SƯỜN BÁO CÁO ĐỒ ÁN HỆ HỖ TRỢ RA QUYẾT ĐỊNH
Đề tài: Predicting stock market index using fusion of machine learning techniques
Giảng viên hướng dẫn:
Sinh viên thực hiện:
TP. Hồ Chí Minh, tháng 5 năm 2021

LỜI CẢM ƠN

• Lời cảm ơn chân thành gửi đến quý Thầy cô giảng viên Trường Đại học Công nghệ thông tin - Đại học Quốc gia TP.HCM đã cung cấp kiến thức nền tảng.
• Lời cảm ơn sâu sắc đến
• Nhóm đã vận dụng kiến thức nền tảng và học hỏi, nghiên cứu kiến thức mới từ thầy cô, bạn bè, và tài liệu tham khảo để hoàn thành báo cáo.
• Công nhận những thiếu sót do kiến thức chuyên môn và kinh nghiệm thực tiễn còn hạn chế, mong nhận được góp ý để hoàn thiện.
• Gửi lời cảm ơn chân thành đến thầy cô, bạn bè.

NHẬN XÉT CỦA GIẢNG VIÊN

MỤC LỤC

CHƯƠNG I. GIỚI THIỆU

1.1. Tổng quan đề tài

1.1.1. Bối cảnh nghiên cứu

Dự đoán giá cổ phiếu là một bài toán kinh điển trong lĩnh vực tài chính và đã thu hút sự quan tâm của nhiều nhà nghiên cứu trong suốt nhiều thập kỷ qua. Theo quan điểm của các nhà phân tích kỹ thuật, hầu hết thông tin về cổ phiếu được phản ánh trong giá gần đây, và việc phân tích xu hướng chuyển động có thể giúp dự đoán giá cách hiệu quả.

Thị trường chứng khoán là một hệ thống phức tạp chịu tác động của nhiều yếu tố kinh tế vĩ mô và vi mô. Diễn biến của thị trường chứng khoán bị ảnh hưởng bởi một loạt các yếu tố bao gồm:

- Yếu tố chính trị: Các sự kiện chính trị, chính sách của chính phủ
- Yếu tố kinh tế: Điều kiện kinh tế chung, chỉ số giá hàng hóa, tỷ giá ngân hàng, tỷ giá hối đoái Yếu tố doanh nghiệp: Chính sách và hiệu suất hoạt động của các công ty
- Yếu tố tâm lý: Kỳ vọng và tâm lý của nhà đầu tư, lựa chọn của nhà đầu tư tổ chức
- Yếu tố thị trường: Diễn biến của các thị trường chứng khoán khác Giá trị các chỉ số chứng khoán được tính toán dựa trên các cổ phiếu có giá trị vốn hóa thị trường cao, cung cấp một bức tranh tổng thể về tình hình nền kinh tế và phụ thuộc vào nhiều yếu tố phức tạp.

1.1.2. Vấn đề nghiên cứu

Mặc dù đã có nhiều nghiên cứu về dự đoán giá cổ phiếu, các phương pháp hiện có vẫn tồn tại những hạn chế đáng kể:

Hạn chế chính: Các phương pháp truyền thống thường chỉ sử dụng một lớp dự đoán duy nhất, trong đó các tham số thống kê của ngày 't' được sử dụng làm đầu vào để dự đoán giá đóng cửa của ngày '(t+n)'.

Vấn đề cụ thể: Khi giá trị 'n' (số ngày dự đoán trước) tăng lên, việc dự đoán dựa trên các giá trị ngày càng cũ của tham số thống kê dẫn đến độ chính xác không đủ và sai số tăng cao.

1.1.3. Nhu cầu giải quyết

Để khắc phục các hạn chế nêu trên, cần phát triển một phương pháp tiếp cận mới có khả năng:

- Thu hẹp khoảng cách thời gian giữa dữ liệu đầu vào và kết quả dự đoán
- Giảm thiểu lỗi dự đoán theo từng giai đoạn
- Cải thiện độ chính xác khi dự đoán cho các khoảng thời gian dài hơn
- Giải pháp được đề xuất là sử dụng sơ đồ dự đoán hai giai đoạn kết hợp các kỹ thuật máy học khác nhau để tối ưu hóa quá trình dự đoán.

1.2. Mục tiêu, Phạm vi và Đối tượng nghiên cứu

1.2.1. Mục tiêu nghiên cứu

Mục tiêu tổng quát:

Nghiên cứu tập trung vào nhiệm vụ dự đoán giá trị tương lai của chỉ số thị trường chứng khoán bằng cách sử dụng kết hợp các kỹ thuật máy học tiên tiến.

Mục tiêu cụ thể:

- Phát triển mô hình dự đoán hai giai đoạn để cải thiện độ chính xác dự đoán
- So sánh hiệu suất của các thuật toán máy học khác nhau trong dự đoán chỉ số chứng khoán
- Đánh giá hiệu quả của phương pháp kết hợp (fusion) so với các phương pháp đơn lẻ
- Xây dựng hệ thống hỗ trợ ra quyết định cho việc dự đoán chỉ số thị trường chứng khoán

1.2.2. Phạm vi nghiên cứu

Phạm vi thời gian: Nghiên cứu sử dụng dữ liệu lịch sử mười năm từ tháng 01/2003 đến tháng 12/2012 cho nghiên cứu gốc, và dữ liệu từ 01/01/2006 đến 01/04/2021 cho nghiên cứu thực nghiệm.

Phạm vi địa lý: Tập trung vào thị trường chứng khoán Ấn Độ với hai chỉ số chính:

- CNX Nifty (National Stock Exchange)
- S&P BSE Sensex (Bombay Stock Exchange)

Phạm vi kỹ thuật:

- Sử dụng 10 chỉ báo kỹ thuật chính
- Áp dụng 3 thuật toán máy học: ANN, SVR, Random Forest
- Phát triển mô hình kết hợp hai giai đoạn

1.2.3. Đối tượng nghiên cứu Đối tượng chính: Giá cổ phiếu và các chỉ số thị trường chứng khoán, cụ thể:

Chỉ số CNX Nifty Chỉ số S&P Bombay Stock Exchange Sensex

Các chỉ báo kỹ thuật liên quan (SMA, EMA, MACD, RSI, v.v.)

Đối tượng nghiên cứu phụ: Các mô hình và thuật toán máy học được áp dụng trong dự đoán tài chính.

1.3. Mô tả bài toán

1.3.1. Input (Đầu vào)

Đầu vào của hệ thống bao gồm 10 chỉ báo kỹ thuật mô tả tình trạng thị trường tại ngày thứ 't':

- Simple Moving Average (SMA) - Đường trung bình động đơn giản
- Exponential Moving Average (EMA) - Đường trung bình động hàm mũ
- Momentum (MOM) - Chỉ báo đo lường tốc độ thay đổi giá
- Stochastic (STC) - Chỉ số phản ánh vùng quá mua/quá bán
- Relative Strength Index (RSI) - Chỉ số sức mạnh tương đối
- Moving Average Convergence Divergence (MACD) - Phân kỳ hội tụ đường trung bình
- Larry Williams %R (WR) - Chỉ báo xung lượng
- Accumulation/Distribution Oscillator (ADO) - Đường tích lũy/phân phối
- Commodity Channel Index (CCI) - Chỉ số kênh hàng hóa
- Volume và các thông tin giao dịch khác

1.3.2. Process (Quá trình xử lý)

Đề tài đề xuất cách tiếp cận hợp nhất hai giai đoạn (Two-stage Fusion Approach):

Giai đoạn 1:

Sử dụng Support Vector Regression (SVR) để dự đoán giá trị tương lai của các tham số thống kê

- Đầu vào: 10 chỉ báo kỹ thuật của ngày 't'
- Đầu ra: 10 chỉ báo kỹ thuật dự đoán cho ngày '(t+n)'

Giai đoạn 2:

Sử dụng các thuật toán Artificial Neural Networks (ANN), Random Forest (RF) và SVR

- Đầu vào: 10 chỉ báo kỹ thuật dự đoán từ giai đoạn 1
- Đầu ra: Giá đóng cửa dự đoán cho ngày '(t+n)'

Các mô hình kết hợp được tạo ra:

- SVR – ANN: Kết hợp SVR giai đoạn 1 với ANN giai đoạn 2
- SVR – RF: Kết hợp SVR giai đoạn 1 với Random Forest giai đoạn 2
- SVR – SVR: Sử dụng SVR cho cả hai giai đoạn

1.3.3. Output (Đầu ra)

Kết quả đầu ra của hệ thống là giá đóng cửa/giá trị dự đoán của chỉ số thị trường chứng khoán cho ngày thứ '(t+n)', trong đó:

't' là ngày hiện tại (ngày có dữ liệu đầu vào) 'n' là số ngày dự đoán trước (n = 5, 10, 30 ngày)

Kết quả bao gồm các chỉ số đánh giá độ chính xác: MAE, MAPE, RMSE, MSE

1.3.4. Ý nghĩa của đề tài

Đề tài mang lại những đóng góp quan trọng:

Về mặt lý thuyết: Đề xuất phương pháp tiếp cận hai giai đoạn mới trong dự đoán tài chính

Về mặt thực tiễn: Cung cấp công cụ hỗ trợ ra quyết định cho các nhà đầu tư và tổ chức tài chính

Về mặt ứng dụng: Phương pháp có thể được tổng quát hóa cho các bài toán dự báo khác như dự báo thời tiết, tiêu thụ năng lượng, GDP

CHƯƠNG II. CÁC HƯỚNG NGHIÊN CỨU VÀ HƯỚNG TIẾP CẬN LIÊN QUAN

2.1. Tổng quan các nghiên cứu liên quan

2.1.1. Xu hướng nghiên cứu chính

Trong lĩnh vực dự đoán tài chính, các nghiên cứu có thể được phân loại theo nhiều tiêu chí khác nhau. Về mặt phương pháp tiếp cận, các nghiên cứu chủ yếu tập trung vào:

Nhóm 1: Các phương pháp học máy đơn lẻ

- Sử dụng một thuật toán cụ thể như ANN, SVM, Random Forest
- Tối ưu hóa tham số và cấu trúc mô hình
- Áp dụng trực tiếp lên dữ liệu thô hoặc đã qua tiền xử lý

Nhóm 2: Các phương pháp kết hợp (Hybrid Methods)

- Kết hợp nhiều thuật toán khác nhau
- Sử dụng một thuật toán để tối ưu hóa tham số của thuật toán khác
- Áp dụng các kỹ thuật ensemble learning

Nhóm 3: Các phương pháp tối ưu hóa

- Sử dụng các thuật toán tối ưu hóa meta-heuristic
- Kết hợp với các phương pháp học máy truyền thống
- Tập trung vào việc tìm kiếm tham số tối ưu

2.1.2. Các kỹ thuật chính được sử dụng

Các nghiên cứu trước đây đã sử dụng đa dạng các kỹ thuật bao gồm:

- **Artificial Neural Networks (ANN)**: Mạng nơ-ron nhân tạo với các biến thể khác nhau
- **Support Vector Machine/Regression (SVM/SVR)**: Máy vectơ hỗ trợ cho phân lớp và hồi quy
- **Fuzzy Logic**: Logic mờ để xử lý tính không chắc chắn
- **Genetic Algorithms (GA)**: Thuật toán di truyền cho tối ưu hóa
- **Ensemble Learning**: Các phương pháp học tập hợp
- **Time Series Analysis**: Phân tích chuỗi thời gian truyền thống

2.2. Phân tích chi tiết các nghiên cứu tiêu biểu

2.2.1. Nhóm nghiên cứu sử dụng Neural Networks

Zhang và Wu (2009): Tối ưu hóa Neural Networks với IBCO

Mục tiêu: Cải thiện hiệu suất dự đoán chỉ số chứng khoán ngắn hạn và dài hạn.

Phương pháp tiếp cận:

- Kết hợp mạng nơ-ron truyền ngược (Backpropagation Neural Network)
- Sử dụng Improved Bacterial Colony Optimization (IBCO) để tối ưu hóa trọng số
- Áp dụng cho dự đoán ngắn hạn (1 ngày) và dài hạn (15 ngày)

Đóng góp: Chứng minh rằng việc tối ưu hóa trọng số mạng nơ-ron bằng thuật toán sinh học có thể cải thiện đáng kể độ chính xác dự đoán.

Asadi et al. (2012): Kết hợp tiền xử lý và tối ưu hóa

Mục tiêu: Nâng cao độ chính xác mô hình Neural Networks thông qua tiền xử lý dữ liệu và tối ưu hóa thuật toán học.

Phương pháp tiếp cận:

- Áp dụng các phương pháp tiền xử lý dữ liệu tiên tiến
- Sử dụng Genetic Algorithm để tối ưu hóa cấu trúc mạng
- Kết hợp với thuật toán Levenberg-Marquardt (LM) cho quá trình học

Đóng góp: Cho thấy tầm quan trọng của việc tiền xử lý dữ liệu và tối ưu hóa đồng thời cấu trúc và thuật toán học.

Shen et al. (2011): AFSA-RBFNN

Mục tiêu: Phát triển mô hình dự đoán hiệu quả sử dụng Radial Basis Function Neural Network.

Phương pháp tiếp cận:

- Sử dụng Artificial Fish Swarm Algorithm (AFSA) để đào tạo RBFNN
- Tối ưu hóa các tham số của hàm cơ sở xuyên tâm
- So sánh với các phương pháp truyền thống

Đóng góp: Đề xuất phương pháp kết hợp mới cho thấy độ chính xác vượt trội trong dự đoán chỉ số chứng khoán.

2.2.2. Nhóm nghiên cứu sử dụng Support Vector Machines

Ou và Wang (2009): So sánh toàn diện các phương pháp Data Mining

Mục tiêu: Đánh giá hiệu suất của 10 kỹ thuật khai thác dữ liệu khác nhau trong dự đoán chỉ số Hang Seng.

Phương pháp tiếp cận:

- So sánh 10 phương pháp: LDA, QDA, K-NN, Naive Bayes, Logit, Tree-based, ANN, Bayesian, SVM, LS-SVM
- Đánh giá trên cùng một bộ dữ liệu với các chỉ số đánh giá thống nhất
- Phân tích hiệu suất trong các điều kiện thị trường khác nhau

Kết quả: SVM và LS-SVM cho hiệu suất vượt trội so với các phương pháp khác.

Đóng góp: Cung cấp bằng chứng thực nghiệm mạnh mẽ về ưu việt của SVM trong dự đoán tài chính.

Kazem et al. (2013): SVR-CFA Hybrid Model

Mục tiêu: Phát triển mô hình kết hợp SVR với các thuật toán tối ưu hóa meta-heuristic.

Phương pháp tiếp cận:

- Kết hợp Support Vector Regression (SVR) với:
    - Chaotic mapping (ánh xạ hỗn loạn)
    - Firefly Algorithm (thuật toán đom đóm)
- So sánh với SVR-GA, SVR-CGA, SVR-FA, ANN, ANFIS

Kết quả: SVR-CFA cho hiệu suất tốt nhất trong các mô hình được thử nghiệm.

Đóng góp: Chứng minh hiệu quả của việc kết hợp SVR với các thuật toán tối ưu hóa sinh học.

Pai et al. (2010): Seasonal SVR (SSVR)

Mục tiêu: Xử lý tính chu kỳ và mùa vụ trong dữ liệu chuỗi thời gian tài chính.

Phương pháp tiếp cận:

- Phát triển mô hình Seasonal Support Vector Regression (SSVR)
- Tích hợp yếu tố thời vụ vào mô hình SVR truyền thống
- So sánh với SVR chuẩn và SARIMA

Kết quả: SSVR vượt trội hơn cả SVR và SARIMA trong dự báo dữ liệu theo mùa.

Đóng góp: Mở rộng khả năng ứng dụng của SVR cho dữ liệu có tính chu kỳ mạnh.

2.2.3. Nhóm nghiên cứu về Feature Engineering và Optimization

Huang và Wu (2008): GA-SVM với Time-scale Feature Extraction

Mục tiêu: Cải thiện hiệu suất dự đoán thông qua tối ưu hóa đặc trưng và tham số.

Phương pháp tiếp cận:

- Sử dụng Genetic Algorithm để trích xuất đặc trưng quy mô thời gian tối ưu
- Kết hợp với Support Vector Machine cho dự đoán
- So sánh với Neural Networks, SVM thuần túy và mô hình GARCH

Kết quả: Mô hình lai GA-SVM hoạt động tốt hơn các phương pháp truyền thống.

Đóng góp: Nhấn mạnh tầm quan trọng của việc lựa chọn đặc trưng trong dự đoán tài chính.

Aldin et al. (2012): Đánh giá hiệu quả Technical Indicators

Mục tiêu: Phân tích tác động của các chỉ báo kỹ thuật đến độ chính xác dự đoán.

Phương pháp tiếp cận:

- Đánh giá riêng lẻ các chỉ báo: Moving Average, RSI, CCI, MACD
- Phân tích tương quan và khả năng dự đoán của từng chỉ báo
- Áp dụng trên chỉ số TEPIX (Tehran Price Index)

Kết quả: Xác định được các chỉ báo có hiệu quả cao nhất cho dự đoán biến động giá.

Đóng góp: Cung cấp cơ sở lựa chọn đặc trưng cho các mô hình dự đoán.

2.2.4. Nhóm nghiên cứu về Ensemble Learning

Cheng et al. (2012): Ensemble Learning Algorithms

Mục tiêu: Cải thiện hiệu suất dự đoán thông qua kết hợp nhiều mô hình.

Phương pháp tiếp cận:

- Phát triển các thuật toán học tập hợp (ensemble learning)
- Kết hợp dự đoán từ nhiều mô hình cơ sở khác nhau
- Sử dụng các phương pháp voting và averaging

Kết quả: Ensemble methods cho hiệu suất ổn định và cao hơn các mô hình đơn lẻ.

Đóng góp: Khẳng định hiệu quả của phương pháp học tập hợp trong dự đoán tài chính.

2.3. Phân tích so sánh các phương pháp tiếp cận

2.3.1. Bảng so sánh tổng quan

Nghiên cứu	Phương pháp chính	Ưu điểm	Hạn chế	Độ chính xác
Zhang & Wu (2009)	IBCO-ANN	Tối ưu hóa trọng số hiệu quả	Phức tạp tính toán	Cao
Asadi et al. (2012)	GA-LM-ANN	Kết hợp tiền xử lý và tối ưu	Thời gian huấn luyện lâu	Rất cao
Ou & Wang (2009)	SVM/LS-SVM	Hiệu suất ổn định	Phụ thuộc tham số	Cao
Kazem et al. (2013)	SVR-CFA	Kết hợp đa thuật toán	Độ phức tạp cao	Rất cao
Huang & Wu (2008)	GA-SVM	Tối ưu hóa đặc trưng	Tốn thời gian tính toán	Cao

2.3.2. Phân loại theo cách tiếp cận

Phương pháp đơn lẻ (Single Method):

- Ưu điểm: Đơn giản, dễ cài đặt, thời gian tính toán ngắn
- Nhược điểm: Hiệu suất hạn chế, khó xử lý dữ liệu phức tạp

Phương pháp kết hợp (Hybrid Method):

- Ưu điểm: Hiệu suất cao, tận dụng ưu điểm của nhiều thuật toán
- Nhược điểm: Phức tạp, khó điều chỉnh tham số, thời gian tính toán lâu

Phương pháp tối ưu hóa (Optimization-based):

- Ưu điểm: Tìm được tham số tối ưu, hiệu suất ổn định
- Nhược điểm: Dễ rơi vào tối ưu cục bộ, phụ thuộc vào khởi tạo

2.4. Kết quả đạt được, hạn chế và khả năng kế thừa

2.4.1. Kết quả đạt được trong các nghiên cứu liên quan

Về mặt độ chính xác:

- Các phương pháp được đề xuất đã chứng minh khả năng dự đoán chỉ số chứng khoán với độ chính xác cao
- Các mô hình kết hợp thường cho hiệu suất tốt hơn mô hình đơn lẻ
- Việc tối ưu hóa tham số mang lại cải thiện đáng kể

Về mặt khả năng thích ứng:

- Các mô hình cho thấy khả năng đối phó tốt với biến động thị trường
- Hiệu suất ổn định trong các điều kiện thị trường khác nhau
- Khả năng xử lý noise và outliers được cải thiện

Về mặt thời gian dự đoán:

- Hầu hết các nghiên cứu tập trung vào dự đoán ngắn hạn (1-15 ngày)
- Một số nghiên cứu mở rộng cho dự đoán trung hạn
- Hiệu suất giảm dần khi tăng khoảng thời gian dự đoán

2.4.2. Hạn chế của các nghiên cứu hiện có

Hạn chế về phương pháp luận:

- **Vấn đề chính**: Hầu hết các phương pháp hiện có chỉ sử dụng **một lớp dự đoán duy nhất**
- **Hệ quả**: Dẫn đến giảm độ chính xác khi dự đoán cho thời gian xa hơn (khi n tăng)
- **Nguyên nhân**: Dự đoán dựa trên các giá trị ngày càng cũ của tham số thống kê

Hạn chế về kỹ thuật kết hợp:

- Một số nghiên cứu đã cố gắng kết hợp các kỹ thuật học máy
- Tuy nhiên, **không nhằm mục đích thu hẹp khoảng cách thời gian**
- Thường chỉ sử dụng một kỹ thuật để **điều chỉnh tham số thiết kế** của kỹ thuật khác
- Chưa có cách tiếp cận hệ thống để giải quyết vấn đề khoảng cách thời gian

Hạn chế về dữ liệu và đặc trưng:

- Phụ thuộc nhiều vào lựa chọn đặc trưng đầu vào
- Chưa có phương pháp tự động cập nhật đặc trưng theo thời gian
- Khó xử lý dữ liệu thiếu hoặc nhiễu

Hạn chế về đánh giá:

- Thiếu chuẩn mực đánh giá thống nhất
- Ít nghiên cứu so sánh trực tiếp trên cùng bộ dữ liệu
- Chưa có đánh giá chi tiết về tính robust của các mô hình

2.4.3. Khả năng kế thừa và phát triển

Khả năng tổng quát hóa:

- Cách tiếp cận hai giai đoạn được đề xuất có tiềm năng **tổng quát hóa cao**
- Có thể áp dụng cho các nhiệm vụ dự báo khác ngoài tài chính:

- **Dự báo thời tiết**: Dự đoán các thông số khí tượng trong tương lai
- **Dự báo tiêu thụ năng lượng**: Dự đoán nhu cầu điện, gas, nhiên liệu
- **Dự báo GDP**: Dự đoán các chỉ số kinh tế vĩ mô
- **Dự báo y tế**: Dự đoán xu hướng dịch bệnh, nhu cầu y tế

Khả năng mở rộng kỹ thuật:

- Có thể tích hợp thêm các thuật toán học máy mới
- Khả năng kết hợp với deep learning và AI hiện đại
- Tiềm năng ứng dụng big data và real-time processing

Khả năng cải tiến:

- **Cải tiến về thuật toán**: Sử dụng các thuật toán tối ưu hóa tiên tiến hơn
- **Cải tiến về đặc trưng**: Tích hợp thêm dữ liệu phi cấu trúc (tin tức, mạng xã hội)
- **Cải tiến về kiến trúc**: Phát triển thành mô hình đa giai đoạn (multi-stage)

2.5. Định hướng nghiên cứu và khoảng trống cần lấp đầy

2.5.1. Khoảng trống nghiên cứu được xác định

Dựa trên phân tích các nghiên cứu liên quan, có thể xác định được những **khoảng trống nghiên cứu** chính cần được giải quyết:

1. **Vấn đề khoảng cách thời gian**: Chưa có nghiên cứu nào giải quyết hiệu quả vấn đề giảm độ chính xác khi tăng khoảng thời gian dự đoán
2. **Thiếu phương pháp kết hợp hệ thống**: Các nghiên cứu kết hợp hiện tại chưa có cách tiếp cận hệ thống để tối ưu hóa toàn bộ quy trình dự đoán
3. **Chưa có đánh giá toàn diện**: Thiếu nghiên cứu so sánh toàn diện các phương pháp trên cùng bộ dữ liệu với nhiều khoảng thời gian dự đoán khác nhau

2.5.2. Định hướng giải pháp

Đề tài này định hướng giải quyết các khoảng trống trên thông qua:

- **Phương pháp hai giai đoạn**: Giải quyết trực tiếp vấn đề khoảng cách thời gian
- **Kết hợp hệ thống**: Tối ưu hóa cả giai đoạn chuẩn bị dữ liệu và dự đoán cuối cùng
- **Đánh giá toàn diện**: So sánh chi tiết các phương pháp một và hai giai đoạn

---

CHƯƠNG III. MÔ HÌNH/THUẬT TOÁN ĐỀ XUẤT

1. **TIẾP CẬN MỘT GIAI ĐOẠN (Single Stage Approach)**

◦ **Ý tưởng cơ bản:** Đối với nhiệm vụ dự đoán trước 'n' ngày, đầu vào là mười chỉ báo kỹ thuật mô tả ngày 't', và đầu ra là giá đóng cửa của ngày '(t+n)'.

◦ **1.1 Các chỉ báo kỹ thuật (Technical Indicators).**

▪ **Simple Moving Average (SMA)**: Đường trung bình động đơn giản (10 ngày). Công thức: (Tổng giá đóng cửa của 'n' ngày trước đó) / 'n'.

▪ **Exponential Moving Average (EMA)**: Đường trung bình động hàm mũ (10 ngày). Giúp xác định trọng số của dữ liệu gần nhất, làm đường đi chuẩn xác hơn SMA. Nguyên tắc sử dụng cho ngắn/trung/dài hạn.

▪ **Momentum (MOM)**: Chỉ báo đo lường tốc độ thay đổi của giá chứng khoán. Mức chỉ báo > 100 (giá hiện tại cao hơn giá 'n' phiên trước), < 100 (giá hiện tại thấp hơn giá 'n' phiên trước).

▪ **Stochastic (STC)**: Chỉ số cơ bản cho khuynh hướng thị trường, giới hạn từ 0-100, phản ánh vùng quá bán (oversold) và quá mua (overbought). Bao gồm Đường Nhanh %K và Đường Chậm %D.

▪ **Relative Strength Index (RSI)**: Chỉ số Sức mạnh Tương đối, đo lường thay đổi giá trong 14 giai đoạn mặc định, biểu diễn trên thang điểm 0-100. Sử dụng để dự đoán xu hướng đảo chiều, mức hỗ trợ/kháng cự thông qua phân kỳ dương/âm.

▪ **Moving Average Convergence Divergence (MACD)**: Phân kỳ hội tụ đường trung bình. Thể hiện tín hiệu mua/bán cổ phiếu, xác định độ mạnh xu hướng, và đánh giá tình trạng quá mua/quá bán.

▪ **Larry Williams %R (WR)**: Chỉ báo xung lượng để đo mức quá mua/quá bán, dao động từ 0% đến -100% với 3 vùng tín hiệu.

▪ **Accumulation/Distribution Oscillator (ADO)**: Đường tích lũy/phân phối. Xác định xu hướng và phát hiện dấu hiệu phân kỳ (suy yếu của xu hướng tăng/giảm).

▪ **Commodity Channel Index (CCI)**: Chỉ số kênh hàng hóa. Giống Momentum, dao động quanh trục 0 (nhưng trong khoảng -100 đến 100). CCI cao/tăng báo hiệu xu hướng tăng; CCI thấp/giảm báo hiệu xu hướng giảm.

◦ **1.2 Artificial Neural Networks (ANN)**

▪ Mô hình xử lý thông tin mô phỏng hệ thống thần kinh sinh vật, học hỏi từ kinh nghiệm và dự đoán dữ liệu chưa biết.

▪ Cấu trúc gồm 3 thành phần chính: **Input layer, Output layer và Hidden layer** (có thể có nhiều layer).

▪ Sử dụng **truyền ngược ba lớp nạp về phía trước ANN**. Lớp đầu vào có 10 nơ-ron (một cho mỗi tham số kỹ thuật), lớp đầu ra có một nơ-ron biểu thị giá trị chỉ số dự đoán.

▪ **Thuật toán cập nhật trọng số:** Giảm dần độ dốc thích ứng (Adaptive Gradient Descent).

▪ Chức năng truyền: Sigmoid tiếp tuyến cho lớp ẩn, tuyến tính cho lớp đầu ra.

▪ Tham số thiết kế: Số lượng nơ-ron trong lớp ẩn (n=1,2,3,4) và số chu kỳ (epochs=1000, 2000,...10000).

◦ **1.3 Support Vector Regression (SVR)**

▪ Mô hình hồi quy sử dụng thuật toán Support Vector Machine (SVM) để dự đoán giá trị biến liên tục.

▪ Đặt biên dung sai ε, lỗi được coi là 0 đến ngưỡng ε. Ý tưởng chính là giảm thiểu lỗi bằng cách cá nhân hóa siêu mặt phẳng tối đa hóa lợi nhuận.

▪ Xây dựng hàm hồi quy tuyến tính: $f(x, w) = \ *x +$.

▪ Tối thiểu hóa rủi ro R bằng cách giảm thiểu mô phỏng >||>2 và tổng của tổn thất độ nhạy ε tuyến tính.

▪ Sử dụng lý thuyết Lagrangian và điều kiện Karush – Kuhn – Tucker để giải quyết bài toán tối ưu hóa.

▪ Có thể ánh xạ các vectơ đầu vào vào không gian đặc trưng chiều cao sử dụng hàm nhân (kernel function). Các hàm nhân cơ sở đa thức (Polynomial Function) và xuyên tâm (Radial Basis Function - RBF) được sử dụng.

▪ Tham số thiết kế: Lựa chọn hàm nhân, bậc của hàm đa thức (d=1,2,3,4), gamma trong hàm nhân (γ=0, 0.5, ... 100) và hằng số chính quy (c=1).

◦ **1.4 Random Forest (RF)**

▪ Thuộc thể loại thuật toán học tập hợp (ensemble learning).

▪ Sử dụng cây quyết định (regression tree) làm người học cơ bản của nhóm.

▪ Lý do sử dụng học tổng hợp: một cây hồi quy đơn lẻ không đủ chính xác để xác định giá trị dự đoán của biến phụ thuộc do không phân biệt được nhiễu và mẫu.

▪ Thực hiện lấy mẫu có thay thế để học 'n' cây dựa trên các mẫu tập dữ liệu.

▪ Trong nghiên cứu này, mỗi cây được học bằng cách sử dụng 3 đặc điểm được chọn ngẫu nhiên.

▪ Giá trị dự đoán cuối cùng là giá trị trung bình của giá trị dự đoán từ mỗi cây trong tập hợp, giúp loại trừ vấn đề quá khớp (over-fitting).

▪ Tham số thiết kế: Số cây (n=50, 100, 150).

2. **TIẾP CẬN HAI GIAI ĐOẠN (Two Stage Fusion Approach)**

◦ **2.1 Mô tả hướng tiếp cận**

▪ **Ý tưởng cơ bản:** Giai đoạn đầu tiên sử dụng SVR để chuẩn bị đầu vào cho các mô hình dự đoán ở giai đoạn thứ hai.

▪ Đầu vào cho SVR trong giai đoạn đầu tiên mô tả ngày 't', trong khi đầu ra mô tả ngày '(t+n)' dưới dạng mười chỉ số kỹ thuật.

▪ Những đầu ra này từ giai đoạn đầu tiên đóng vai trò là đầu vào cho các mô hình dự đoán trong giai đoạn thứ hai.

▪ Điều này giúp các mô hình dự đoán ở giai đoạn hai xác định chuyển đổi ánh xạ từ các thông số kỹ thuật mô tả ngày '(t+n)' sang giá đóng cửa của ngày '(t+n)', khác với cách tiếp cận một giai đoạn (từ ngày 't' đến ngày '(t+n)').

▪ ANN, SVR và RF được sử dụng làm mô hình dự đoán trong giai đoạn thứ hai, với số lượng thử nghiệm toàn diện bằng cách thay đổi giá trị tham số tương tự như phương pháp một giai đoạn.

▪ Để quyết định sự kết hợp tốt nhất của các giá trị tham số cho mỗi SVR trong giai đoạn đầu tiên, một phần dữ liệu (20% của toàn bộ tập dữ liệu, chia 80% huấn luyện, 20% thử nghiệm) được dùng để điều chỉnh tham số.

◦ **2.2 Giai đoạn 1 (Mười mô hình SVR)**

▪ Xác định sự kết hợp tốt nhất của các giá trị tham số cho mỗi SVR thông qua thử nghiệm điều chỉnh tham số trên tập huấn luyện và thử nghiệm.

▪ Kết quả cho thấy phép biến đổi không gian đầu vào thông qua **nhân RBF thực hiện tốt hơn nhân Polynomial**.

▪ Mục đích của các thử nghiệm điều chỉnh tham số là xác định tổ hợp tham số tốt nhất cho mỗi SVR ở giai đoạn đầu, nhằm giảm thiểu sai số trong các tham số thống kê được sử dụng làm đầu vào cho các mô hình dự đoán ở giai đoạn thứ hai.

▪ Bảng kết hợp tham số tốt nhất cho SVRs giai đoạn đầu (ví dụ: SVR-1 RBF γ=2, SVR-4 RBF γ=100, v.v.).

▪ **Cải tiến của nhóm:** Sử dụng thêm SVR với `kernel="poly"` thay đổi so với tác giả chỉ sử dụng `kernel="rbf"`. Lý do là dữ liệu của nhóm có những chỉ báo có dạng tuyến tính, nên việc thay đổi tham số bên trong mô hình SVR có thể dẫn tới kết quả tốt hơn. Bảng chi tiết các thay đổi tham số của nhóm so với tác giả.

◦ **2.3 Giai đoạn 2**

▪ Từ 10 chỉ báo được dự báo ở giai đoạn 1, sử dụng các thuật toán SVR, RF, ANN để dự báo giá đóng cửa cho ngày '(t+n)'.

▪ Tạo ra các mô hình kết hợp: **SVR-SVR, SVR-RF, SVR-ANN**.

CHƯƠNG IV. CÀI ĐẶT THỰC NGHIỆM

1. **Dữ liệu (Dataset)**

◦ **Dữ liệu của tác giả:** Tổng dữ liệu lịch sử mười năm (tháng 01/2003 - tháng 12/2012) của hai chỉ số thị trường chứng khoán Ấn Độ: CNX Nifty và S&P BSE Sensex. Dữ liệu được lấy từ các

trang web `http://www.nseindia.com/` và `http://www.bseindia.com/`. Mười chỉ số kỹ thuật được tính toán từ giá đóng cửa, giá cao, giá thấp và giá mở cửa.

◦ **Bộ dữ liệu của nhóm:** Bộ dữ liệu NIFTY 50 lấy từ nguồn của tác giả (`https://www1.nseindia.com/products/content/equities/indices/historical_index_data.htm`).

▪ NIFTY 50 là chỉ số chuẩn của thị trường chứng khoán Ấn Độ, đại diện cho mức trung bình có trọng số của 50 công ty lớn nhất niêm yết trên Sở Giao dịch Chứng khoán Quốc gia.

▪ **Thời gian của bộ dữ liệu nhóm:** Từ 01/01/2006 đến 01/04/2021.

▪ **Thuộc tính của bộ dữ liệu:** Date, Open, High, Low, Close, Shares Traded (Volume), Turnover (Rs. Cr).

▪ **Kích thước bộ dữ liệu:** 3780 dòng và 7 cột (tương ứng 3780 mẫu).

▪ Nhóm đã tạo ra 10 bộ chỉ số kỹ thuật (SMA, EMA, MOM, STCK, STCD, MACD, RSI, WR, A/D, CCI) từ dữ liệu gốc.

◦ **Rủi ro dữ liệu:** Sử dụng bộ chỉ số và mốc thời gian khác với tác giả gốc.

2. **Công cụ**

◦ **Google Colab:** Môi trường phát triển để tạo Notebook, khởi động, cài đặt GPU miễn phí, bắt đầu code, và liên kết Google Drive.

◦ **Weka:** Công cụ để cài đặt và sử dụng Weka Explorer, cài đặt các package Timeseries.

◦ **Jupyter Notebook:** Được sử dụng cho các mô hình hai giai đoạn.

3. **Tiền xử lý dữ liệu**

◦ Import thư viện và dữ liệu.

◦ Đổi tên cột "Shares Traded" thành "Volume" và "Turnover" thành "Amount".

◦ Kiểm tra dữ liệu trống và kiểu dữ liệu.

◦ Trực quan hóa dữ liệu bằng đồ thị Heatmap và các thuộc tính khác (Giá đóng cửa, Khối lượng giao dịch).

◦ Install và sử dụng thư viện `StockDataframe` để cài đặt 10 chỉ số dự báo, với demo `n=10` ngày và thực nghiệm với `n=5` và `n=30` ngày.

4. **Cài đặt thuật toán**

◦ **a. Cài đặt 10 chỉ số dự báo:** Khởi tạo các chỉ số RSI, MACD, EMA, SMA, MOM, STCK, STCD, WR, A/D, CCI.

◦ **b. Mô hình Một giai đoạn:** Sử dụng Weka Explorer.

▪ **Áp dụng Thuật toán SVR:** Kernel "RBF", Target Selection là "Close", thời gian dự đoán 10 ngày. Chọn độ đo MAE, MSE, RMSE, MAPE. Chia dữ liệu train-test 80-20. Ví dụ với `gamma = 0.01`.

▪ **Áp dụng Thuật Toán Random Forest:** Ví dụ với `n=150` cây.

▪ **Áp dụng Thuật Toán ANN:** Ví dụ với `epochs=6000`.

◦ **c. Mô hình Hai giai đoạn:** Sử dụng Jupyter Notebook.

▪ Xét 3 mô hình SVR để tìm mô hình tối ưu nhất dự báo cho từng chỉ số.

▪ Đồ thị so sánh giữa 3 mô hình.

▪ Kết quả dự đoán của 3 mô hình với kết quả thực tế.

▪ Mô hình sử dụng **Polynomial SVR mang lại kết quả tối ưu nhất đối với dữ liệu có dạng tuyến tính**.

5. **Kết quả đạt được**

◦ **a. Dự đoán 10 ngày:** Bảng so sánh kết quả (MAE, MAPE, RMSE, MSE) của ANN, SVR-ANN, SVR (rbf), SVR-SVR, RF, SVR-RF cho cả tác giả và nhóm thực hiện.

◦ **b. Kết quả sau khi cải tiến:** Bảng kết quả (MAE, MAPE, RMSE, MSE) cho `n=5` ngày và `n=30` ngày với các tham số cụ thể.

◦ **c. So sánh các thuật toán:**

▪ **Hiệu suất dự đoán trung bình:** Bảng tổng hợp hiệu suất trung bình (MAE, MAPE, RMSE, MSE) của các mô hình một giai đoạn (ANN, SVR, RF) và hai giai đoạn (SVR-ANN, SVR-SVR, SVR-RF).

▪ **Hiệu suất cải thiện của Mô hình 2 giai đoạn so với 1 giai đoạn:** Bảng phần trăm cải thiện (MAE, MAPE, RMSE, MSE) cho các cặp thuật toán (ANN vs SVR-ANN, SVR vs SVR-SVR, RF vs SVR-RF).

▪ **Phân tích kết quả:**

• Mô hình của nhóm cho kết quả hiệu suất trung bình có một số độ đo tốt hơn tác giả (MAE cho SVR-ANN, SVR-RF; MAPE cho ANN, SVR; MSE cho SVR-ANN, SVR-RF).

• Kết quả các chỉ số của nhóm phần lớn tốt hơn cho Giai đoạn 2 so với Giai đoạn 1, trừ độ đo MAPE mà Giai đoạn 1 hiệu quả hơn. Thuật toán SVR giai đoạn 1 đem lại hiệu suất tốt hơn SVR-SVR giai đoạn 2.

• Nhóm có các thuật toán có độ đo tốt hơn hẳn so với tác giả ở ANN vs. SVR-ANN và RF vs. SVR-RF.

• Nhìn chung, **mô hình SVR – ANN đạt được hiệu suất dự đoán tổng thể tốt nhất**.

• Hiệu suất của SVR–ANN và SVR–RF cải thiện đáng kể so với ANN và SVR. SVR–SVR thể hiện cải thiện vừa phải so với SVR.

• Lợi ích của mô hình hai giai đoạn trở nên rõ ràng khi dự đoán cho nhiều ngày hơn.

• Việc giới thiệu giai đoạn bổ sung trong cách tiếp cận hai giai đoạn chịu trách nhiệm chuẩn bị dữ liệu cho giai đoạn hai, giúp chuyển đổi giá đóng cửa và giá mở cửa, thấp, cao của ngày 't' thành các thông số kỹ thuật đại diện cho ngày '(t+n)', từ đó giảm sai số dự đoán.

CHƯƠNG V. KẾT LUẬN VÀ HƯỚNG CẢI TIẾN

1. KẾT LUẬN

◦ Nhiệm vụ dự đoán giá trị tương lai của chỉ số thị trường chứng khoán được tập trung nghiên cứu.

◦ Các phương pháp hiện có chỉ sử dụng một lớp dự đoán, dẫn đến giảm độ chính xác khi dự đoán cho thời gian xa hơn.

◦ Bài báo đề xuất cách tiếp cận tổng hợp hai giai đoạn sử dụng SVR ở giai đoạn đầu và ANN, RF, SVR ở giai đoạn thứ hai để giải quyết vấn đề đã xác định.

◦ Kết quả thực nghiệm cho thấy mô hình lai hai giai đoạn hoạt động tốt hơn mô hình dự đoán giai đoạn đơn, với cải thiện đáng kể cho SVR-ANN và SVR-RF, và cải thiện vừa phải cho SVR-SVR.

◦ Mô hình SVR-ANN đạt hiệu suất dự đoán tổng thể tốt nhất.

◦ Đề xuất sơ đồ dự báo hai giai đoạn là một đóng góp nghiên cứu đáng kể vì nó cung cấp một phương pháp mới để cung cấp thông tin đầy đủ cho các mô hình dự báo, có thể tổng quát hóa cho các nhiệm vụ dự báo khác.

2. HƯỚNG CẢI TIẾN

◦ **Nội dung 1: Phương pháp tiếp cận một giai đoạn**.

▪ Mô tả: Từ 10 chỉ báo ngày 't', dự báo giá đóng cửa ngày '(t+n)'.

▪ ANN, SVR, RF đã được sử dụng.

◦ **Nội dung 2: Phương pháp tiếp cận hai giai đoạn**.

▪ Mô tả: Từ 10 chỉ báo và giá đóng cửa ngày 't', dự báo ra 10 chỉ báo ngày '(t+n)'. Sau đó từ 10 chỉ báo ngày '(t+n)', dự báo ra giá đóng cửa ngày '(t+n)'.

▪ Mười mô hình SVR cho giai đoạn 1 (mỗi mô hình dự báo 1 chỉ báo làm đầu vào cho giai đoạn 2).

▪ **Đề xuất cải tiến:** Sử dụng thêm SVR với `kernel="poly"` thay vì chỉ `kernel="rbf"` của tác giả, vì dữ liệu có chỉ báo dạng tuyến tính, có thể mang lại kết quả tốt hơn. Bảng thay đổi tham số của nhóm so với tác giả.

▪ Ba mô hình dự đoán cho giai đoạn hai tạo ra các mô hình kết hợp SVR-SVR, SVR-RF, SVR-ANN.

◦ **Hướng nghiên cứu tương lai:**

▪ Có thể khám phá các thuật toán như thuật toán di truyền để điều chỉnh các tham số thiết kế của SVRs trong giai đoạn đầu tiên, có thể dẫn đến dự đoán chính xác hơn.

▪ Sử dụng nhiều tham số thống kê hơn làm đầu vào để tìm ra mối tương quan tốt hơn.

▪ Kết hợp các tin tức liên quan đến chính sách chính phủ, hiệu suất công ty, sự quan tâm của nhà đầu tư (phân loại thành 'tốt', 'rất tốt', 'xấu', 'tệ hơn') vào hệ thống. Một hệ thống bán giám sát như vậy có thể làm cho hệ thống mạnh mẽ hơn và dự đoán chính xác hơn.