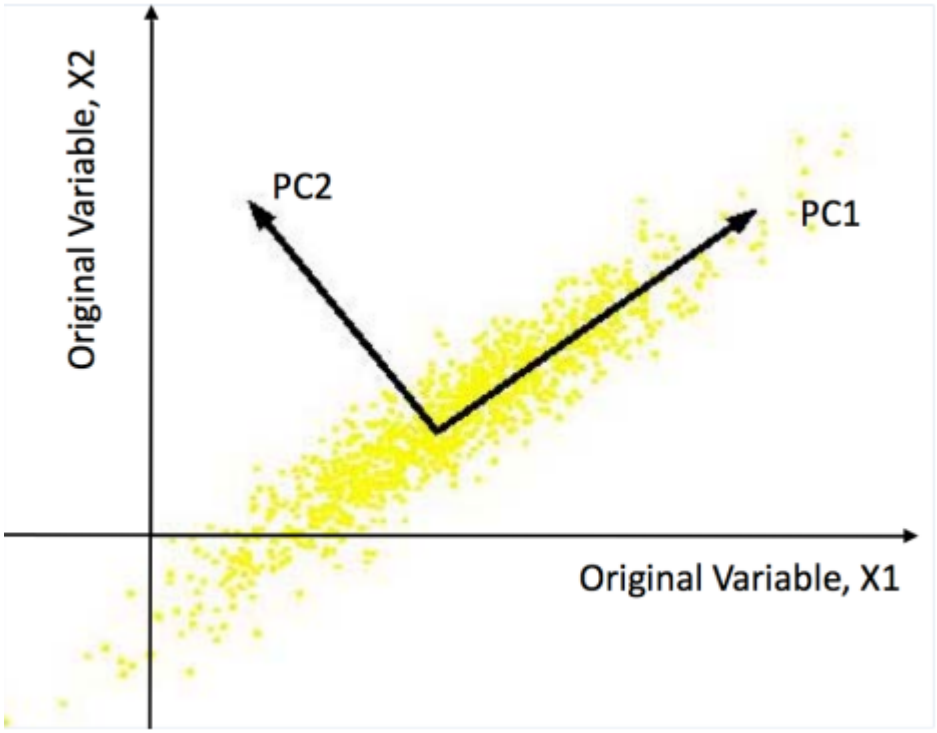
Cơ sở lý thuyết

2.1 Thuật toán Phân tích thành phần chính

Phân tích thành phần chính (PCA) là một phương pháp giảm kích thước thống kê được sử dụng để giảm độ phức tạp của tập dữ liệu đồng thời giảm thiểu mất mát thông tin. Nó chuyển đổi một tập dữ liệu trong đó có một số lượng lớn các biến có liên quan đến nhau thành một tập hợp mới gồm các biến không tương quan được gọi là các thành phần chính và được sắp xếp theo trình tự với thành phần đầu tiên giải thích càng nhiều biến thể càng tốt. Mỗi thành phần chính là sự kết hợp tuyến tính của các biến ban đầu, trong đó các hệ số chỉ ra tầm quan trọng tương đối của biến trong thành phần.

PCA rất hữu ích khi có một số lượng lớn các kích thước tương quan chứa nhiều dữ liệu dư thừa. PCA có thể được sử dụng để giảm sự dư thừa này, dẫn đến việc giảm dữ liệu có độ tương quan cao thành một số lượng nhỏ các thành phần chính không tương quan và chiếm phần lớn phương sai [8].

Xét một ví dụ không thực tế nhưng đơn giản, giả sử có hai biến và dữ liệu được vẽ theo hai chiều. Hình 2.1 hiển thị diễn giải hình học của phân tích thành phần chính trong trường hợp hai biến. x_1 và x_2 đại diện cho các biến và trục ban đầu. PC_1 và PC_2 là các biến và trục được biến đổi. Hướng của các trục chính biểu thị các thành phần chính. Thành phần chính đầu tiên biểu diễn tổ hợp



Hình 2.1: Diễn giải hình học của phân tích thành phần chính [8]

tuyến tính $\alpha_1' x$ giải thích nhiều biến thể nhất, đó là:

$$PC_1 = \alpha_1' x = \alpha_{11} x_1 + \alpha_{12} x_2$$

trong đó $\alpha_1' x$ là giá trị riêng của thành phần chính thứ nhất PC_1, x_1 và x_2 là hai biến ban đầu, α_{1i} là hệ số của chứng khoán i trong thành phần chính thứ nhất.

Tương tự, thành phần chính thứ hai biểu diễn một tổ hợp tuyến tính $\alpha_2' x$, giải thích phần lớn biến thể còn lại; tuy nhiên thành phần chính thứ 2 phải không trực giao với thành phần chính ban đầu hay thành phần chính thứ 2 không có mối tương quan tuyến tính với thành phần chính đầu tiên. Trong trường hợp có hai biến, thành phần chính thứ 2 được biểu diễn như sau:

$$PC_2 = \alpha_2' x = \alpha_{21} x_1 + \alpha_{22} x_2$$

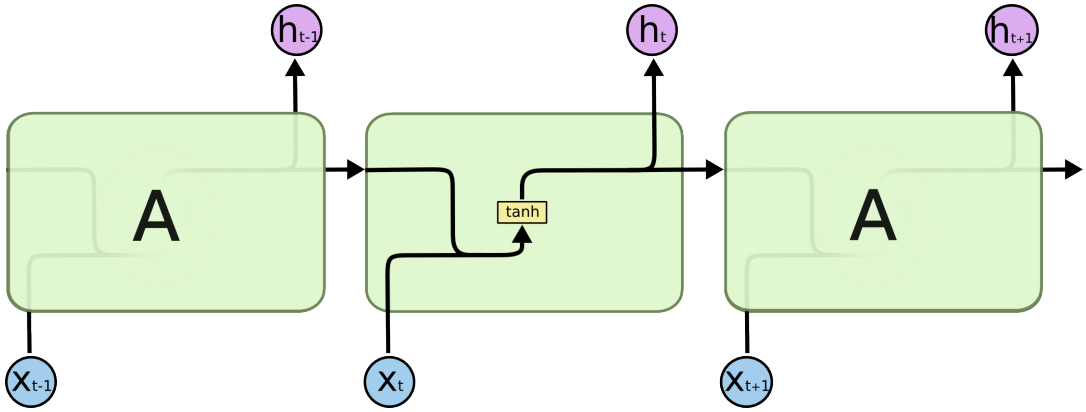
Rõ ràng, trong Hình 2.1, hướng của thành phần chính một biến thiên nhiều hơn so với hai biến ban đầu và có ít biến đổi hơn theo hướng của thành phần chính thứ hai.

2.2 Lý thuyết mạng ngắn-dài hạn (LSTM)

LSTM là một mạng cải tiến của RNN nhằm giải quyết vấn đề nhớ các bước dài của RNN. LSTM được giới thiệu bởi Hochreiter và Schmidhuber (1997), và sau đó đã được cải tiến và phổ biến bởi rất nhiều người trong ngành. Chúng hoạt động cực kì hiệu quả trên nhiều bài toán khác nhau nên dần đã trở nên phổ biến như hiện nay.

LSTM được thiết kế để tránh được vấn đề phụ thuộc xa. Việc ghi nhớ thông tin trong suốt thời gian dài là đặc tính mặc định của chúng, chứ ta không cần phải huấn luyện để có thể nhớ được. Tức là ngay nội tại của nó đã có thể ghi nhớ được mà không cần bất kì can thiệp nào.

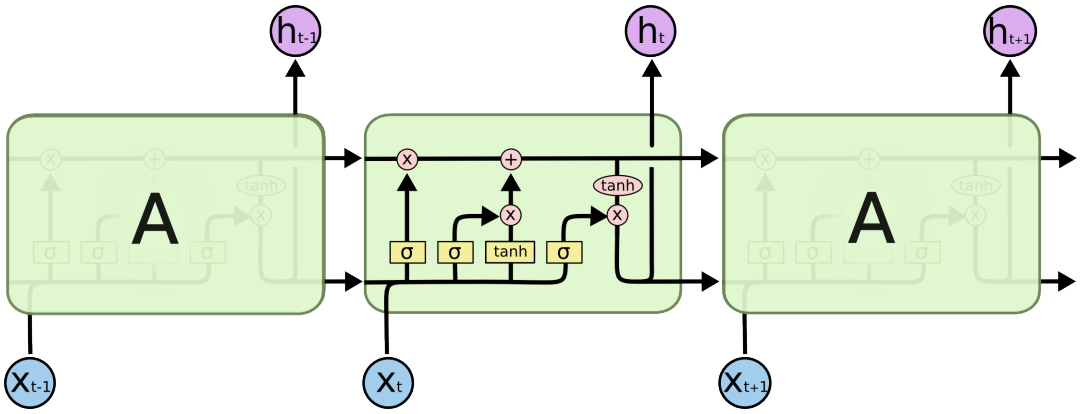
Mọi mạng hồi quy đều có dạng là một chuỗi các mô-đun lặp đi lặp lại của mạng nơ-ron. Với mạng RNN chuẩn, các mô-đun này có cấu trúc rất đơn giản, thường là một tầng tanh.



Hình 2.2: Cấu trúc mô-đun lặp lại trong mạng RNN tiêu chuẩn chứa một tầng duy nhất. (Colah, 2015, `https://colah.github.io/posts/2015-08-Understanding-LSTMs/`)

LSTM cũng có kiến trúc dạng chuỗi như vậy, nhưng các mô-đun trong nó có cấu trúc khác với mạng RNN chuẩn. Thay vì chỉ có một tầng mạng nơ-ron, chúng có tới 4 tầng tương tác với nhau một cách rất đặc biệt.

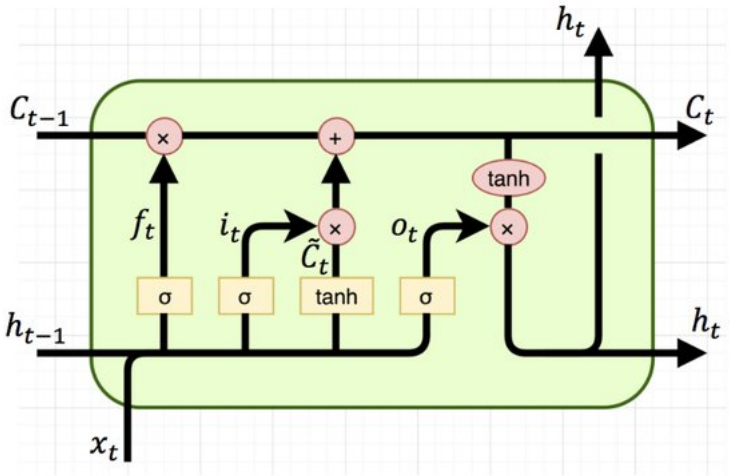
Ở sơ đồ 2.3, mỗi một đường mang một véc-tơ từ đầu ra của một nút tới đầu vào của một nút khác. Các hình trong màu hồng biểu diễn các phép toán như phép cộng véc-tơ chẳng hạn, còn các ô màu vàng được sử dụng để học trong các



Hình 2.3: Cấu trúc mô-đun lặp lại trong mạng LSTM chứa bốn tầng tương tác. (Colah, 2015, `https://colah.github.io/posts/2015-08-Understanding-LSTMs/`)

từng mạng nơ-ron. Các đường hợp nhau kí hiệu việc kết hợp, còn các đường rẽ nhánh ám chỉ nội dung của nó được sao chép và chuyển tới các nơi khác nhau.

Một đơn vị LSTM tiêu chuẩn bao gồm ba cổng: cổng đầu vào, đầu ra và cổng quên. Các cổng này tìm hiểu trọng số của chúng và xác định nên nhớ bao nhiêu mẫu dữ liệu hiện tại và bao nhiêu nội dung đã học trong quá khứ nên bị lãng quên. Cấu trúc đơn giản này là một cải tiến so với mô hình RNN trước đó.



Hình 2.4: Cấu trúc bốn tầng LSTM. (Colah, 2015, `https://colah.github.io/posts/2015-08-Understanding-LSTMs/`)

Như đã thấy trong hình 2.4, f, i và o đại diện cho ba cổng: đầu vào, đầu ra và cổng quên. C là trạng thái ô (cell state) lưu trữ dữ liệu đã học, được đưa ra

dưới dạng đầu ra là trạng thái ẩn (hidden state) h. Tất cả dữ liệu được tính cho mốc thời gian t, xét theo dữ liệu đã học từ mốc thời gian $(t-1)$.

Bước đầu tiên của LSTM là quyết định xem thông tin nào cần bỏ đi từ trạng thái tế bào. Quyết định này được đưa ra bởi tầng sigmoid - gọi là tầng cổng quên. Nó sẽ lấy đầu vào là h_{t-1} và x_t rồi đưa ra kết quả là một số trong khoảng [0,1] cho mỗi số trong trạng thái tế bào C_{t-1}. Đẩu ra là 1 thể hiện rằng nó giữ toàn bộ thông tin lại, còn đầu ra là 0 chỉ rằng toàn bộ thông tin sẽ bị bỏ đi.

$$f_t = \sigma(x_t U^f + h_{t-1} W^f)$$

Bước tiếp theo là quyết định xem thông tin mới nào ta sẽ lưu vào trạng thái tế bào. Việc này gồm 2 phần: Đầu tiên là sử dụng một tầng sigmoid được gọi là tầng cổng vào để quyết định giá trị nào ta sẽ cập nhập; Tiếp theo, sử dụng một tầng tanh tạo ra một véc-tơ cho giá trị mới $\tilde{C}_t$ nhằm sửa đổi bộ nhớ.

$$i_t = \sigma(x_t U^i + h_{t-1} W^i)$$
$$\tilde{C}_t = \tanh(x_t U^g + h_{t-1} W^)$$

Tiếp theo, ta sẽ cập nhập trạng thái tế bào cũ C_{t-1} thành trạng thái mới C_t. Ta sẽ nhân trạng thái cũ với f_t để bỏ đi những thông tin ta quyết định quên lúc trước. Sau đó cộng thêm $i_t * \tilde{C}_t$. Trạng thái mới thu được này phụ thuộc vào việc ta quyết định cập nhập mỗi giá trị trạng thái ra sao.

$$C_t = \sigma(f_t * C_{t-1} + i_t * \tilde{C}_t)$$

Cuối cùng, ta cần quyết định xem ta muốn đầu ra là gì. Giá trị đầu ra sẽ dựa vào trạng thái tế bào C_t, nhưng sẽ được tiếp tục sàng lọc. Đầu tiên, ta chạy một tầng sigmoid để quyết định phần nào của trạng thái tế bào ta muốn xuất ra. Sau đó, ta đưa trạng thái tế bảo qua một hàm *tanh* để co giá trị nó về khoảng [-1,1], và nhân với đầu ra của cổng sigmoid để được giá trị đầu ra mong muốn.

$$o_t = \sigma(x_t U^o + h_{t-1} W^o)$$

$$h_t = o_t * \tanh(C_t)$$

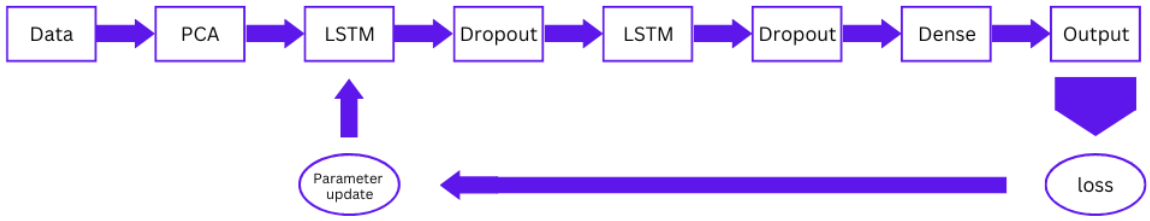
Từ cách thức hoạt động trên, LSTM được đánh giá là vượt trội hơn so với RNN. LSTM có thể truy vấn được thông tin từ một tập thông tin lớn hơn. Vì vậy, LSTM rất thích hợp để dự báo chuỗi thời gian trong dài hạn.

Chương 3

Mô hình dự đoán

3.1 Kiến trúc mô hình PCA – LSTM kết hợp

Để giải quyết vấn đề dự đoán giá chứng khoán theo dữ liệu lịch sử của N ngày giao dịch, một mô hình học sâu bao gồm hai lớp mô-đun LSTM và một lớp Dropout được đặt ở giữa để tránh over-fitting. Cuối cùng, một lớp Dense được thêm vào mô hình để đưa ra một giá trị cụ thể. Quá trình mô hình hoạt động được trình bày trong hình 3.1 dưới đây.



Hình 3.1: Quá trình hoạt động của mô hình

3.1.1 Tiền xử lý dữ liệu

Trong nghiên cứu, có nhiều yếu tố ảnh hưởng đến giá chứng khoán. Chúng có thể tương quan với nhau và có tác động khác nhau đến kết quả. Phương pháp PCA có thể tập trung các yếu tố ảnh hưởng này vào một số thành phần chính, sao cho các thành phần chính này phản ánh càng nhiều thông tin càng tốt so với các biến ban đầu, loại bỏ sự dư thừa dữ liệu và giảm kích thước dữ liệu.

3.1.2 Dự báo giá chứng khoán

Như đã trình bày trong chương trước, hầu hết dữ liệu của thị trường chứng khoán là dữ liệu chuỗi thời gian và mạng thần kinh LSTM có lợi thế rõ ràng trong việc xử lý thông tin chuỗi thời gian. LSTM đã được ứng dụng thành công trong nhiều lĩnh vực như xử lý ảnh, nhận dạng giọng nói, nhận dạng chữ viết tay,... Do đó, Đồ án sẽ ứng dụng mạng thần kinh LSTM vào mô hình dự báo giá chứng khoán.

Kỹ thuật Drop Out

Drop Out là việc bỏ qua các đơn vị (tức là 1 nút mạng) trong quá trình đào tạo 1 cách ngẫu nhiên. Trong neural network, ta cần tối ưu các tham số để làm giảm giá trị hàm loss, nhưng đôi khi có một vài đơn vị thay đổi và ảnh hưởng tới các đơn vị khác dẫn đến việc over-fitting làm giảm tính dự đoán của mô hình.

Tại mỗi bước trong quá trình học, khi thực hiện Forward Propagation (Lan truyền xuôi) đến layer sử dụng Drop Out, thay vì tính toán tất cả các đơn vị có trên layer, ta sẽ "gieo xúc xắc" xem đơn vị đó có được tính hay không dựa trên xác suất p. Theo đó, p được gọi là xác suất giữ lại 1 nút mạng trong mỗi giai đoạn huấn luyện, vì thế xác suất nó bị loại bỏ là (1 - p).

Dense Layer

Dense layer là một trong những lớp có sẵn trong framework Keras, thường được thêm vào trong neural network. Lớp này chứa các nơ-ron được kết nối dày đặc. Mỗi nơ-ron nhận đầu vào từ tất cả nơ-ron của lớp trước đó.

Bên trong, Dense Layer là nơi thực hiện các phép nhân khác nhau của các vectơ ma trận. Chúng ta có thể huấn luyện các giá trị bên trong ma trận vì chúng thực chất chỉ là các tham số.

Dense Layer tạo ra kết quả đầu ra dưới dạng vectơ có kích thước m chiều. Đây là lý do tại sao Dense Layer thường được sử dụng nhất để thao tác với vectơ nhằm thay đổi kích thước hoặc thực hiện các phép toán vectơ. Do đầu ra của

LSTM layer là một véctơ nên ta dễ dàng truyền qua một Dense Layer.

3.2 Phương pháp đánh giá

Trong phần này, chúng ta sẽ giới thiệu các hàm mất mát để đánh giá chất lượng mô hình dự đoán. Cụ thể trong Đồ án này, Sai số trung bình tuyệt đối (MAE) và Sai số phần trăm trung bình tuyệt đối (MAPE) được chọn để đánh giá định lượng hiệu suất của mô hình PCA-LSTM.

3.2.1 Sai số trung bình tuyệt đối

Sai số trung bình tuyệt đối (MAE) đo độ lớn trung bình của các lỗi trong một tập hợp các dự đoán mà không cần xem xét hướng của chúng. Đó là giá trị trung bình trên mẫu thử nghiệm về sự khác biệt tuyệt đối giữa dự đoán và quan sát thực tế, trong đó tất cả các khác biệt riêng lẻ có trọng số bằng nhau. Ta có công thức MAE cho bài toán dự báo giá chứng khoán như sau:

$$MAE = \frac{\sum_{i=1}^{n} |\overline{x_i} - x_i|}{n}$$

trong đó,

$\overline{x_i}$: giá trị dự đoán giá chứng khoán,

x_i : dữ liệu thực tế giá chứng khoán,

n : số ngày dùng để dự đoán.

3.2.2 Sai số phần trăm trung bình tuyệt đối

Sai số phần trăm trung bình tuyệt đối (MAPE) đo lường độ chính xác của một mô hình dự báo. Nó đo lường độ chính xác theo tỷ lệ phần trăm và có thể được tính bằng sai số phần trăm tuyệt đối trung bình cho mỗi khoảng thời gian trừ đi các giá trị thực tế chia cho các giá trị thực tế.

MAPE là thước đo phổ biến nhất được sử dụng để dự báo sai số, có thể là do các đơn vị của biến được chia tỷ lệ thành đơn vị phần trăm, giúp biến trở nên

dễ hiểu hơn. Nó hoạt động tốt nhất nếu không có dữ liệu cực đoan (và không có số không). Nó thường được sử dụng như một hàm mất mát trong phân tích hồi quy và đánh giá mô hình. Ta có công thức MAPE cho mô hình dự báo giá chứng khoán như sau:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}|\frac{\overline{x_i} - x_i}{x_i}|,$$

trong đó,

$\overline{x_i}$: giá trị dự đoán giá chứng khoán,

x_i : dữ liệu thực tế giá chứng khoán,

N : số ngày dùng để dự đoán.

Chương 4

Thực nghiệm

4.1 Mô tả dữ liệu

Đồ án sử dụng dữ liệu chứng khoán của tập đoàn Google từ ngày 11/01/2018 đến ngày 11/01/2023 (1259 ngày giao dịch). Dữ liệu bao gồm 6 chỉ số cơ bản ảnh hưởng đến giá của chứng khoán như đã trình bày tại phần 1.1.2.

Date	Open	High	Low	Close	Adj Close	Volume
2018-01-11	55.314999	55.326248	54.979500	55.276001	55.276001	19566000
2018-01-12	55.120499	56.214500	55.057499	56.112999	56.112999	34410000
2018-01-16	56.625500	56.995499	55.891602	56.088001	56.088001	31506000
2018-01-17	56.311001	56.630001	55.850498	56.598999	56.598999	24052000
2018-01-18	56.570499	56.625500	55.875000	56.489498	56.489498	23964000

Hình 4.1: Tập dữ liệu gốc

Sau khi sử dụng thư viện có sẵn để tính toán các chỉ báo kỹ thuật, ta thu được bộ dữ liệu mới như hình 4.2.

4.2 Tiền xử lý dữ liệu

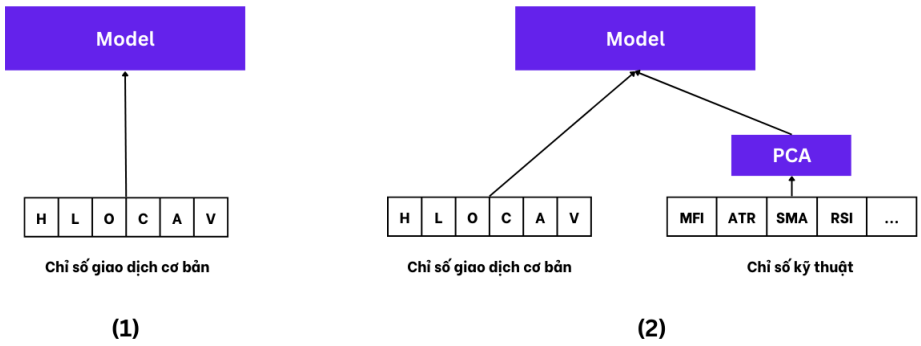
Trong mô hình này, chỉ một số biến đầu vào được ứng dụng phương pháp. Các biến đầu vào cơ bản sẽ không được xử lý (các biến High, Low, Open, Close,

Date	Open	High	Low	Close	Adj Close	Volume	volume_adi	volume_obv	volume_cmf	volume_fi	...	momentum_ppo
2018-01-11	55.314999	55.326248	54.979500	55.276001	55.276001	19566000	1.389541e+07	19566000	NaN	NaN	...	NaN
2018-01-12	55.120499	56.214500	55.057499	56.112999	56.112999	34410000	4.226799e+07	53976000	NaN	NaN	...	NaN
2018-01-16	56.625500	56.995499	55.891602	56.088001	56.088001	31506000	2.197272e+07	22470000	NaN	NaN	...	NaN
2018-01-17	56.311001	56.630001	55.850498	56.598999	56.598999	24052000	4.411156e+07	46522000	NaN	NaN	...	NaN
2018-01-18	56.570499	56.625500	55.875000	56.489498	56.489498	23964000	5.939027e+07	22558000	NaN	NaN	...	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...
2023-01-04	91.010002	91.239998	87.800003	88.709999	88.709999	27046500	8.346381e+08	1184943900	-0.216589	-5.184483e+06	...	-2.203422
2023-01-05	88.070000	88.209999	86.559998	86.769997	86.769997	23136100	8.173912e+08	1161807800	-0.221787	-1.085585e+07	...	-2.305211
2023-01-06	87.360001	88.470001	85.570000	88.160004	88.160004	26604400	8.383078e+08	1188412200	-0.130318	-4.022118e+06	...	-2.236191
2023-01-09	89.195000	90.830002	88.580002	88.800003	88.800003	22996700	8.198083e+08	1211408900	-0.124042	-1.344977e+06	...	-2.099875
2023-01-10	86.720001	89.474998	86.699997	89.239998	89.239998	22846300	8.387851e+08	1234255200	-0.042847	2.831992e+05	...	-1.929929

Hình 4.2: Tập dữ liệu mới thu được

Volume). Các dữ liệu chỉ báo kỹ thuật sẽ được giảm số lượng để trích xuất các tính năng quan trọng và tương quan bằng phương pháp PCA. Cụ thể, trong nghiên cứu này, tôi sẽ so sánh 2 mô hình với 2 đầu vào được xử lý khác nhau: (1) Đầu vào bao gồm các chỉ số cơ bản và các chỉ báo kỹ thuật. Tất cả các biến đầu vào không được xử lý bởi PCA (2) Đầu vào bao gồm các chỉ số cơ bản và các chỉ báo kỹ thuật. PCA chỉ ứng dụng xử lý các chỉ báo kỹ thuật.

Các phương thức xử lý đầu vào khác nhau này được thể hiện trong hình 4.3.



Hình 4.3: 2 phương thức xử lý đầu vào

Ngoài ra, các biến đầu vào cần được thay đổi thành dữ liệu chuỗi. Việc phân

đoạn dữ liệu đầu vào được thực hiện bằng một cửa sổ trượt (sliding window) trong 30 ngày. Quá trình này được mô tả trong Hình 2. Một cửa sổ trượt được áp dụng cho toàn bộ tập dữ liệu để trích xuất dữ liệu đầu vào mà mô hình dự báo sử dụng. Sau đó, mỗi biến đầu vào có 30 ngày quan sát. Khối màu xanh đại diện cho dữ liệu đầu vào, bao gồm dữ liệu dạng 30 ngày giao dịch. Khối màu trắng thể hiện dữ liệu đầu ra, giá đóng cửa của ngày hôm sau. Các mẫu đào tạo và mẫu thử nghiệm được lấy tuần tự.

4.3 Xây dựng mô hình thực nghiệm

Tôi sẽ dự đoán giá đóng cửa ngày hôm sau của tập đoàn GOOGLE, quá trình học và thử nghiệm sẽ được lặp lại với chu kỳ 400 ngày. Đối với mỗi loại tập kiểm tra, dữ liệu còn lại là dữ liệu huấn luyện. Quy trình đào tạo học máy có thể được chia thành ba phần: xử lý dữ liệu (chuẩn hóa dữ liệu), khởi tạo trọng số và tối ưu hóa tham số.

Xử lý dữ liệu

Trong Đồ án này, xử lý dữ liệu chỉ đề cập đến việc chuẩn hóa dữ liệu. Dễ nhận thấy, đơn vị đo khác nhau có thể ảnh hưởng đến việc phân tích dữ liệu. Ví dụ: thay đổi đơn vị đo từ mét sang inch cho chiều cao hoặc từ kilôgam sang pound cho cân nặng, có thể dẫn đến các kết quả rất khác nhau. Để tránh sự phụ thuộc vào việc lựa chọn đơn vị đo, dữ liệu nên được chuẩn hóa trước khi đưa vào mô hình. Điều này liên quan đến việc chuyển đổi dữ liệu để nằm trong một phạm vi nhỏ hơn hoặc phổ biến hơn, chẳng hạn như nằm trong khoảng [-1, 1] hoặc [0, 1].

Hiện tại có một vài phương pháp để chuẩn hóa dữ liệu như chuẩn hóa min-max, chuẩn hóa z-score, chuẩn hóa bởi thang chia 10. Trong mô hình này, tôi sẽ sử dụng phương pháp chuẩn hóa min-max.

Quá trình chuẩn hóa min-max thực hiện phép biến đổi tuyến tính trên dữ liệu gốc. Quá trình này sẽ ánh xạ giá trị ν_i thành giá trị đã được chuẩn hóa ν_i'

sao cho $\nu_i^{'} \in [new_{min}, new_{max}]$ [10]. Phương trình chuẩn hóa min-max được trình bày như sau:

$$\nu_i^{'} = \frac{\nu_i - \nu_{min}}{\nu_{max} - \nu_{min}}.(new_{max} - new_{min}) + new_{min}$$

trong đó,

$$\nu = (\nu_1, \nu_2, ..., \nu_n),$$

$\nu_i^{'}$: dữ liệu chuẩn hóa thứ i,

$$new_{max} = 1, \quad new_{min} = 0,$$

Ở bước cuối cùng của quá trình xử lý, tôi sẽ trộn dữ liệu đào tạo một cách ngẫu nhiên sau mỗi giai đoạn huấn luyện.

Khởi tạo trọng số

Trong mô hình dự đoán, hai phương pháp được sử dụng để khởi tạo trọng số. Tôi sử dụng Glorot uniform làm phương thức khởi tạo của ma trận đầu vào và khởi tạo trực giao cho ma trận hồi quy.

Khởi tạo Glorot uniform tạo ra các trọng số và độ lệch ngẫu nhiên bằng cách lấy mẫu từ hàm phân phối đồng nhất (Glorot & Bengio, 2010). Trong mô hình này, các trọng số truyền thẳng được khởi tạo bằng cách lấy mẫu từ quy trình khởi tạo phân phối đồng nhất để đáp ứng mục tiêu duy trì phương sai của hàm kích hoạt và phương sai độ dốc lan truyền ngược khi di chuyển lên hoặc xuống lớp mạng:

$$W \sim U[-\frac{\sqrt{6}}{\sqrt{numUnits1 + numUnits2}}, \frac{\sqrt{6}}{\sqrt{numUnits1 + numUnits2}}],$$

trong đó,

$numUnits1 =$ số lượng đơn vị thần kinh trong lớp mạng thấp

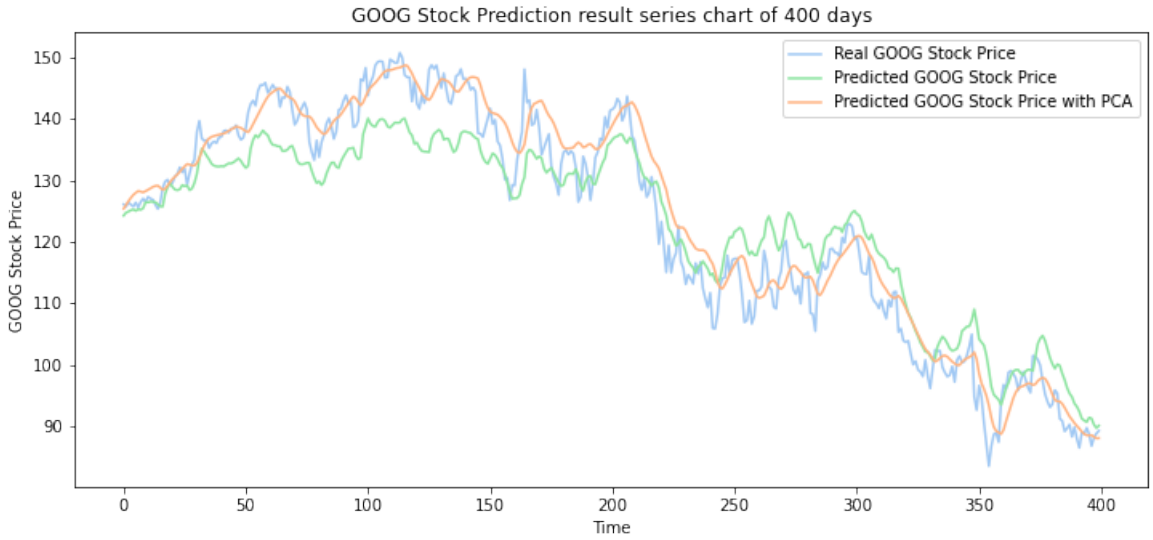
$numUnits2 =$ số lượng đơn vị thần kinh trong lớp mạng cao

Trong phần tử bên trong của lớp LSTM, việc khởi tạo sử dụng phương pháp ma trận trực giao tạo ra một ma trận trực giao ngẫu nhiên. Phương pháp này đã được chứng minh là có hiệu suất tốt trong RNN (Le, Jaitly, & Hinton, 2015). Ma trận trực giao có nhiều tính chất thú vị, nhưng tính chất quan trọng nhất là tất cả các giá trị riêng của nó có giá trị tuyệt đối bằng một. Điều này có nghĩa là dù ta thực hiện phép nhân ma trận lặp đi lặp lại thì ma trận kết quả cũng không bị phá vỡ. Điều này cho phép việc tính toán gradient bằng thuật toán lan truyền ngược trở nên hiệu quả hơn.

Thuật toán tối ưu hóa để cập nhật tham số

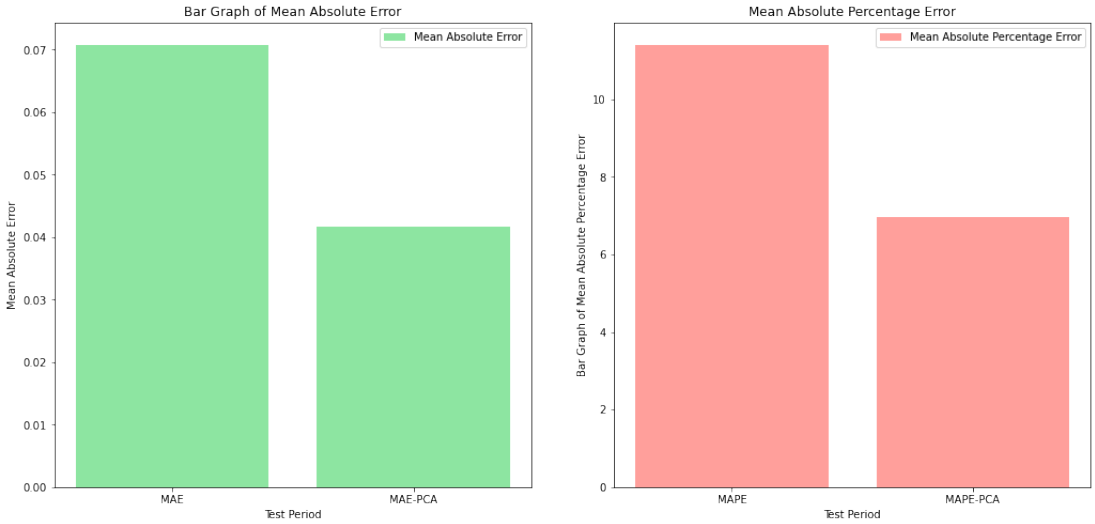
Mô hình sử dụng tối ưu Adam để thực hiện tối ưu hóa. Phương pháp này hiệu quả về mặt tính toán, ít yêu cầu bộ nhớ và không thay đổi tỷ lệ theo đường chéo của các gradient (Kingma & Ba, 2014). Cốt lõi của thuật toán Adam là ước tính thời điểm đầu tiên và thứ hai của độ dốc để thực hiện cập nhật tham số.

4.4 Đánh giá kết quả



Hình 4.4: Kết quả thực nghiệm (Mẫu thử nghiệm = 400)

Đồ thị trong hình 4.4 mô tả kết quả dự đoán giá chứng khoán theo 2 phương thức xử lý đầu vào khác nhau.



Hình 4.5: Tính toán MAE và MAPE của mô hình

Hình 4.5 tính toán riêng MAE và MAPE theo các đầu vào khác nhau và so sánh hiệu quả của từng mô hình.

Các kết quả ở trên phản ánh mô hình mục tiêu (mô hình LSTM kết hợp ứng dụng PCA tiền xử lý dữ liệu) có khả năng dự đoán giá chứng khoán trong tương lai tốt hơn và hiệu quả hơn. Bên cạnh đó, chỉ số MAE và MAPE của mô hình được tính toán tương đối thấp so với mô hình không xử dụng PCA, vì vậy xu hướng dự đoán của mô hình về giá chứng khoán là phù hợp với thực tế.

Kết luận

Mô hình LSTM kết hợp PCA cho chúng ta thấy hiệu quả trong việc dự đoán giá chứng khoán, tuy nhiên nhược điểm của nó cũng rất rõ ràng. Nếu muốn thu được kết quả dự đoán xu hướng giá có độ chính xác cao, chúng ta có thể phải chạy mô hình rất nhiều lần và tốn rất nhiều thời gian. Mô hình này có thể không dự đoán được chính xác điểm giá mà mã chứng khoán sẽ giao động, tuy nhiên, nó lại có ích cho các nhà đầu tư *Chứng khoán phái sinh*.

Tại thị trường Việt Nam, các sản phẩm chứng khoán phái sinh cho phép nhà đầu tư đặt cược vào sự "tăng" hoặc "giảm" của tài sản cơ sở trong tương lai. Nếu sự thay đổi đó diễn ra đúng như dự đoán, nhà đầu tư sẽ có lời. Do đó, tôi đề xuất bốn chiến lược cho các nhà đầu tư như sau.

- Chiến lược 1: Nếu dự đoán giá chứng khoán ngày hôm sau cao hơn giá hôm nay, ta sẽ mua và giữ mã chứng khoán đó.
- Chiến lược 2: Nếu dự đoán giá chứng khoán ngày hôm sau thấp hơn giá hôm nay, ta sẽ thanh lý lệnh chứng khoán.
- Chiến lược 3: Nếu hôm nay dự đoán giá ngày mai cao hơn dự đoán giá hôm nay (dự đoán này ta đã đưa ra vào ngày hôm qua), thì ta mua và giữ mã chứng khoán.
- Chiến lược 4: Nếu hôm nay dự đoán giá ngày mai thấp hơn dự đoán giá hôm nay (dự đoán này ta đã đưa ra vào ngày hôm qua), ta sẽ thanh lý lệnh chứng khoán.

Trong quá trình nghiên cứu thực hiện Đồ án, tôi đã học hỏi và nâng cao được nhiều kiến thức bổ ích, quan trọng. Tôi hy vọng những kiến thức được trình bày trong báo cáo Đồ án 2 này sẽ có ích cho những ai muốn tìm hiểu, nghiên cứu về mô hình học máy LSTM kết hợp tiền xử lý dữ liệu bằng PCA, và ứng dụng trong bài toán dự đoán giá chứng khoán.

Do còn nhiều hạn chế về kiến thức nên nội dung trong Đồ án khó tránh khỏi những thiếu sót. Vì vậy tôi rất mong nhận được những nhận xét và ý kiến đóng góp của thầy cô để Đồ án được hoàn thiện hơn.

Tài liệu tham khảo

[1] M. Wen, P. Li, L. Zhang, and Y. Chen (2019), *Stock market trend prediction using high-order information of time series*, IEEE Access, vol. 7, pp. 28299–28308.

[2] M. Z. Asghar, F. Rahman, F. M. Kundi, and S. Ahmad (2019), *Development of stock market trend prediction system using multiple regression*, Computational and Mathematical Organization Theory, vol. 25, no. 3, pp. 271–301.

[3] H. Ni, S. Wang, and P. Cheng (2021), *A hybrid approach for stock trend prediction based on tweets embedding and historical prices*, World Wide Web, vol. 24, no. 3, pp. 849–868.

[4] A. Moghar and M. Hamiche (2020), *Stock market prediction using LSTM recurrent neural network*, Procedia Computer Science, vol. 170, pp. 1168–1173.

[5] S. Paul and S. Vishnoi (2018), *Real-time stock trend prediction via sentiment analysis of news article*, Computer Engineering and Intelligent Systems, vol. 9, no. 7, pp. 21–28.

[6] V. Kranthi Sai Reddy (2018), *Stock Market Prediction Using Machine Learning*, International Research Journal of Engineering and Technology, vol. 5, pp. 1032–1035.

[7] Jingyi Shen and M. Omair Shafq (2020), *Short-term stock market price trend prediction using a comprehensive deep learning system*, Journal of Big Data.

[8] Libin Yang (2015), *An Application of Principal Component Analysis to Stock Portfolio Management*, University of Canterbury; Department of economics and finance.

[9] Robert W. Colby (2002), *The Encyclopedia Of Technical Market Indicators*, McGraw-Hill Education.

[10] Jiawei Han, Micheline Kamber, Jian Pei (2012), *Data Mining (Third Edition)*, pp. 83-124.