|  |  |
| --- | --- |
| logo (CMYK)-01 | **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM** |

**BÁO CÁO ĐỒ ÁN TRÍ TUỆ NHÂN TẠO**

**CHỦ ĐỀ:**

**Nghiên cứu hệ thống dự đoán giá chứng khoán trên môi trường song song/phân tán**

Ngành: **Công Nghệ Thông tin**

Giảng viên hướng dẫn: **Huỳnh Quốc Bảo**

Sinh viên thực hiện: MSSV: Lớp:

Phan Hoàng Anh 2080601184 20DTHE4

Nguyễn Hiền Thiện 2080601397 20DTHE4

TP. Hồ Chí Minh, năm 2022

MỤC LỤC

[DANH MỤC VIẾT TẮT 3](#_Toc123132936)

[DANH MỤC BẢNG 4](#_Toc123132937)

[DANH MỤC HÌNH 4](#_Toc123132938)

[BẢNG PHÂN CÔNG 5](#_Toc123132939)

[LỜI CẢM ƠN 5](#_Toc123132940)

[Chương 1. TỔNG QUAN 6](#_Toc123132941)

[**1.1.** **Tính cấp thiết của đề tài** 6](#_Toc123132942)

[**1.2.** **Mục Tiêu Nghiên Cứu** 7](#_Toc123132943)

[**1.3.** **Phạm vi Giới Hạn** 8](#_Toc123132944)

[**1.4.** **Cấu Trúc Đồ Án** 8](#_Toc123132945)

[**Chương 1: Tổng quan** 8](#_Toc123132946)

[**Chương 2: Cơ sở lý thuyết** 8](#_Toc123132947)

[**Chương 3: kết quả thực nghiệm** 8](#_Toc123132948)

[**Chương 4: kết luận và kiến nghị** 8](#_Toc123132949)

[Chương 2. CƠ SỞ LÝ THUYẾT 9](#_Toc123132950)

[**2.1.** **Giới thiệu một vài phương pháp nổi tiếng** 9](#_Toc123132951)

[**2.1.1.** **Giới thiệu về mạng CNN** 10](#_Toc123132952)

[**2.1.2.** **Kiến trúc cơ bản của CNN** 10](#_Toc123132953)

[**2.1.3.** **Giới thiệu về mạng RNN** 13](#_Toc123132954)

[**2.2.** **giải pháp đề xuất** 15](#_Toc123132955)

[**2.2.1.** **Giải pháp đề xuất** 15](#_Toc123132956)

[**2.3.** **Kỹ thuật chi tiết Mô hình đề xuất** 19](#_Toc123132957)

[**2.3.1.** **Áp dụng feature extension** 21](#_Toc123132958)

[**2.3.2.** **Áp dụng recursive feature elimination** 21](#_Toc123132959)

[**2.3.3.** **Áp dụng phân tích thành phần chính (PCA)** 21](#_Toc123132960)

[**2.3.4.** **Lắp mô hình long short-term memory (LSTM)** 22](#_Toc123132961)

[**2.4.** **Phương pháp đánh giá** 22](#_Toc123132962)

[**2.4.1.** **MAPE** 22](#_Toc123132963)

[**2.4.2.** **RMSE** 23](#_Toc123132964)

[Chương 3: KẾT QUẢ THỰC NGHIỆM 24](#_Toc123132965)

[**3.1.** **Tình hình thực nghiệm** 24](#_Toc123132966)

[**3.2.** **Kết quả áp dụng thuật toán RFE.** 25](#_Toc123132967)

[**3.2.1.** **Kết quả** 25](#_Toc123132968)

[**3.2.2.** **Tiện ích mở rộng tính năng RFE.** 28](#_Toc123132969)

[**3.2.3.** **Giảm tính năng bằng cách sử dụng principal component analysis(PCA)** 29](#_Toc123132970)

[**3.3.** **So sánh với các công trình đã công bố** 33](#_Toc123132971)

[**3.4.** **Phân tích độ phức tạp của giải pháp đề xuất** 37](#_Toc123132972)

[Chương 4: KẾT LUẬN VÀ KIẾN NGHỊ 39](#_Toc123132973)

[**4.1.** **Kết luận** 39](#_Toc123132974)

[**4.2.** **Đề xuất mới** 39](#_Toc123132975)

[**4.3.** **Tài liệu tham khảo với trích dẫn** 39](#_Toc123132976)

# DANH MỤC VIẾT TẮT

|  |  |
| --- | --- |
| **KÝ HIỆU** | **NGHĨA** |
| TKGD | Tài khoản giao dịch |
| SVM | Support Vector Machine |
| LSTM | Long-Short term memory |
| CNN | Convolutional Neural Network |
| RNN | Recurrent Neural Network |
| GA-ANN | Genetic Algorithm-Artificial Neural Network |
| HMM | Hidden Markov Model |
| WNN | Wavelet Neural Network |
| MLP | multi layer perceptron |
| SOFM | self-organizing feature map |
| SVR | support vector regression |
| BN-NMAS | bat-neural network multi-agent system |
| ARIMA | Autoregressive Integrated Moving Average |
| RFE | recursive feature elimination |
| PCA | principal component analysis |
| FE | feature extension |
| DNN | Deep neural network |
| DTW | Dynamic Time Warping |

# DANH MỤC BẢNG

[Bảng 2. 1. tổng quan về bộ dữ liệu với các danh mục và tập hợp con khác nhau của các trường 17](#_Toc122119694)

[bảng 3. 1. Các tính năng hiệu quả tương ứng với thời hạn 26](#_Toc122119699)

[bảng 3. 2. Mối quan hệ giữa số lượng các thành phần chính và hiệu quả đào tạo 29](#_Toc122119700)

[bảng 3. 3. Số lượng các tính năng được chọn ảnh hưởng như thế nào đến độ chính xác của dự đoán 31](#_Toc122119701)

[bảng 3. 4. Phân tích độ chính xác và hiệu quả của các quy trình tiền xử lý tính năng 32](#_Toc122119702)

[bảng 3. 5. So sánh giải pháp đề xuất với các công trình liên quan 36](#_Toc122119703)

# DANH MỤC HÌNH

[Hình 2. 1*.* Mô hình tổng quát bài toán dự đoán xu hướng giá chứng khoán 9](#_Toc123135393)

[Hình 2. 2: Mô hình CNN cơ bản 11](#_Toc123135394)

[Hình 2. 3 Biểu diễn hình ảnh đầu vào và bộ lọc dưới dạng ma trận 12](#_Toc123135395)

[Hình 2. 4 Minh hoạt Max pooling với bộ lọc 2x2 và stride là 2 13](#_Toc123135396)

[Hình 2. 5 Mô hình RNN 14](#_Toc123135397)

[Hình 2. 6 Mô tả bên trong mô hình RNN 14](#_Toc123135398)

[Hình 2. 7. mô tả kiến ​​trúc cấp cao của LSTM 18](#_Toc123135399)

[Hình 2. 8. thiết kế kỹ thuật chi tiết từ xử lý dữ liệu đến dự đoán, bao gồm cả khám phá dữ liệu 19](#_Toc123135400)

[Hình 3. 1. Biểu diễn RFE 24](#_Toc122117552)

[Hình 3. 2. Ma trận nhầm lẫn xác nhận tính hiệu quả của tiện ích mở rộng tính năng 28](#_Toc122117553)

[Hình 3. 3. Mối quan hệ giữa số tính năng và thời gian đào tạo 30](#_Toc122117554)

[Hình 3. 4. Số lượng các thành phần chính ảnh hưởng đến kết quả đánh giá như thế nào 31](#_Toc122117555)

[Hình 3. 5. So sánh dự đoán mô hình—ma trận nhầm lẫn 33](#_Toc122117556)

[Hình 3. 6. Đường cong học tập của giải pháp đề xuất 33](#_Toc122117557)

[Hình 3. 7. So sánh chính xác dự đoán mô hình được đề xuất—ma trận nhầm lẫn 34](#_Toc122117558)

# BẢNG PHÂN CÔNG

|  |  |  |
| --- | --- | --- |
| Tên nhóm: AI STOCK |  |  |
| Họ tên | MSSV | Đóng góp |
| Phan Hoàng Anh(nhóm trưởng) | 2080601184 | Làm báo cáo + demo+ thuyết trình |
| Nguyễn Hiền Thiện | 2080601397 | Làm slide + thuyết trình |

# LỜI CẢM ƠN

Lời đầu tiên, nhóm em xin gửi lời cảm ơn chân thành đến Trường ĐH Công Nghệ TP.HCM đã đưa môn học trí tuệ nhân tạo vào chương trình giảng dạy. Đặc biệt, em xin gửi lời cảm ơn chân thành đến giảng viên bộ môn – Thầy Huỳnh Quốc bảo đã dạy dỗ và tâm huyết truyền đạt những kiến thức quý giá cho nhóm em trong suốt thời gian học tập vừa qua. Trong thời gian tham gia lớp học trí tuệ nhân tạo của Thầy, nhóm em đã trau dồi cho bản thân nhiều kiến thức bổ ích, tinh thần học tập nghiêm túc và hiệu quả. Đây chắc chắn sẽ là những kiến thức có giá trị sâu sắc, là hành trang để nhóm em vững chắc sau này.

Bộ môn trí tuệ nhân là môn học thú vị, bổ ích và có tính thực tế cao. Đảm bảo cung cấp đầy đủ kiến thức, kỹ năng, giúp sinh viên có thể ứng dụng vào thực tế. Tuy nhiên, do khả năng tiếp thu thực tế còn nhiều hạn hẹp, kiến thức chưa sâu rộng. Mặc dù bản thân đã cố gắng hết sức nhưng chắc chắn đồ án khó tránh khỏi những thiếu sót, kính mong Thầy xem xét và góp ý để đồ án của nhóm em được hoàn thiện và tốt hơn.

Nhóm chúng em xin chân thành cảm ơn!.

# Chương 1. TỔNG QUAN

* 1. **Tính cấp thiết của đề tài**

Thị trường chứng khoán vẫn luôn chiếm vị thế nhất định trong thị trường đầu tư, không chỉ đem lại nguồn thu nhập thụ động mà còn giúp cho nền kinh tế phát triển một cách mạnh mẽ. Do đó đã thu hút được đông đảo các nhà đầu tư từ các cá nhân nhỏ lẻ cho đến các công ty lớn, bởi đây là thị trường tiềm năng đem đến nguồn tiền tích cực. Cùng với sự phát triển của công nghệ, việc truy xuất các dữ liệu giao dịch, các thông tin trực quan của thị trường đã giúp các nhà đầu tư có cơ sở và quyết định đúng đắn để bắt đầu công cuộc đầu tư của mình, chính vì lẽ đó mà sự phát triển của các sàn chứng khoán trở nên mạnh mẽ và năng động hơn bao giờ hết.

Theo số liệu được cung cấp bởi Statista – Công ty chuyên về thị trường và dữ liệu người dùng và diễn đàn Asktraders , trên thế giới có xấp xỉ 9.6 triệu nhà đầu tư trực tuyến, chiếm vỏn vẹn gần 0.12% dân số trên thế giới. Thị trường chứng khoán ở Mỹ chiếm gần 52% tổng thị trường trên thế giới, tiếp đó là Nhật Bản và Trung Quốc lần lượt là 7.4% và 5.4%.

Ở thị trường Việt Nam, theo thống kê số lượng tài khoản nhà đầu tư chứng khoán tại Việt Nam ngày 30/09/2021 :

• Số lượng tài khoản giao dịch (TKGD) trong nước là 3.693.553 tài khoản, trong đó:

– Số lượng TKGD của nhà đầu tư cá nhân: 3.681.168 tài khoản

– Số lượng TKGD của nhà đầu tư tổ chức: 12.385 tài khoản

Chiếm khoảng 3.8% tổng dân số trên cả nước. Ngoài ra dưới sự ảnh hưởng của dịch bệnh Covid 19, lượng tài khoản chứng khoán mới mở cũng tăng mạnh do nhu cầu khai thác tài chính thụ động được tăng cao. “Lũy kế 8 tháng đầu năm, nhà đầu tư trong nước mở mới 842.405 tài khoản chứng khoán, lớn hơn tổng số tài khoản mở mới trong 3 năm 2018-2020 cộng lại là 837.345 tài khoản”

– Theo CƠ QUAN TRUNG ƯƠNG ĐẢNG CỘNG SẢN VIỆT NAM .

Tuy nhiên, thị trường chứng khoán có tính chất ngẫu nhiên và phi tuyến tính, điều đó có nghĩa là việc dự đoán xu hướng của chứng khoán chỉ bằng những dữ liệu lịch sử giá là một bài toán đầy thách thức, đồng thời thị trường chứng khoán cũng bị ảnh hưởng bởi nhiều yếu tố như thị trường cung cầu, tình hình thế giới, tình hình chính trị xã hội, biến động kinh tế, báo cáo tài chính. Dẫn đến việc giải quyết bài toán này gần như là không thể ở những năm về trước.

Nhưng nhờ có sự bùng nổ của lĩnh vực trí tuệ nhân tạo, cụ thể là sự phát triển nhảy vọt của những kỹ thuật học sâu, các bài toán đầy thách thức ở các lĩnh vực khác nhau đã lần lượt được giải quyết. Các ứng dụng về xử lý ảnh trong lĩnh vực y khoa như dự đoán bệnh qua ảnh CT, ứng dụng xử lý ngôn ngữ tự nhiên, nhận diện giọng nói đã được phát triển và cải thiện vượt bậc. Với những đột phá về công nghệ và kỹ thuật này đã góp phần tạo nên một nền tảng vững chắc để xây dựng mô hình dự đoán và các tri thức để cải tiến.

* 1. **Mục Tiêu Nghiên Cứu**

Cùng với sự phát triển của nền kinh tế thị trường, nhu cầu tăng thêm thu nhập của con người ngày càng cao. Đối với tầng lớp lao động, đi làm thuê thì việc vươn lên tầng lớp doanh nhân, giàu có gần như là rất khó chỉ với nguồn thu nhập bị động ít ỏi. Trong khi đó thị trường chứng khoán lại vô cùng năng động và tiềm năng lại chỉ có những chuyên gia tài chính, những cá nhân có vốn hoặc những doanh nghiệp dám tham gia vào thị trường đầu tư. Với lượng kiến thức chuyên môn hạn hẹp cùng với số vốn ít ỏi, những tầng lớp phổ thông ít có cơ hội chen chân vào thị trường chứng khoán dồi dào và màu mỡ đó để cải thiện cuộc sống. Để giảm thiểu rủi ro thua lỗ và tối ưu hóa lợi nhuận khi đầu tư, giúp cho tất cả các nhà đầu tư thuộc mọi tầng lớp có thể tham gia vào thị trường, các mô hình học sâu về bài toán dự đoán xu hướng giá chứng khoán đã được nghiên cứu và phát triển.

Tuy nhiên để có được một mô hình có độ tin cậy cao đòi hỏi phải có một phương pháp tối ưu kết hợp với các phương pháp phân tích cơ bản (Fundamental Analysis) và phương pháp phân tích kỹ thuật (Technical Analysis) cũng như dữ liệu giá tương ứng. Đến nay, đã có nhiều hướng tiếp cận để giải quyết bài toán như các mô hình máy học Support Vector Machine, Decision Tree, Random Forest. Các mô hình học sâu như: Mạng Nơron tích chập (Convolutional Neural Network), mạng Nơron hồi quy (Recurrent Neural Network). Và một trong những mô hình tiên tiến được cho là mô hình thích hợp và có độ hiệu quả cao với bài toán có dữ liệu tuần tự (sequence hay time-series data) là mạng LSTM (Long-Short term memory - bộ nhớ dài-ngắn hạn) . Như đã phân tích ở trên, bài toán xây dựng mô hình học sâu dự đoán xu hướng giá chứng khoán có ý nghĩa thực tiễn lớn, không những giúp mang lại thu nhập thụ động cho nhà đầu tư mà còn giúp thúc đẩy nền kinh tế phát triển bằng việc thúc đẩy dòng tiền chạy liên tục trên thị trường.

* 1. **Phạm vi Giới Hạn**
* Về dự đoán chuỗi thời gian tài chính được biết đến là một nhiệm vụ cực kỳ khó khăn do hình thức hiệu quả thị trường bán mạnh được chấp nhận rộng rãi và mức độ nhiễu cao
* Về hướng tiếp cận của bài toán: Có nhiều hướng tiếp cận trong lĩnh vực trí tuệ nhân tạo như Học máy (Machine Learning), Học sâu (Deep Learning) và Học tăng cường (Reinforcement Learning) nhưng đề tài mà nhóm nghiên cứu thực hiện sẽ tập trung chủ yếu vào các mô hình học sâu kết hợp với phương pháp phân tích kỹ thuật
  1. **Cấu Trúc Đồ Án**

### ***Chương 1: Tổng quan***

Trong chương này đồ án sẽ giới thiệu về đề tài. Trình bày về mục tiêu nghiên cứu và phạm vi giới hạn của đề tài này

### ***Chương 2: Cơ sở lý thuyết***

Trong chương này đồ án sẽ trình bày một vài phương pháp nổi tiếng, các công trình nghiên cứu trước đây , các giải pháp mới , mô tả cấu trúc dữ liệu và giải pháp đề xuất của tác giả.

### ***Chương 3: kết quả thực nghiệm***

Trong chương này đồ án mô tả kết quả nghiên cứu của tác giả và so sánh mô hình của tác giả với các mô hình khác

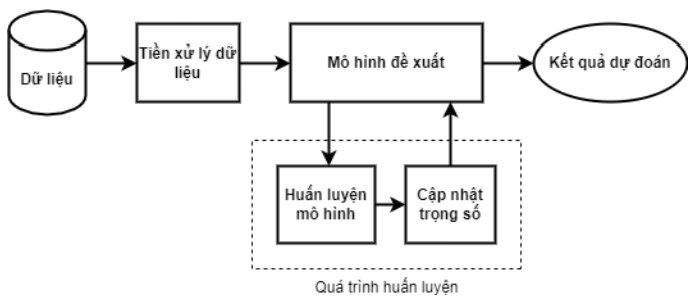
### ***Chương 4: kết luận và kiến nghị***

Trong chương đồ án sẽ nêu kết luận của tác giả và đề xuất mới của nhóm chúng em

# Chương 2. CƠ SỞ LÝ THUYẾT

* 1. **Giới thiệu một vài phương pháp nổi tiếng**

Từ các phương pháp máy học truyền thống đến các mô hình học sâu, có rất nhiều kĩ thuật khác nhau dùng để dự đoán chuỗi thời gian tài chính nói chung



Hình 2. 1*.* Mô hình tổng quát bài toán dự đoán xu hướng giá chứng khoán

Đầu tiên có thể kể đến RNN - một mạng nơ ron được thiết kế đặc biệt với khả năng xử lý được dữ liệu thời gian và trích xuất thông tin cũng như dự đoán dựa trên chuỗi thông tin trước đó. Vì vậy, các mô hình RNN và biến thể của nó dường như phù hợp để dự đoán cho chuỗi thời gian và cụ thể là chuỗi thời gian về tài chính. CNN là một trong những mô hình học sâu khác có thể áp dụng cho bài toán dự đoán giá chứng khoán nhờ vào khả năng trích xuất thông tin hiệu quả qua các bộ lọc. Theo một số kết quả thực nghiệm, CNN có một vai trò đáng kể trong việc xử lý dữ liệu đầu vào và trích xuất đặc trưng. Ví dụ, theo có thể áp dụng cho nhiều dữ liệu từ các nguồn khác nhau, bao gồm nhiều thị trường khác nhau, và trích xuất đặc trưng để dự đoán cho các thị trường này. Các kết quả đánh giá cho thấy so với thuật toán cơ sở (baseline algorithm) tốt nhất, hiệu suất dự đoán khi sử dụng CNN cho cải thiện đáng kể. Ngoài ra các nhà nghiên cứu cũng đã áp dụng mô hình Học tăng cường - một phương pháp học sâu, đưa ra hành động dựa trên tình huống hiện tại để tối đa hóa lợi nhuận. Học tăng cường học cách thực hiện giao dịch (có thể là mua, bán hoặc giữ cổ phiếu) theo tình hình của thị trường hiện tại. Mô hình xem thông tin ngữ cảnh (giá cả, tin tức, dư luận, phí giao dịch, hành động giao 21 dịch, lợi nhuận, thua lỗ) như một môi trường để học tăng cường. Coi lãi hoặc lỗ có thể coi như phần thưởng, các giao dịch (bán, mua, giữ) được coi như các hành động và mô hình cần tối ưu các hành động này để đưa ra lợi nhuận tối ưu.

* + 1. *Giới thiệu về mạng CNN*

Mô hình mạng nơron truyền thẳng ra đời đã được áp dụng vào các bài toán nhận dạng. Tuy nhiên, đối với dữ liệu hình ảnh, mạng nơron truyền thẳng cho kết quả không thực sự tốt bởi vì sự liên kết quá đầy đủ tạo nên những hạn chế cho mô hình. Nếu coi một điểm ảnh (pixel) là một đặc trưng thì một bức ảnh hệ màu RGB có kích thước (64 x 64) sẽ có 64 x 64 x 3 = 12288 đặc trưng. Nếu tăng kích thước ảnh lên 1000 x 1000 thì ta sẽ được một vectơ đặc trưng với 1000 x 1000 x 3 = 3000000 chiều. Rõ ràng nếu ta sử dụng thêm các tầng kết nối đầy đủ thì số lượng trọng số sẽ tăng lên một cách bùng nổ. Ví dụ với lớp kế tiếp sử dụng 100 units thì ma trận trọng số sẽ có kích thước 3000000 x 100 tương đương với 3.108 trọng số cần huấn luyện, nếu thêm nhiều tầng kết nối đầy đủ thì kích thước ma trận sẽ càng ngày càng lớn. Điều này yêu cầu khối lượng tính toán cực lớn và lãng phí, thường sẽ dẫn đến hiện tượng quá khớp (overfitting) do không đủ dữ liệu huấn luyện. Việc liên kết đầy đủ của các điểm ảnh trong mạng là dư thừa và lãng phí vì sự phụ thuộc lẫn nhau giữa các điểm ảnh xa nhau là không nhiều, mà chủ yếu là sự phụ thuộc giữa các điểm ảnh lân cận nó, dựa trên ý tưởng này, mô hình học sâu CNN ra đời với cấu trúc khác hẳn mạng nơron truyền thẳng. Đầu tiên, các lớp của CNN được tổ chức theo 3 chiều: Chiều rộng, chiều cao và chiều sâu và thay vì mỗi lớp được kết nối hoàn toàn với tất cả các nơron trong tầng trước đó như mạng truyền thẳng thì các nơron trong một tầng của CNN không kết nối với tất cả các nơron trong tầng tiếp theo mà chỉ kết nối đến một phần nhỏ của nó.

* + 1. *Kiến trúc cơ bản của CNN*

Về cơ bản mô hình CNN bao gồm các lớp: Lớp tích chập (Convolutional layer), lớp kích hoạt (Activation layer), lớp tổng hợp (Pooling layer) và cuối cùng là lớp kết nối đầy đủ (Fully-connected layer) Hình ảnh được đưa về dạng ma trận điểm ảnh để bắt đầu đưa qua các lớp

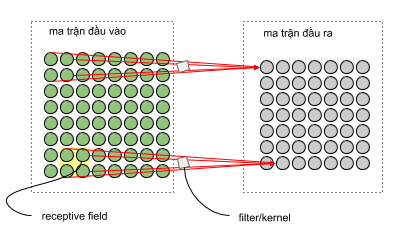


Hình 2. 2**:** Mô hình CNN cơ bản

trong mô hình CNN. Thông qua lớp tích chập, các đặc trưng của dữ liệu được rút trích và thông qua tầng kích hoạt để cho ra các biểu đồ đặc trưng (feature map). Sau đó tiếp tục đi qua tầng tổng hợp để thu nhỏ số chiều dữ liệu nhưng vẫn giữ lại được lượng thông tin cần thiết, điều này giúp cho mô hình có thể học nhanh hơn nhờ giảm được chi phí tính toán. Cuối cùng sau khi đã có được biểu đồ đặc trưng thì ta làm phẳng (flatten) nó để vào tầng kết nối đầy đủ từ đó đưa ra kết quả dự đoán cuối cùng

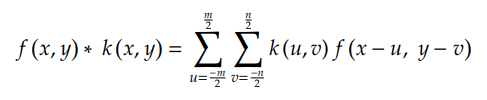
* + - 1. Lớp tích chập (Convolutional Layer)

Lớp tích chập là lớp cốt lõi trong CNN. Chức năng của lớp tích chập là trích xuất đặc trưng của các dữ liệu đầu vào thông qua các bộ lọc. Bộ lọc (filter/kernel) là tập hợp các tham số học được trong quá trình huấn luyện mô hình nhằm trích xuất đặc trưng cho dữ liệu. Trong lĩnh vực xử lý ảnh, tích chập được sử dụng khá rộng rãi và hết sức quan trọng. Ví dụ: Đối với phép lọc ảnh, phép tích chập giữa ma trận lọc và ảnh, cho ra kết quả ảnh đã được xoá nhiễu.



Hình 2. 3 Biểu diễn hình ảnh đầu vào và bộ lọc dưới dạng ma trận

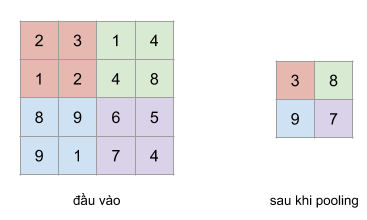
Công thức tích chập giữa hàm ảnh f x, y và bộ lọc k x, y (kích thước m x n) như sau:



Trong phép tính tích chập, ta bắt đầu với cửa sổ tích chập ở góc trên bên trái của mảng dữ liệu đầu vào, di chuyển cửa sổ tích chập theo chiều từ trái sang phải, từ trên xuống dưới. Với mỗi vị trí mà bộ lọc nằm hoàn toàn trong mảng đầu vào, ta thực hiện phép tích chập bằng cách nhân từng phần tử của cửa sổ tích chập và mảng con ở vị trí tương ứng, sau đó cộng tất cả các giá trị lại với nhau để được một giá trị duy nhất. Mỗi bộ lọc khác nhau sẽ cho ra các đặc trưng khác nhau nên mỗi lớp tích chập sẽ dùng nhiều bộ lọc để tìm ra nhiều đặc trưng của dữ liệu. Giả sử có toàn bộ bộ lọc trong lớp tích chập (ví dụ: k bộ lọc), mỗi bộ lọc cho ra kết quả là một ma trận đặc trưng hai chiều riêng biệt (ví dụ: Ma trận kích thước n x n), ta xếp chồng các ma trận này tạo ra ma trận mới có độ sâu là k làm ma trận đầu ra (kích thức n x n x k)

* + - 1. Lớp Pooling (Lớp tổng hợp)

Thông thường, giữa các lớp tích chập trong một mô hình CNN sẽ chèn giữa là một lớp pooling. Chức năng của nó là giảm dần kích thước dữ liệu, giảm số lượng tham số và tính toán trong mạng do đó cũng hạn chế việc overfitting (trang bị quá mức). Lớp pooling chỉ làm giảm kích thước chiều rộng và chiều cao, không làm thay đổi kích thước chiều sâu dữ liệu đầu vào. Lớp pooling có nhiều loại khác nhau: Max pooling - lấy phần tử lớn nhất trong ma trận con, average pooling - lấy giá trị trung bình của các phần tử, sum pooling - lấy tổng giá trị các phần tử, trong đó max pooling thường được sử dụng rộng rãi nhất.



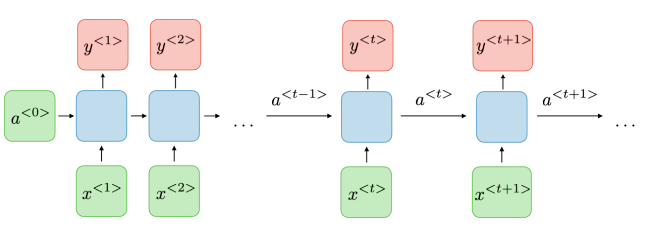
Hình 2. 4 Minh hoạt Max pooling với bộ lọc 2x2 và stride là 2

* + - 1. Lớp Fully connected

Trong mạng nơron truyền thẳng truyền thống, lớp ẩn (hidden layer) giống với lớp fully connected của mạng CNN. Lớp fully connected nằm ở phần cuối của các lớp ẩn trong mạng CNN. Lớp tích chập và pooling có thể thực hiện việc trích xuất đặc trưng từ dữ liệu đầu vào, trong khi đó lớp fully connected kết hợp các tính năng được trích xuất để có được đầu ra, đóng vai trò ánh xạ các đặc trưng đã được học vào không gian nhãn.

* + 1. *Giới thiệu về mạng RNN*

Trong những năm gần đây, công nghệ không ngừng phát triển và tiến bộ, các tác vụ xử lý chuỗi dữ liệu thời gian (Time series data/ Sequence data) cũng đạt được nhiều thành tựu vượt bậc. Các hệ thống dự báo chuỗi thời gian như dự báo thời tiết, dịch máy đã được các tập đoàn lớn như Facebook, Google đưa vào sản phẩm của mình để phục vụ người dùng. Mô hình mạng nơron truyền thẳng ra đời đã được áp dụng vào nhiều bài toán khác nhau. Tuy nhiên, đối với những dữ liệu thời gian hay dữ liệu tuần tự thì mô hình nơron truyền thẳng không đáp ứng được yêu cầu của bài toán, do đó dẫn đến sự ra đời của mạng hồi quy (Recurrent Neural Network - RNN) - mô hình mạng nơron có thể xử lý được được các loại dữ liệu tuần tự này.

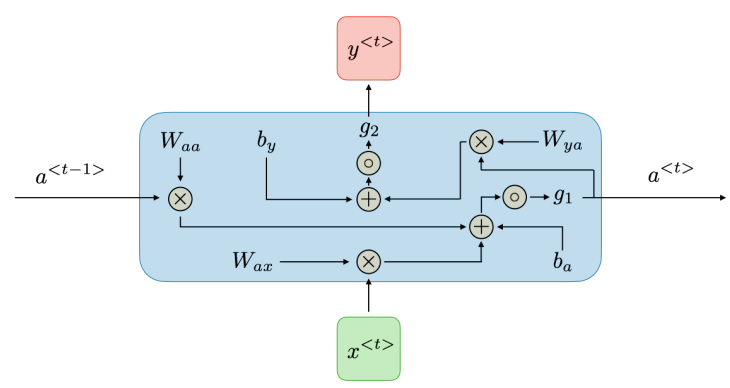
****

Hình 2. 5 Mô hình RNN

Dữ liệu được đưa vào mô hình một cách tuần tự, ở mỗi bước thời gian t ta sẽ đưa dữ liệu đầu vào xt tương ứng. Và lượng thông tin từ xt sẽ được truyền tiếp tục sang các bước tiếp theo, điều này cho phép mạng RNN kết nối được những thông tin từ dữ liệu trước đó và dự đoán thông tin ở bước tiếp theo.

* + - 1. Kiến trúc mô hình

Ở mỗi bước xử lý, ta gọi đó là một trạng thái ẩn (hidden state), sau khi tiếp nhận thông tin từ đầu vào xt và thông tin trước đó a ta được kết quả đầu ra y và thông tin cho bước tiếp theo a với:

****

Hình 2. 6 Mô tả bên trong mô hình RNN

* + - 1. Các vấn đề của mạng RNN

Điểm nổi bật của RNN so với các mạng nơron khác chính là ý tưởng kết nối các thông tin trước đó để đưa ra dự đoán kế tiếp. Nhưng nếu dữ liệu quá dài hoặc lượng thông tin quá lớn thì sẽ mạng RNN sẽ không thể “nhớ” được những thông tin quá xa trước đó, hiện tượng này gọi là Vanishing gradient (đạo hàm bị triệt tiêu). Hiện tượng này xảy ra là do hàm kích hoạt (tanh hay sigmoid) của mô hình sẽ cho kết quả output nằm trong khoảng giá trị [−1, 1] (với sigmoid là [0, 1]) nên giá trị đạo hàm của nó cũng sẽ nằm trong khoảng [0, 1] (với sigmoid là [0, 0.25]). Ở trên, chúng ta đã dùng quy tắc chuỗi để tính đạo hàm. Có một vấn đề ở đây là, hàm tanh lẫn sigmoid đều có đạo hàm bằng 0 tại 2 đầu. Mà khi đạo hàm bằng 47 0 thì nút mạng tương ứng tại đó sẽ bị bão hòa. Lúc đó các nút phía trước cũng sẽ bị bão hoà theo. Nên với các giá trị nhỏ trong ma trận, khi ta thực hiện phép nhân ma trận sẽ đạo hàm tương ứng sẽ xảy ra Vanishing gradient, có nghĩa đạo hàm bị triệt tiêu chỉ sau vài bước nhân. Như vậy, các bước ở xa sẽ không còn tác dụng với nút hiện tại nữa, làm cho RNN không thể học được các phụ thuộc xa. Ngoài ra với cách tiếp cận trên, ta còn thấy một hiện tượng khác là Exploding gradient (bùng nổ đạo hàm), tùy thuộc vào hàm kích hoạt và tham số của mạng, vấn đề này xảy ra khi các giá trị của ma trận là lớn (lớn hơn 1) dẫn tới các giá trị trọng số cực lớn trong quá trình huấn luyện. Tuy nhiên người ta thường nhắc tới Vanishing gradient nhiều hơn bởi vì hai lý do sau:

• Exploding gradient có thể theo dõi được vì khi đạo hàm bị bùng nổ thì ta sẽ thu được kết quả là một giá trị phi số NaN làm cho chương trình của ta bị dừng hoạt động.

• Exploding gradient có thể ngăn chặn được khi ta đặt một ngưỡng giá trị trên bằng kỹ thuật Gradient Clipping, còn Vanishing Gradient thì không.

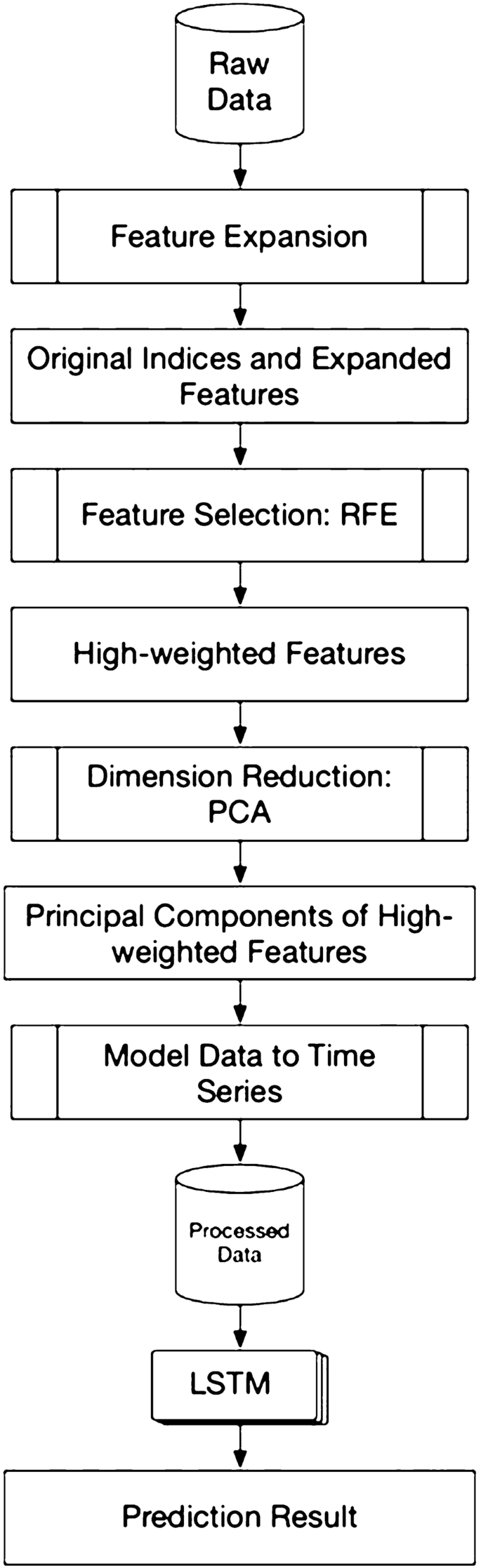
Tuy nhiên vẫn có những phương pháp để xử lý hiện tượng Vanishing Gradient, một trong những cách phổ biến đó là sử dụng hàm kích hoạt ReLU (hoặc các biến thể của nó) vì đạo hàm của ReLu hoặc là 0 hoặc là 1, nên ta có thể kiểm soát phần nào vấn đề mất mát đạo hàm. Ta có thể thấy RNN không có cách nào lọc được các thông tin không cần thiết vì vậy các biến thể khác của RNN ra đời với kiến trúc giúp cho mạng có thể nhớ dài hạn hơn, phổ biến nhất là mạng LSTM (Long – Short Term Memory) và mạng GRU (Gated Recurrent Unit).

* 1. **giải pháp đề xuất**

Trong phần này, chúng em sẽ nói về các phương pháp của tác giả đề xuất và thiết kế của giải pháp được đề xuất. Tiếp đó chúng em giới thiệu thiết kế kiến ​​trúc cũng như các chi tiết về thuật toán và triển khai.

* + 1. *Giải pháp đề xuất*

Kiến trúc cấp cao của giải pháp đề xuất của tác giả có thể được tách thành ba phần. Đầu tiên là phần lựa chọn tính năng, để đảm bảo tính năng được chọn có hiệu quả cao. Thứ hai, tác giả xem xét dữ liệu và thực hiện giảm kích thước. Và phần cuối cùng, đóng góp chính trong công việc của chúng tôi là xây dựng một mô hình dự đoán về các cổ phiếu mục tiêu. Hình  **2.3** mô tả kiến ​​trúc cấp cao của giải pháp được đề xuất.



Hình 2. 7. mô tả kiến ​​trúc cấp cao của LSTM

Có nhiều cách để phân loại các loại cổ phiếu khác nhau. Một số nhà đầu tư thích đầu tư dài hạn, trong khi những người khác tỏ ra quan tâm hơn đến các khoản đầu tư ngắn hạn. Người ta thường thấy các báo cáo liên quan đến cổ phiếu cho thấy hiệu suất trung bình, trong khi giá cổ phiếu đang tăng mạnh; Đây là một trong những hiện tượng cho thấy việc dự đoán giá cổ phiếu không có quy luật cố định, do đó việc tìm ra các đặc trưng hiệu quả trước khi huấn luyện mô hình trên dữ liệu là cần thiết.

Trong nghiên cứu này, nhóm nghiên cứu tập trung vào dự đoán xu hướng giá ngắn hạn. Hiện tại, tác giả chỉ có dữ liệu thô không có nhãn. Vì vậy, bước đầu tiên là gắn nhãn dữ liệu. tác giả đánh dấu xu hướng giá bằng cách so sánh giá đóng cửa hiện tại với giá đóng cửa của n ngày giao dịch trước, phạm vi của n là từ 1 đến 10 do nghiên cứu của chúng tôi tập trung vào ngắn hạn. Nếu xu hướng giá tăng lên, chúng tôi đánh dấu nó là 1 hoặc đánh dấu là 0 trong trường hợp ngược lại. Để cụ thể hơn, chúng tôi sử dụng các chỉ số từ các chỉ số của ngày thứ *n* -  ***thứ****1* để dự đoán xu hướng giá của ngày ***thứ****n* .

Theo các công trình trước đây, một số nhà nghiên cứu áp dụng cả kiến ​​thức lĩnh vực tài chính và phương pháp kỹ thuật trên dữ liệu chứng khoán đã sử dụng các quy tắc để lọc các cổ phiếu chất lượng cao. Nhóm nghiên cứu đã tham khảo các tác phẩm của họ và khai thác các quy tắc của họ để đóng góp cho thiết kế tiện ích mở rộng tính năng.

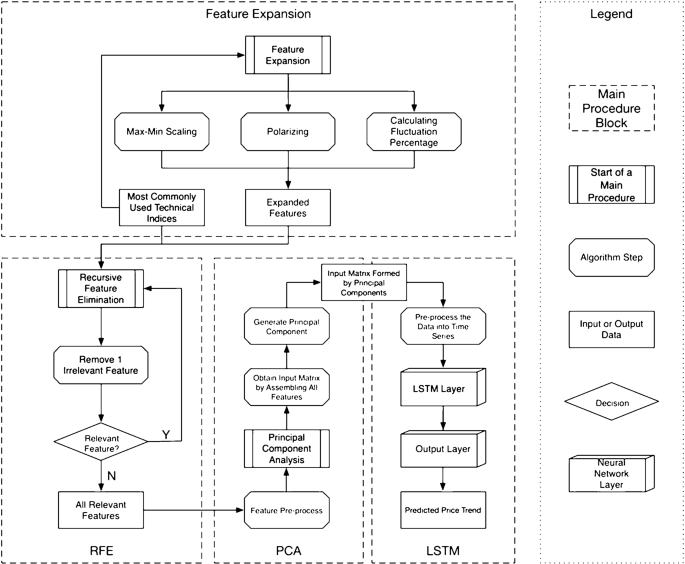
Tuy nhiên, để đảm bảo hiệu suất tốt nhất của mô hình dự đoán, tác giả đã xem xét dữ liệu trước. Có một số lượng lớn các tính năng trong dữ liệu thô; nếu tác giả xem xét tất cả các tính năng, nó sẽ không chỉ làm tăng đáng kể độ phức tạp tính toán mà còn gây ra tác dụng phụ nếu chúng tôi muốn thực hiện học tập không giám sát trong nghiên cứu sâu hơn. Vì vậy, chúng tôi tận dụng việc loại bỏ tính năng đệ quy (RFE) để đảm bảo tất cả các tính năng đã chọn đều hiệu quả.

Tác giả nhận thấy hầu hết các công trình trước đây trong lĩnh vực kỹ thuật đều phân tích tất cả các cổ phiếu, trong khi trong lĩnh vực tài chính, các nhà nghiên cứu thích phân tích kịch bản đầu tư cụ thể hơn, để lấp đầy khoảng cách giữa hai lĩnh vực, quyết định của tác giả là áp dụng tiện ích mở rộng tính năng dựa trên về những phát hiện mà họ thu thập được từ lĩnh vực tài chính trước khi tác giả bắt đầu thủ tục RFE.

Vì nhóm nghiên cứu dự định mô hình hóa dữ liệu thành chuỗi thời gian, số lượng tính năng càng nhiều, quy trình đào tạo sẽ càng phức tạp. Vì vậy, sẽ tận dụng việc giảm kích thước bằng cách sử dụng PCA ngẫu nhiên khi bắt đầu kiến ​​trúc giải pháp được đề xuất.

* 1. **Kỹ thuật chi tiết Mô hình đề xuất**

Phần này cung cấp thông tin chi tiết về thiết kế kỹ thuật chi tiết của mô hình đề xuất như một giải pháp toàn diện dựa trên việc sử dụng, kết hợp và tùy chỉnh một số kỹ thuật tiền xử lý dữ liệu, kỹ thuật tính năng và kỹ thuật học sâu hiện có. Hình  **2.4** cung cấp thiết kế kỹ thuật chi tiết từ xử lý dữ liệu đến dự đoán, bao gồm cả khám phá dữ liệu. Tác giả chia nội dung theo quy trình chính và mỗi quy trình chứa các bước thuật toán. Chi tiết thuật toán được xây dựng trong phần tiếp theo. Nội dung của phần này sẽ tập trung vào việc minh họa quy trình xử lý dữ liệu.



Hình 2. 8. thiết kế kỹ thuật chi tiết từ xử lý dữ liệu đến dự đoán, bao gồm cả khám phá dữ liệu

Dựa trên đánh giá tài liệu, tác giả chọn các chỉ số kỹ thuật được sử dụng phổ biến nhất và sau đó đưa chúng vào quy trình mở rộng tính năng để có được bộ tính năng mở rộng. tác giả đã chọn các tính năng *i* hiệu quả nhất từ ​​bộ tính năng mở rộng. Sau đó, sẽ cung cấp dữ liệu với *i* tính năng đã chọn vào thuật toán PCA để giảm kích thước thành *j* tính năng. Sau khi chúng tôi có được sự kết hợp tốt nhất giữa *i* và *j* , rồi xử lý dữ liệu thành bộ tính năng đã hoàn thiện và đưa chúng vào mô hình LSTM để có được kết quả dự đoán xu hướng giá.

Tính mới của giải pháp được đề xuất của tác giả là sẽ không chỉ áp dụng phương pháp kỹ thuật trên dữ liệu thô mà còn thực hiện các phần mở rộng tính năng được sử dụng giữa các nhà đầu tư trên thị trường chứng khoán. Chi tiết về phần mở rộng tính năng được đưa ra trong tiểu mục tiếp theo. Kinh nghiệm thu được từ việc áp dụng và tối ưu hóa các giải pháp dựa trên học sâu trong đã được tính đến khi thiết kế và tùy chỉnh kỹ thuật tính năng và giải pháp học sâu trong công việc này.

* + 1. *Áp dụng feature extension*

Quy trình chính đầu tiên trong Hình  **2.4** là phần mở rộng tính năng. Trong khối này, dữ liệu đầu vào là các chỉ số kỹ thuật được sử dụng phổ biến nhất được đúc kết từ các công việc liên quan. Ba phương pháp mở rộng tính năng là chia tỷ lệ tối đa tối thiểu, phân cực và tính phần trăm dao động. Không phải tất cả các chỉ số kỹ thuật đều có thể áp dụng cho cả ba phương pháp mở rộng tính năng; quy trình này chỉ áp dụng các phương pháp mở rộng có ý nghĩa đối với các chỉ số kỹ thuật. Chúng tôi chọn các phương pháp mở rộng có ý nghĩa trong khi xem xét cách tính toán các chỉ số. Sau quy trình mở rộng tính năng, các tính năng mở rộng sẽ được kết hợp với các chỉ số kỹ thuật được sử dụng phổ biến nhất, tức là dữ liệu đầu vào với dữ liệu đầu ra và đưa vào khối RFE dưới dạng dữ liệu đầu vào trong bước tiếp theo.

* + 1. *Áp dụng recursive feature elimination*

Sau phần mở rộng tính năng ở trên, chúng tôi khám phá các tính năng *i* hiệu quả nhất bằng cách sử dụng thuật toán Loại bỏ tính năng đệ quy (RFE) . Chúng tôi ước tính tất cả các tính năng theo hai thuộc tính, hệ số và tầm quan trọng của tính năng. Chúng tôi cũng giới hạn các tính năng xóa khỏi nhóm theo từng tính năng, có nghĩa là chúng tôi sẽ xóa một tính năng ở mỗi bước và giữ lại tất cả các tính năng có liên quan. Sau đó, đầu ra của khối RFE sẽ là đầu vào của bước tiếp theo, đề cập đến PCA.

* + 1. *Áp dụng phân tích thành phần chính (PCA)*

Bước đầu tiên trước khi tận dụng PCA là xử lý trước tính năng. Bởi vì một số tính năng sau RFE là dữ liệu tỷ lệ phần trăm, trong khi những tính năng khác là số rất lớn, tức là đầu ra từ RFE ở các đơn vị khác nhau. Nó sẽ ảnh hưởng đến kết quả trích xuất thành phần chính. Do đó, trước khi đưa dữ liệu vào thuật toán PCA , cần phải xử lý trước tính năng.

Sau khi thực hiện tiền xử lý tính năng, bước tiếp theo là cung cấp dữ liệu đã xử lý với các tính năng *i* đã chọn vào thuật toán PCA để giảm tỷ lệ ma trận tính năng thành các tính năng *j* . Bước này nhằm giữ lại càng nhiều tính năng hiệu quả càng tốt và đồng thời loại bỏ độ phức tạp tính toán của việc huấn luyện mô hình. Công trình nghiên cứu này cũng đánh giá sự kết hợp tốt nhất của *i* và *j,* có độ chính xác dự đoán tương đối tốt hơn, đồng thời cắt giảm mức tiêu thụ tính toán.

* + 1. *Lắp mô hình long short-term memory (LSTM)*

PCA đã giảm kích thước của dữ liệu đầu vào, trong khi tiền xử lý dữ liệu là bắt buộc trước khi đưa dữ liệu vào lớp LSTM. Lý do thêm bước tiền xử lý dữ liệu trước mô hình LSTM là ma trận đầu vào được hình thành bởi các thành phần chính không có bước thời gian. Trong khi một trong những tham số quan trọng nhất của việc đào tạo LSTM là số bước thời gian. Do đó, chúng ta phải mô hình hóa ma trận thành các bước thời gian tương ứng cho cả tập dữ liệu huấn luyện và kiểm tra.

Sau khi thực hiện phần tiền xử lý dữ liệu, bước cuối cùng là đưa dữ liệu huấn luyện vào LSTM và đánh giá hiệu suất bằng dữ liệu thử nghiệm. Là một mạng nơ-ron biến thể của RNN, ngay cả với một lớp LSTM, cấu trúc NN vẫn là một mạng nơ-ron sâu vì nó có thể xử lý dữ liệu tuần tự và ghi nhớ các trạng thái ẩn của nó theo thời gian. Một lớp LSTM bao gồm một hoặc nhiều đơn vị LSTM và một đơn vị LSTM bao gồm các ô và cổng để thực hiện phân loại và dự đoán dựa trên dữ liệu chuỗi thời gian.

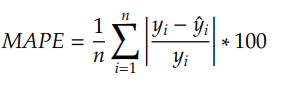
Cấu trúc LSTM được hình thành bởi hai lớp. Kích thước đầu vào được xác định bởi j sau thuật toán PCA. Lớp đầu tiên là lớp LSTM đầu vào và lớp thứ hai là lớp đầu ra. Kết quả cuối cùng sẽ là 0 hoặc 1 cho biết kết quả dự đoán xu hướng giá cổ phiếu là đi xuống hay đi lên, như một gợi ý hỗ trợ cho các nhà đầu tư thực hiện các quyết định đầu tư tiếp theo.

* 1. **Phương pháp đánh giá**

Khi xây dựng một mô hình máy học, chúng ta cần một phép đánh giá để xem mô hình sử dụng có hiệu quả không và để so sánh khả năng của các mô hình. Có rất nhiều cách đánh giá một mô hình hồi quy. Tuỳ vào những bài toán khác nhau mà chúng ta sử dụng các phương pháp khác nhau

* + 1. *MAPE*

Sai số phần trăm tỉ lệ tuyệt đối trung bình (Mean Absolute Percentage Error - MAPE) là một trong các phương pháp đánh giá thường được sử dụng cho bài toán hồi quy. MAPE đo lường hiệu suất của mô hình bằng cách tính toán phần trăm lỗi tuyệt đối, với công thức được định nghĩa như trong hình sau.

****

Ưu điểm:

• Được biểu diễn dưới dạng phần trăm, không phụ thuộc vào kích thước mẫu có thể dùng để so sánh các dự đoán trên các mẫu có kích thước khác nhau.

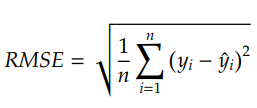
• Dễ hiểu, dễ hình dung.

Nhược điểm:

• Nhận giá trị không xác định khi mẫu (yi) có giá trị bằng 0, hoặc nhận giá trị rất lớn khi giá trị mẫu rất gần 0.

* + 1. *RMSE*

Sai số trung bình bình phương gốc (Root Mean Squared Error - RMSE) tính căn bậc hai của trung bình sai số bình phương giữa giá trị dự đoán và thực tế của mẫu.

****

Ưu điểm:

• RMSE đảm bảo cho mô hình hạn chế dự đoán các giá trị ngoại lai với sai số lớn, vì RMSE đặt trọng số lớn hơn cho các lỗi lớn do phần bình phương trong công thức tính toán.

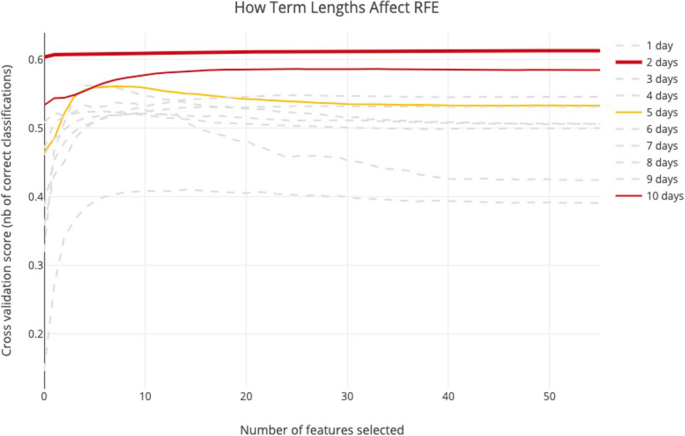
Nhược điểm:

• Nếu mô hình đưa ra một dự đoán rất tệ, phần bình phương của hàm sẽ phóng đại sai số. Tuy nhiên, trong nhiều trường hợp thực tế, không quan 6 tâm nhiều đến những ngoại lệ này và hướng tới một mô hình toàn diện, hoạt động đủ tốt đối với số đông. Điểm giống nhau của MAE và RMSE là cả hai đều là những giá trị không âm, và với những giá trị đánh giá thấp hơn sẽ tốt hơn. Sự khác biệt quan trọng của RMSE so với MAE là do các lỗi được bình phương trước khi lấy trung bình, nên RMSE cho trọng số tương đối cao với các lỗi lớn. Điều này có nghĩa là RMSE hữu ích hơn khi gặp các lỗi lớn.

# Chương 3: KẾT QUẢ THỰC NGHIỆM

* 1. **Tình hình thực nghiệm**
* Một số quy trình tác động đến hiệu quả nhưng không ảnh hưởng đến độ chính xác hoặc độ chụm và ngược lại, trong khi các quy trình khác có thể ảnh hưởng đến cả hiệu quả và kết quả dự đoán. Để đánh giá đầy đủ thiết kế thuật toán của mình, nhóm tác giả cấu trúc phần đánh giá theo các quy trình chính và đánh giá xem mỗi quy trình ảnh hưởng như thế nào đến hiệu suất thuật toán. Đầu tiên, nhóm nghiên cứu đánh giá giải pháp của mình trên máy có bộ xử lý i7 2,2 GHz, RAM 16 GB. Hơn nữa, họ cũng đã đánh giá giải pháp của mình trên phiên bản Amazon EC2, Bộ xử lý 3,1 GHz với 16 vCPU và RAM 64 GB.
* Nhóm nghiên cứu đã chọn ngẫu nhiên hai phần ba dữ liệu chứng khoán theo ID chứng khoán để đào tạo RFE và lưu ý tập dữ liệu là DS\_train\_f; tất cả dữ liệu bao gồm các chỉ số kỹ thuật đầy đủ và các tính năng mở rộng trong suốt năm 2018. Công cụ ước tính của thuật toán RFE là SVR với các hạt nhân tuyến tính. Chúng tôi xếp hạng 54 tính năng bằng cách bỏ phiếu và nhận được 30 tính năng hiệu quả, sau đó xử lý chúng bằng thuật toán PCA để thực hiện giảm kích thước và giảm các tính năng thành 20 thành phần chính. Phần còn lại của dữ liệu chứng khoán tạo thành tập dữ liệu thử nghiệm DS\_test\_f để xác thực tính hiệu quả của các thành phần chính mà chúng tôi đã trích xuất từ ​​các tính năng đã chọn. Chúng tôi đã cải tổ tất cả dữ liệu từ năm 2018 dưới dạng tập dữ liệu huấn luyện của mô hình dữ liệu và được ghi chú là DS\_train\_m. Bộ dữ liệu thử nghiệm mô hình DS\_test\_m bao gồm dữ liệu của 3 tháng đầu năm 2019, không trùng lặp với tập dữ liệu mà chúng tôi đã sử dụng trong các bước trước. Cách tiếp cận này là để ngăn chặn vấn đề tiềm ẩn do trang bị quá mức gây ra.
  1. **Kết quả áp dụng thuật toán RFE.**
     1. *Kết quả*

  Nhóm nghiên cứu quyết định sử dụng dữ liệu chỉ số trước 1 ngày để dự đoán xu hướng giá của ngày hôm sau. Chúng tôi đã thử nghiệm thuật toán RFE trong phạm vi ngắn hạn từ 1 ngày đến 2 tuần (mười ngày giao dịch) để đánh giá mức độ tương quan của các chỉ số kỹ thuật thường được sử dụng với xu hướng giá. Để đánh giá độ dài của thuật ngữ dự đoán, nhóm nghiên cứu đã mở rộng đầy đủ các tính năng và cung cấp chúng cho RFE. Trong quá trình thử nghiệm, chúng tôi nhận thấy rằng độ dài khác nhau của thuật ngữ có mức độ nhạy cảm khác nhau đối với cùng một bộ chỉ số.



Hình 3. 1. Biểu diễn RFE

Ta lấy giá đóng cửa của ngày giao dịch đầu tiên và so sánh với giá đóng cửa của ngày *giao* dịch thứ *n* . Vì chúng tôi đang dự đoán xu hướng giá nên chúng tôi không xem xét thời hạn nếu điểm xác thực chéo dưới 0,5. Và sau khi thử nghiệm, như chúng ta có thể thấy trong Hình  **3.1** , có độ dài ba kỳ hạn nhạy cảm nhất với các chỉ số mà chúng tôi đã chọn từ các công trình liên quan. Chúng là *n* = {2, 5, 10}, điều này cho thấy dự đoán xu hướng giá cách ngày, 1 tuần và 2 tuần bằng cách sử dụng bộ chỉ số có thể đáng tin cậy hơn.

Trong Bảng  **3.1** đã liệt kê 15 tính năng hiệu quả hàng đầu cho các độ dài ba giai đoạn này. Nếu chúng tôi dự đoán xu hướng giá của mỗi ngày, thì điểm xác thực chéo chỉ dao động theo số lượng tính năng được chọn. Vì vậy, trong bước tiếp theo, chúng tôi sẽ đánh giá kết quả RFE cho các độ dài ba kỳ hạn này, như trong Hình  **3.1**.

|  |  |  |  |
| --- | --- | --- | --- |
| **xếp hạng có liên quan** | **hai ngày một lần** | **hàng tuần** | **Hai tuần một lần** |
| 1st | Up\_down | SLOWK\_maxmin | MTM\_10\_plr |
| 2nd | Change | SLOWK | ROC\_10\_plr |
| 3rd | pct\_chg | SLOWD\_maxmin | WNR\_9 |
| 4th | Low | RSI\_5\_maxmin | WNR\_9\_maxmin |
| 5th | RSI\_5\_flc | SLOWD | SLOWK |
| 6th | Open | RSI\_5 | SLOWK\_maxmin |
| 7th | Amount | SLOWK\_flc | ROC\_10 |
| 8th | Amount\_maxmin | WNR\_9\_maxmin | SLOWD\_flc |
| 9th | Vol | WNR\_9 | WNR\_9\_flc |
| 10th | BIAS\_20\_maxmin | CCI\_24 | RSI\_5 |
| 11th | High | BIAS\_20\_maxmin | BIAS\_20\_maxmin |
| 12th | Vol\_maxmin | BIAS\_20 | RSI\_5\_maxmin |
| 13th | ROC\_10 | ADOSC\_maxmin | BIAS\_20 |
| 14th | ADOSC\_maxmin | ADOSC | SMA\_10 |
| 15th | ADOSC | WNR\_9\_flc | SLOWD |
| ….. | …… | …… | …… |
| **Số tính năng được chọn** | **48 tính năng được chọn** | **8 tính năng được chọn** | **29 tính năng được chọn** |

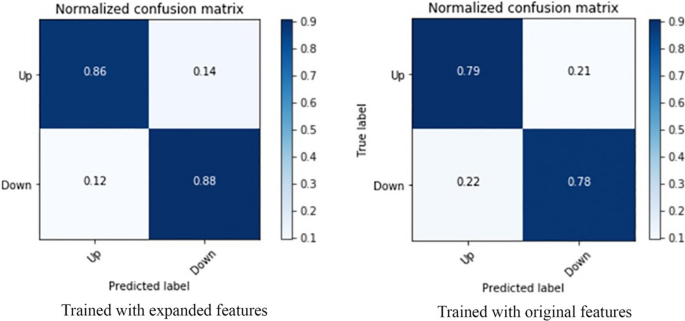
bảng 3. 1. Các tính năng hiệu quả tương ứng với thời hạn

Nhóm nghiên cứu so sánh bộ tính năng đầu ra của RFE với bộ tính năng hoàn toàn gốc làm cơ sở, bộ tính năng hoàn toàn gốc bao gồm n tính năng và chúng tôi chọn *n* tính năng hiệu quả nhất từ ​​các tính năng đầu ra RFE để đánh giá kết quả bằng cách sử dụng SVR tuyến tính. Chúng tôi đã sử dụng hai cách tiếp cận khác nhau để đánh giá tính hiệu quả của tính năng. Phương pháp đầu tiên là kết hợp tất cả dữ liệu vào một ma trận lớn và đánh giá chúng bằng cách chạy thuật toán RFE một lần. Một phương pháp khác là chạy RFE cho từng cổ phiếu riêng lẻ và tính toán các tính năng hiệu quả nhất bằng cách bỏ phiếu.

* + 1. *Tiện ích mở rộng tính năng RFE.*

Từ kết quả của tiểu mục trước, chúng ta có thể thấy rằng khi dự đoán xu hướng giá cho hai ngày một lần hoặc hai tuần một lần, kết quả tốt nhất đạt được bằng cách chọn một số lượng lớn các tính năng. Trong các đặc trưng được chọn, một số đặc trưng được xử lý từ phương pháp mở rộng có thứ hạng tốt hơn so với đặc trưng gốc, điều này chứng tỏ phương pháp mở rộng đặc trưng rất hữu ích cho việc tối ưu mô hình. Phần mở rộng tính năng ảnh hưởng đến cả độ chính xác và hiệu quả, trong phần này, chúng tôi chỉ thảo luận về khía cạnh độ chính xác và để lại phần hiệu quả trong bước tiếp theo vì PCA là phương pháp hiệu quả nhất để tối ưu hóa hiệu quả đào tạo trong thiết kế của chúng tôi. Chúng tôi đã tham gia đánh giá mức độ ảnh hưởng của tiện ích mở rộng tính năng đến RFE và sử dụng kết quả thử nghiệm để đo lường sự cải thiện của tiện ích mở rộng tính năng liên quan.

Nhóm nghiên cứu kiểm tra thêm tính hiệu quả của tiện ích mở rộng tính năng, tức là nếu phân cực, thang đo tối đa–tối thiểu và tính toán tỷ lệ phần trăm dao động hoạt động tốt hơn các chỉ số kỹ thuật ban đầu. Trường hợp tốt nhất để tận dụng thử nghiệm này là dự đoán hàng tuần vì nó có tính năng kém hiệu quả nhất được chọn.



Hình 3. 2. Ma trận nhầm lẫn xác nhận tính hiệu quả của tiện ích mở rộng tính năng

Ma trận nhầm lẫn đã chuẩn hóa của thử nghiệm hai bộ tính năng được minh họa trong Hình  **3.2** . Phần bên trái là ma trận nhầm lẫn của bộ tính năng với các tính năng mở rộng và phần bên phải bên cạnh là kết quả thử nghiệm của việc chỉ sử dụng các tính năng gốc. Cả hai độ chính xác của dương tính thực và âm tính thực đã được cải thiện lần lượt là 7% và 10%, điều này chứng tỏ rằng thiết kế phương pháp mở rộng tính năng của chúng tôi có hiệu quả hợp lý.

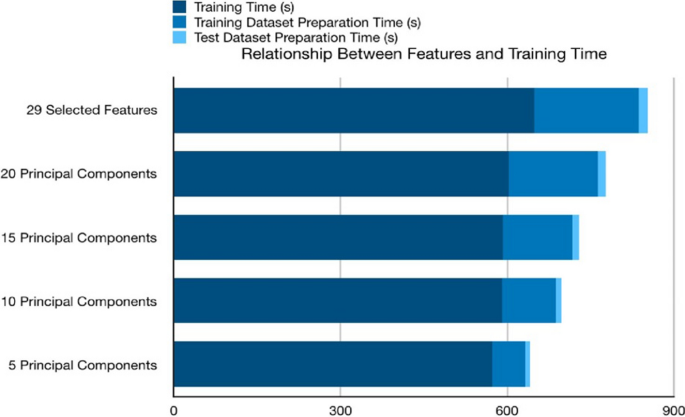
* + 1. *Giảm tính năng bằng cách sử dụng principal component analysis(PCA)*

PCA sẽ ảnh hưởng đến hiệu suất thuật toán về cả độ chính xác của dự đoán và hiệu quả đào tạo, trong khi phần này sẽ được đánh giá bằng mô hình NN, vì vậy nhóm nghiên cứu cũng đã xác định mô hình DNN đơn giản nhất với ba lớp như chúng tôi đã sử dụng trong bước trước để thực hiện đánh giá.

Trong phần này, nhóm nghiên cứu sẽ chọn dự đoán hai tuần một lần để thực hiện phân tích trường hợp sử dụng, vì nó có đường cong điểm số xác thực chéo tăng dần, hơn nữa, không giống như mọi dự đoán hàng ngày, nó đã loại bỏ hơn 20 tính năng không hiệu quả. Ở bước đầu tiên, chúng tôi chọn tất cả 29 tính năng hiệu quả và đào tạo mô hình NN mà không cần thực hiện PCA. Nó tạo ra một đường cơ sở về độ chính xác và thời gian đào tạo để so sánh. Để đánh giá độ chính xác và hiệu quả, chúng tôi giữ nguyên số lượng thành phần chính là 5, 10, 15, 20, 25. Bảng  **3.2** ghi lại cách số lượng tính năng ảnh hưởng đến hiệu quả đào tạo mô hình, sau đó sử dụng biểu đồ thanh ngăn xếp trong Hình  **3.3** để minh họa cách PCA ảnh hưởng đến hiệu quả đào tạo. Bảng  **3.4** hiển thị phân tích độ chính xác và hiệu quả trên các quy trình khác nhau để xử lý trước các tính năng. Thời gian được hiển thị trong Bảng  **3.2, 3.4** dựa trên các thử nghiệm được thực hiện trong máy người dùng tiêu chuẩn để cho thấy khả năng tồn tại của giải pháp của nhóm nghiên cứu với nguồn lực sẵn có hạn chế hoặc trung bình.

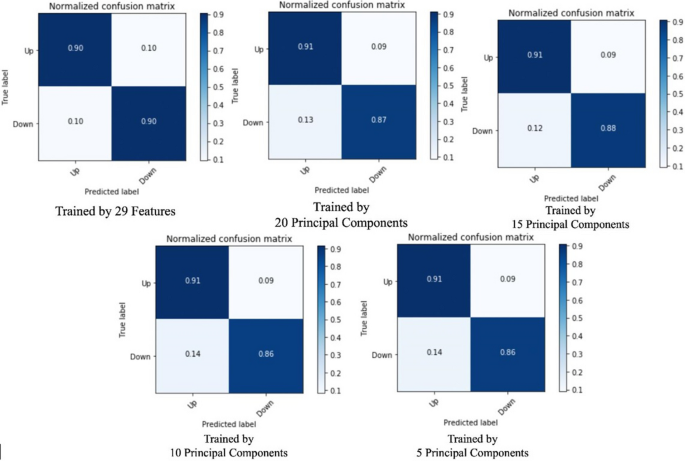
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Số lượng tính năng** | **Thời gian chuẩn bị tập dữ liệu huấn luyện (s)** | **Thời gian chuẩn bị tập dữ liệu thử nghiệm (s)** | **(Các) thời gian đào tạo** | **Tổng (s)** |
| 29 tính năng được chọn | 187.46 | 16h30 | 648.53 | 852.29 |
| 20 thành phần chủ yếu | 160,29 | 14.24 | 602.68 | 777.21 |
| 15 thành phần chủ yếu | 125.20 | 12.18 | 591.93 | 729.31 |
| 10 thành phần chủ yếu | 96,54 | 0.37 | 590.76 | 697.67 |
| 5 thành phần chủ yếu | 59,37 | 8,22 | 572.88 | 640.47 |

bảng 3. 2. Mối quan hệ giữa số lượng các thành phần chính và hiệu quả đào tạo



Hình 3. 3. Mối quan hệ giữa số tính năng và thời gian đào tạo

Nhóm nghiên cứu cũng liệt kê ma trận nhầm lẫn của từng bài kiểm tra trong Hình  **3.4** . Biểu đồ thanh ngăn xếp cho thấy tổng thời gian dành cho việc đào tạo mô hình đang giảm theo số lượng tính năng được chọn, trong khi phương pháp PCA có hiệu quả đáng kể trong việc tối ưu hóa việc chuẩn bị tập dữ liệu đào tạo. Đối với thời gian dành cho giai đoạn đào tạo, PCA không hiệu quả bằng giai đoạn chuẩn bị dữ liệu. Mặc dù có khả năng hiệu quả tối ưu hóa của PCA không đủ mạnh do cấu trúc đơn giản của mô hình NN.



Hình 3. 4. Số lượng các thành phần chính ảnh hưởng đến kết quả đánh giá như thế nào

Bảng  **3.3** chỉ ra rằng độ chính xác dự đoán tổng thể không bị ảnh hưởng nhiều khi giảm kích thước. Tuy nhiên, độ chính xác không thể hỗ trợ đầy đủ nếu PCA không có tác dụng phụ đối với dự đoán mô hình, vì vậy chúng tôi đã xem xét ma trận nhầm lẫn của kết quả thử nghiệm.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Số tính năng đã chọn** | **5 thành phần chủ yếu** | **10 thành phần chủ yếu** | **15 thành phần chủ yếu** | **20 thành phần chủ yếu** | **29 thành phần chủ yếu** |
| Sự chính xác | 89,35% | 89,35% | 89,39% | 89,30% | 90,29% |

bảng 3. 3. Số lượng các tính năng được chọn ảnh hưởng như thế nào đến độ chính xác của dự đoán

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tiền xử lý tính năng** | **Độ chính xác tổng thể (%)** | **Thời gian chuẩn bị tập dữ liệu huấn luyện (s)** | **Kiểm tra thời gian chuẩn bị tập dữ liệu (s)** | **thời gian huấn luyện(s)** | **Tổng (s)** |
| Max–min scaling | 89.30 | 160.28 | 14.24 | 602.68 | 777.20 |
| Normalization | 78.17 | 157.63 | 14.73 | 596.22 | 768.58 |
| N/A | 78.88 | 142.17 | 13.00 | 595.52 | 750.69 |

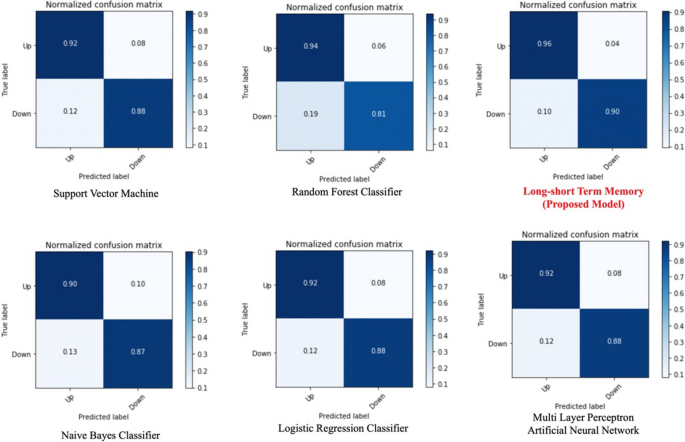
bảng 3. 4. Phân tích độ chính xác và hiệu quả của các quy trình tiền xử lý tính năng

nhóm nghiên cứu đã tận dụng một thử nghiệm khác về việc thêm các thủ tục trước khi trích xuất 20 thành phần chính từ tập dữ liệu gốc và so sánh các khía cạnh về thời gian trôi qua của giai đoạn đào tạo và độ chính xác của dự đoán. Tuy nhiên, kết quả kiểm tra dẫn đến những kết luận khác nhau. Trong Bảng  **3.4** , chúng ta có thể kết luận rằng tiền xử lý tính năng không có tác động đáng kể đến hiệu quả đào tạo, nhưng nó ảnh hưởng đến độ chính xác dự đoán của mô hình.

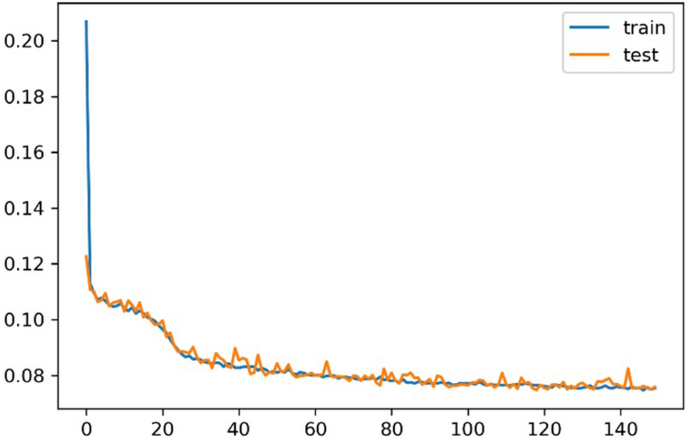
* 1. **So sánh với các công trình đã công bố**

Từ các công trình trước đây, nhóm nghiên nhận thấy các mô hình được khai thác phổ biến nhất để dự đoán xu hướng giá thị trường chứng khoán ngắn hạn là support vector machine (SVM), mạng thần kinh nhân tạo multilayer perceptron (MLP), bộ phân loại Naive Bayes (NB), bộ phân loại random forest (RAF). và phân loại logistic regression (LR). Trường hợp so sánh thử nghiệm cũng là dự đoán xu hướng giá hai tuần một lần, để đánh giá kết quả tốt nhất của tất cả các mô hình, chúng tôi giữ tất cả 29 tính năng được chọn bởi thuật toán RFE. Đối với đánh giá MLP, để kiểm tra xem số lượng lớp ẩn có ảnh hưởng đến điểm số hay không, chúng tôi đã ghi số lớp là *n* và kiểm tra *n*= {1, 3, 5}, 150 giai đoạn huấn luyện cho tất cả các bài kiểm tra, đã tìm thấy sự khác biệt nhỏ trong hiệu suất của mô hình, điều này cho thấy rằng biến số lớp MLP hầu như không ảnh hưởng đến điểm số.

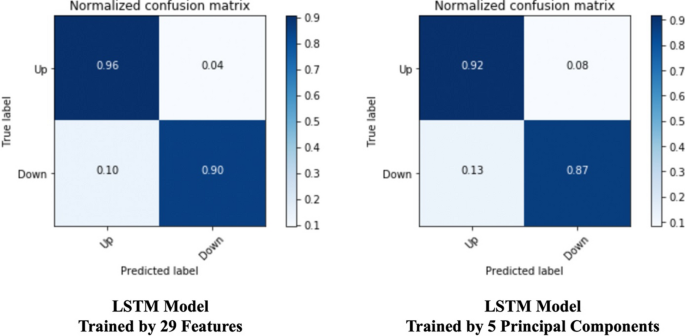
Từ các ma trận nhầm lẫn trong Hình  **3.5**, chúng ta có thể thấy tất cả các mô hình máy học hoạt động tốt khi đào tạo với bộ tính năng đầy đủ mà nhóm nghiên cứu đã chọn bởi RFE. Từ góc độ thời gian đào tạo, đào tạo mô hình NB đạt hiệu quả tốt nhất. Thuật toán LR tốn ít thời gian đào tạo hơn so với các thuật toán khác trong khi nó có thể đạt được kết quả dự đoán tương tự với các mô hình tốn kém khác như SVM và MLP. Thuật toán RAF đạt được tỷ lệ dương tính thực tương đối cao trong khi hiệu suất kém trong việc dự đoán các nhãn âm tính. Đối với mô hình LSTM được đề xuất của nhóm nghiên cứu, nó đạt được độ chính xác nhị phân là 93,25%, đây là độ chính xác cao đáng kể để dự đoán xu hướng giá hai tuần một lần. nhóm nghiên cứu cũng đã xử lý trước dữ liệu thông qua PCA và có năm thành phần chính, sau đó được đào tạo trong 150 kỷ nguyên. Đường cong học tập của giải pháp được đề xuất của chúng tôi, dựa trên kỹ thuật tính năng và mô hình LSTM, được minh họa trong Hình **3.6** . Ma trận nhầm lẫn là hình bên phải trong Hình  **3.7** .



Hình 3. 5. So sánh dự đoán mô hình—ma trận nhầm lẫn



Hình 3. 6. Đường cong học tập của giải pháp đề xuất



Hình 3. 7. So sánh chính xác dự đoán mô hình được đề xuất—ma trận nhầm lẫn

Nhóm nghiên cứu đã liệt kê hiệu suất của mô hình dự đoán xu hướng giá cổ phiếu hàng đầu trong Bảng  **3.5** , từ các số liệu có thể so sánh được, điểm số của giải pháp được đề xuất của chúng tôi thường tốt hơn so với các công việc liên quan khác. Thay vì tùy tiện kết luận rằng mô hình được đề xuất của chúng tôi vượt trội so với các mô hình khác trong các công trình liên quan, trước tiên chúng tôi xem xét cột tập dữ liệu của Bảng  **3.5** . Bằng cách xem xét bộ dữ liệu được sử dụng bởi từng công việc, chỉ đào tạo và thử nghiệm giải pháp đề xuất của họ trên ba cổ phiếu riêng lẻ, điều này khó chứng minh tính tổng quát của mô hình đề xuất của họ. Ayo sử dụng phân tích đòn bẩy trên dữ liệu chứng khoán từ Sở giao dịch chứng khoán New York (NYSE), trong khi điểm yếu là họ chỉ thực hiện phân tích về giá đóng cửa, đây là một tính năng có độ nhiễu cao. Zubair và cộng sự, đã đào tạo mô hình đề xuất của họ trên cả cổ phiếu riêng lẻ và giá chỉ số, nhưng như chúng tôi đã đề cập trong phần trước, giá chỉ số chỉ bao gồm một số tính năng và ID cổ phiếu hạn chế, điều này sẽ ảnh hưởng nhiều hơn đến chất lượng đào tạo mô hình. Đối với giải pháp đề xuất của chúng tôi, chúng tôi đã thu thập đầy đủ dữ liệu từ thị trường chứng khoán Trung Quốc và áp dụng thuật toán FE + RFE trên các chỉ số ban đầu để có được các tính năng hiệu quả hơn, kết quả đánh giá toàn diện của 3558 ID cổ phiếu có thể giải thích một cách hợp lý tính tổng quát và hiệu quả của giải pháp đề xuất của chúng tôi trên thị trường chứng khoán Trung Quốc. Tuy nhiên, các tác giả của Khaidem và Dey và Ayo đã chọn phân tích thị trường chứng khoán ở Hoa Kỳ, Zubair et al.  đã thực hiện phân tích về giá thị trường chứng khoán Pakistan và chúng tôi đã thu được bộ dữ liệu từ thị trường chứng khoán Trung Quốc, chính sách của các quốc gia khác nhau có thể ảnh hưởng đến hiệu suất của mô hình, cần nghiên cứu thêm để xác thực.

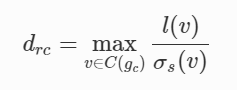
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Công việc liên quan** | **tập dữ liệu** | **Mô hình** | **Sự chính xác** | **Độ chính xác** | **Hồi tưởng** |
| Khaidem và Dey | Dữ liệu giá cổ phiếu của AAPL, GE và Samsung Electronics Co. Ltd. | Random forest | 0.83 | 0.82 | 0.81 |
| Ayo | Giá đóng cửa của dữ liệu chứng khoán từ Sở giao dịch chứng khoán New York (NYSE) | ARIMA | 0.90 | 0.91 | 0.92 |
| Zubair và cộng sự | Chỉ số KSE 100 Lucky Cement Stock Engro Fertilizer Limited | Multiple regression | 0.94 | 0.95 | 0.93 |
| (Giải pháp đề xuất) | Dữ liệu giá cổ phiếu 3558 ID từ năm 2017 đến 2018 được thu thập từ thị trường chứng khoán Trung Quốc | Proposed Model—FE + RFE + PCA + LSTM | 0.93 | 0.96 | 0.96 |

bảng 3. 5. So sánh giải pháp đề xuất với các công trình liên quan

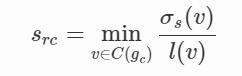
* 1. **Phân tích độ phức tạp của giải pháp đề xuất**

Phần này phân tích sự phức tạp của giải pháp đề xuất của nhóm nghiên cứu. Bộ nhớ ngắn hạn dài khác với các NN khác và nó là một biến thể của RNN tiêu chuẩn, cũng có các bước thời gian với kiến ​​trúc cổng và bộ nhớ. Trong công trình trước tác giả đã thực hiện phân tích độ phức tạp của kiến ​​trúc RNN. Họ đã giới thiệu một phương pháp coi RNN là một đồ thị tuần hoàn có hướng và đề xuất một khái niệm về độ sâu hồi quy, giúp thực hiện phân tích về độ phức tạp của RNN.

Độ sâu lặp lại là một số hữu tỷ dương và chúng tôi biểu thị nó là . Khi sự tăng trưởng của các biện pháp n số đo, số trung bình chuyển đổi phi tuyến tính tối đa của mỗi bước thời gian. Sau đó, chúng tôi mở ra biểu đồ tuần hoàn có hướng của RNN và biểu thị biểu đồ  được xử lý là , trong khi đó, biểu thị (C(gc)) là tập hợp các chu trình có hướng trong biểu đồ này. Đối với đỉnh v , chúng tôi lưu ý là tổng của các trọng số cạnh và (l(v)) như chiều dài. Phương trình dưới đây được chứng minh theo một giả định nhẹ.



Họ cũng phát hiện ra rằng một yếu tố quan trọng khác ảnh hưởng đến hiệu suất của LSTM, đó là hệ số bỏ qua lặp lại. Chúng tôi lưu ý đối ứng của hệ số bỏ qua định kỳ. vui lòng lưu ý rằng Src cũng là một số hữu tỉ dương.

****

Theo định nghĩa trên, mô hình đề xuất của chúng tôi là LSTM xếp chồng 2 lớp, trong đó đrc= 2 và Src= 1. Từ các thí nghiệm được thực hiện trong công việc trước đây, các tác giả cũng nhận thấy rằng khi đối mặt với các vấn đề về sự phụ thuộc dài hạn, LSTM có thể được hưởng lợi từ việc giảm nghịch đảo của các hệ số bỏ qua lặp lại và tăng độ sâu lặp lại. Các phát hiện thực nghiệm được đề cập ở trên rất hữu ích để nâng cao hiệu suất của mô hình đề xuất của chúng tôi hơn nữa.

# Chương 4: KẾT LUẬN VÀ KIẾN NGHỊ

* 1. **Kết luận**
* Giải pháp đề xuất của nhóm nghiên cứu là một tùy chỉnh độc đáo so với các công trình trước đó vì thay vì chỉ đề xuất một mô hình LSTM hiện đại khác, nhóm nghiên cứu đã đề xuất một hệ thống dự đoán học sâu tùy chỉnh và tinh chỉnh cùng với việc sử dụng kỹ thuật tính năng toàn diện và kết hợp nó với LSTM để thực hiện dự đoán. Bằng cách nghiên cứu các quan sát từ các công trình trước đó, chúng tôi lấp đầy khoảng trống giữa các nhà đầu tư và nhà nghiên cứu bằng cách đề xuất một thuật toán mở rộng tính năng trước khi loại bỏ tính năng đệ quy và nhận được sự cải thiện đáng kể về hiệu suất của mô hình.
* Mặc dù đã đạt được kết quả tốt từ giải pháp đề xuất của mình, nhưng nghiên cứu này có nhiều tiềm năng hơn đối với nghiên cứu trong tương lai. Trong quá trình đánh giá, nhóm nghiên cứu cũng nhận thấy rằng thuật toán RFE không nhạy cảm với các thời hạn khác ngoài 2 ngày, hàng tuần, hai tuần một lần. Nghiên cứu sâu hơn về những chỉ số kỹ thuật nào sẽ ảnh hưởng đến độ dài kỳ hạn bất thường sẽ là một hướng nghiên cứu khả thi trong tương lai. Ngoài ra, bằng cách kết hợp các kỹ thuật phân tích tình cảm mới nhất với kỹ thuật tính năng và mô hình học sâu, cũng có tiềm năng cao để phát triển một hệ thống dự đoán toàn diện hơn, được đào tạo bởi nhiều loại thông tin như tweet, tin tức và dữ liệu dựa trên văn bản khác.
  1. **Đề xuất mới**
* Nghiên cứu của nhóm tác giả đã phân tích huấn luyện mô hình có một độ chính xác cao lên đến 93%. Ngoài ra có thể sử dụng cái phương pháp kết hợp kết hợp các mô hình lại với nhau có thể đưa ra một kết quả chính xác hơn ví dụ như CNN kết hợp với LSTM ….
  1. **Tài liệu tham khảo với trích dẫn**

[1] Brownlee J. Deep learning for time series forecasting: predict the future with MLPs, CNNs and LSTMs in Python. Machine Learning Mastery. 2018. <https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/>

[2] ] Cs 229 - deep learning cheatsheet. https://stanford.edu/~shervine/teaching/ cs-229/cheatsheet-deep-learning.

[3] <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00333-6>

[4] <https://www.asktraders.com/how-many-traders-can-be-found-globally/>

[5] <https://doi.org/10.1016/j.asoc.2014.12.028>.

[6] <https://dominhhai.github.io/vi/2017/10/what-is-lstm/>

[7] <https://doi.org/10.1162/neco.1997.9.8.1735>

[8]<https://www.sciencedirect.com/science/article/abs/pii/S0957417407006069?via%3Dihub>

[9] https:// vanban.chinhphu.vn/default.aspx?pageid=27160&docid=198541.

[10] <https://nttuan8.com/bai-3-neuralnetwork/>

[11] <https://dangcongsan.vn/kinh-te-va-hoi-nhap/thang-8-2021-so-tai-khoan-chung-khoan-mo-lon-thu-hai-trong-lich-su-thi-truong-viet-nam-590799.html>

[12] <https://www.sciencedirect.com/science/article/abs/pii/S0957417409001560?via%3Dihub>

[13] <https://finance.yahoo.com/>

[14] https://www.tiingo.com/