Dự đoán xu thế chỉ số chứng khoán Việt Nam sử dụng phân tích hồi quy quá trình Gauss và mô hình tự hồi quy trung bình động

Huỳnh Quyết Thắng, Phùng Đình Vũ, Tống Văn Vinh
Trường Đại học Bách khoa Hà Nội

Tác giả liên hệ: Huỳnh Quyết Thắng, thanghq@soict.hust.edu.vn
Ngày nhận bài: 28/08/2017, ngày sửa chữa: 26/10/2018, ngày duyệt đăng: 01/11/2018
Xem sớm trực tuyến: 08/11/2018, định danh DOI: 10.32913/rd-ict.vol1.no39.571
Biên tập lĩnh vực điều phối phản biện và quyết định nhận đăng: TS. Trịnh Quốc Anh

Tóm tắt: Trong bài báo, chúng tôi trình bày phương pháp dự đoán xu thế chỉ số chứng khoán Việt Nam (VN-Index) gồm bốn bước, trong đó dữ liệu đầu vào là chuỗi thời gian chứa lịch sử chỉ số giá của VN-Index. Các tác giả thực hiện phân tách dữ liệu đầu vào thành các chuỗi thời gian thành phần bao gồm: xu thế, thời vụ và ngẫu nhiên. Chúng tôi áp dụng mô hình tự hồi quy trung bình động (ARMA: Autoregressive moving average) để dự đoán thành phần thời gian ngẫu nhiên ở một bước kế tiếp, phân tích hồi quy quá trình Gauss (GPR: Gaussian process regression) để dự đoán thành phần thời gian xu thế. Cuối cùng, kết quả dự đoán các thành phần riêng lẻ được tổng hợp lại để đưa ra kết quả dự đoán cuối cùng cho phương pháp kết hợp GPR-ARMA. Trong bài báo cũng trình bày các kết quả cài đặt thử nghiệm và phân tích hiệu quả của phương pháp được đề xuất.

Từ khóa: *Dự đoán xu thế VN-Index; Mô hình chuỗi thời gian; Hồi quy Gauss; Mô hình tự hồi quy trung bình động; Phương pháp kết hợp hồi quy Gauss và mô hình tự hồi quy trung bình động.*

Title: **Vietnam Stock Index Trend Prediction using Gaussian Process Regression and Autoregressive Moving Average Model**

Abstract: In this paper, we present a four-step method to predict the trend of Vietnam Stock Index (VN-Index). The input of the method is a time series which contains price history of VN-Index over the years. We decompose VN-Index price history into three time-series components: trend, seasonal and random. The autoregressive moving average model is used to predict one step ahead for the random component. We apply first difference of the trend series and use Gaussian process regression to predict one step ahead for the trend component. Finally, the predicted results of all component are summed to produce the predicted result of the input series. Performance of the proposed method is also evaluated and presented.

Keywords: *VN-Index trend prediction; Time series model, Gaussian process regression, autoregressive moving average model.*

I. GIỚI THIỆU BÀI TOÁN VÀ TỔNG HỢP CÁC KẾT QUẢ NGHIÊN CỨU LIÊN QUAN

Chỉ số chứng khoán Việt Nam (VN-Index) là chỉ số thể hiện xu hướng biến động giá của tất cả các cổ phiếu niêm yết tại sàn Giao dịch Chứng khoán Thành phố Hồ Chí Minh. Ở tầm vĩ mô, chỉ số này phản ảnh các quy luật cung cầu của thị trường chứng khoán (TTCK) và thường được sử dụng để đánh giá sự phát triển của nền kinh tế Việt Nam. Do đó, việc dự đoán đúng xu thế chỉ số VN-Index sẽ mang lại kết quả tốt cho nhà đầu tư khi tham gia vào thị trường. Phương pháp phân tích định lượng được sử dụng rộng rãi để giải quyết bài toán dự đoán biến động chỉ số chứng khoán. Có rất nhiều các mô hình định lượng khác nhau được áp dụng để giải quyết bài toán này như: phân tích hồi quy quá trình Gauss (GPR: Gaussian process regression) [1–3]; mô hình tự hồi quy trung bình động (ARMA: Autoregressive moving average) [4–6]; mạng nơ-ron nhân tạo [7]; mô hình mạng Bayes [8]; mô hình máy vector hỗ trợ [9].

Các tác giả trong [7] dự đoán giá đóng cửa hàng tuần của chỉ số chứng khoán Bombay TTCK Ấn Độ (BSE SENSEX) sử dụng mạng nơ-ron truyền thẳng nhiều lớp với việc điều chỉnh các trọng số thông qua thuật toán lan truyền ngược sai số. Mô hình mạng có một lớp đầu vào với 800 nơ-ron sử dụng hàm chuyển đổi Tan Sigmoid; ba lớp hàm ẩn tuyến

tính với 600 nơ-ron mỗi lớp và một lớp đầu ra có 1 nơ-ron. Dữ liệu dùng để huấn luyện các trọng số trên mạng nơ-ron có độ dài 200 tuần, bao gồm giá đóng cửa hàng tuần của chỉ số BSE SENSEX; sự di chuyển giá trung bình trong 52 tuần giao dịch; sự di chuyển giá trung bình trong 5 tuần giao dịch; sự biến động giá trong 5 tuần giao dịch; dao động giá trong 10 tuần giao dịch. Kết quả cho thấy căn bậc hai sai số toàn phương trung bình (RMSE: Root mean square error) theo phương pháp này là 4.82% và sai số tuyệt đối trung bình (MAE: Mean absolute error) là 3.93%.

Trong phương pháp sử dụng mạng Bayes, các tác giả trong [8] xây dựng mô hình nhân quả thể hiện sự phụ thuộc của xu thế tăng, giảm của chỉ số chứng khoán FTSE100 ở ngày kế tiếp vào xu thế tăng, giảm của chỉ số đó trong quá khứ, đồng thời trong mối tương quan với chỉ số Dow30 và chỉ số Nikkei225. Xác suất có điều kiện trên mỗi nút của mạng được tính toán dựa trên giải thuật K2 với bộ dữ liệu huấn luyện đầu vào từ tháng 1 năm 2005 đến tháng 12 năm 2006. Các tác giả tiến hành dự đoán cho các ngày giao dịch từ tháng 1 năm 2007 đến tháng 12 năm 2007. Kết quả cho thấy phương pháp này có độ chính xác dự đoán xu thế là 61.4%.

Mô hình máy vector hỗ trợ (SVM: Support vector machine) được giới thiệu trong bài báo [9] để dự đoán xu thế cho chỉ số chứng khoán của 13 công ty khác nhau từ năm 2004-2015. SVM được sử dụng như một công cụ để phân loại giữa hai lớp là lớp tăng và lớp giảm bằng cách học một siêu phẳng để phân lớp dữ liệu, và dựa vào dữ liệu lịch sử để dự đoán chỉ số chứng khoán của năm tiếp theo của một công ty thuộc lớp tăng hay lớp giảm. Kết quả cho thấy các tác giả dự đoán đúng cho xu thế của 10 trên 13 công ty trong năm 2014-2015.

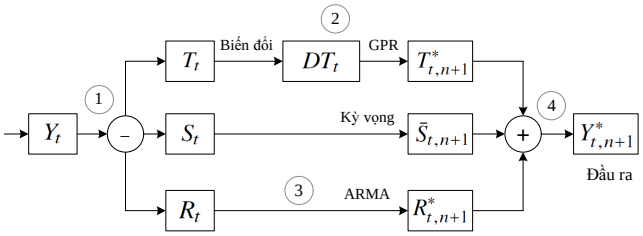
Trong phương pháp hồi quy [2, 3, 10, 11], người ta thường xây dựng mô hình dự báo theo cách tiếp cận kinh tế lượng, sử dụng một số biến kinh tế vĩ mô và biến tài chính tiền tệ mà theo lý thuyết kinh tế có tác động đến biến động thị trường chứng khoán làm biến giải thích trong mô hình hồi quy đa biến.

Phương pháp GPR được trình bày chi tiết trong mục II-2 của bài báo này. Về cơ bản, đây là phương pháp được sử dụng trong học máy nhằm tìm kiếm các mẫu hình lặp lại trong dữ liệu chuỗi thời gian, qua đó thực hiện dự đoán xu thế tiếp theo của các điểm trong chuỗi thời gian. Các tác giả trong bài báo [3] thực nghiệm quá trình Gauss để dự đoán xu thế về giá đóng cửa của các cổ phiếu riêng lẻ theo một số lớp khác nhau các hàm hiệp phương sai như hàm hiệp phương sai lũy thừa bình phương, hàm hiệp phương sai lớp Matern, hàm hữu tỷ bậc hai. Dựa vào đánh giá thực nghiệm, các tác giả khẳng định rằng dữ liệu lịch sử càng dài cho kết quả dự đoán càng chính xác để tìm ra cổ phiếu tốt, và việc sử dụng hàm hiệp phương sai lũy thừa bình phương và hàm hiệp phương sai lớp Matern cho kết quả dự đoán xu thế tốt.

Các tác giả trong [6] sử dụng mô hình tự hồi quy kết hợp trung bình động (ARIMA: Autoregressive integrated moving verage) để dự đoán giá cổ phiếu của 66 công ty từ bảy ngành khác nhau dựa trên bộ dữ liệu lịch sử giá của cổ phiếu các công ty với độ dài khoảng 23 tháng và tiến hành dự đoán cho một tháng kế tiếp. Để đánh giá các tham số cho mô hình các tác giả sử dụng bộ tham số sao cho tiêu chuẩn thông tin Akaike đạt giá trị nhỏ nhất. Chi tiết về mô hình tự hồi quy trung bình động được trình bày trong mục II-3 của bài báo này. Kết quả dự đoán các tác giả thu được có giá trị sai số phần trăm trung bình lớn hơn 85% trong tất cả các trường hợp. Các giả cũng đánh giá đây là hướng tiếp cận khả quan nhất trong dự đoán giá cổ phiếu [2, 3, 6].

Tại Việt Nam hiện có một số nghiên cứu liên quan đến dự báo chỉ số chứng khoán VN-Index [12–14]. Trong [12], các tác giả đề xuất kết hợp phương pháp chỉ số dẫn báo và hệ số tương quan giữa chỉ số thị trường chứng khoán của một sàn giao dịch với các biến dữ liệu giao dịch cổ phiếu trong việc xây dựng mô hình dự báo chỉ số thị trường chứng khoán trên dữ liệu. Tác giả thu thập từ dữ liệu sàn giao dịch Thành phố Hồ Chí Minh: dữ liệu từ 04/01/2010 đến 22/04/2016 được sử dụng để xây dựng mô hình dự báo, dữ liệu kiểm định là từ 25/04/2016 đến ngày 05/05/2016 (gồm 7 ngày giao dịch do các ngày từ 30/04/2016 đến 03/05/2016 là những ngày nghỉ lễ, sàn giao dịch không làm việc). Trong [13, 14], các tác giả áp dụng mô hình tự hồi quy phương sai không đồng nhất tổng quát (GARCH: Generalized autoregressive conditional heteroskedasticity). Mẫu dữ liệu bao gồm hai chỉ số của sàn giao dịch chứng khoán Việt Nam là chỉ số VN-Index và HNX-Index, được cung cấp bởi Sở Giao dịch Chứng khoán Thành phố Hồ Chí Minh (HOSE) và Sở Giao dịch Chứng khoàn Hà Nội (HNX), tương ứng, trong giai đoạn 2007-2015. Kết quả thực nghiệm cho mô hình GARCH, các tác giả khẳng định rằng biến động của các chỉ số chứng khoán trong quá khứ sẽ ảnh hưởng đến biến động trong hiện tại và có thể dự đoán trước, đồng thời cho thấy rằng Mô hình FIAPARCH là mô hình phù hợp nhất cho việc dự báo chỉ số VN-Index và HNX-Index.

Mỗi mô hình ở trên có những ưu điểm và nhược điểm riêng và được áp dụng cho các bộ dữ liệu cụ thể. Trong khuôn khổ bài báo này, chúng tôi tập trung nghiên cứu các mô hình áp dụng trên bộ dữ liệu chuỗi thời gian, đó là GPR và mô hình ARMA. Chúng tôi kế thừa kết quả các phương pháp đã được nghiên cứu trong bài báo [3, 6, 10] bằng cách đề xuất một giải pháp kết hợp mô hình GPR và mô hình ARMA, gọi là GPR-ARMA. Phương pháp kết hợp GPR-ARMA được áp dụng để dự đoán xu thế chỉ số VN-Index dựa trên bộ dữ liệu lịch sử giá đóng cửa chỉ số VN-Index qua các ngày giao dịch.



Hình 1. Phương pháp dự đoán kết hợp GPR-ARMA.

Bố cục tiếp theo của bài báo được trình bày như sau. Mục II trình bày giải pháp đề xuất, mục III trình bày thử nghiệm thực tế đã cài đặt và mục IV là kết luận và hướng nghiên cứu tiếp theo.

II. PHƯƠNG PHÁP KẾT HỢP GPR-ARMA

Hình 1 mô tả tổng quan quá trình gồm bốn bước thực hiện của phương pháp kết hợp GPR-ARMA để dự đoán xu thế chỉ số VN-Index. Đầu vào của phương pháp là một chuỗi thời gian gọi là Y_t.

Bước 1: Phân tách chuỗi thời gian đầu vào Y_t thành các chuỗi thời gian thành phần, bao gồm: chuỗi xu thế (gọi là T_t), chuỗi thời vụ (gọi là S_t), và chuỗi ngẫu nhiên (gọi là R_t). Chuỗi thời gian Y_t được tổng hợp lại theo phương pháp nhân sử dụng công thức nhân [10, 11] sau đây:

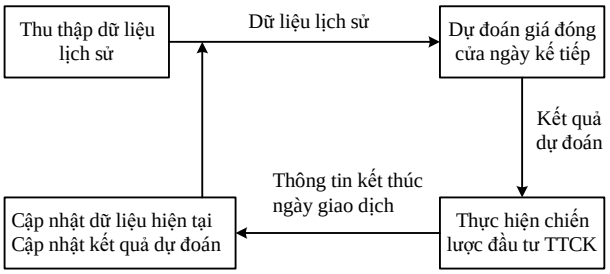
$$Y_t = T_t * S_t * R_t. \tag{1}$$

Bước 2: Áp dụng GPR để dự đoán chuỗi thời gian xu thế T_t. Trước tiên ta biến đổi chuỗi T_t bằng lấy sai phân bậc một của chuỗi xu thế đầu vào T_t để thu được chuỗi xu thế biến đổi DT_t. Việc biến đổi này đảm bảo tính dừng của chuỗi DT_t, là một trong những điều kiện đầu vào của phương pháp. Chuỗi DT_t sẽ là đầu vào cho phương pháp dự đoán theo GPR. Gọi $T^*_{t,n+1}$ là kết quả dự đoán chuỗi xu thế T_t tương ứng tại một thời điểm kế tiếp.

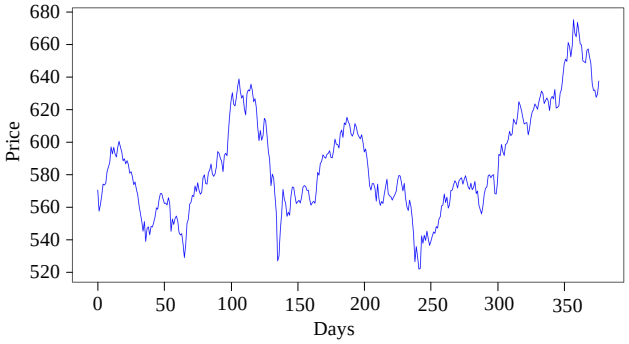
Bước 3: Áp dụng mô hình ARMA để dự đoán chuỗi thời gian ngẫu nhiên R_t. Ta sẽ chỉ ra sau đây rằng chuỗi R_t có tính dừng nên R_t có thể là đầu vào trực tiếp cho phương pháp ARMA. Gọi $R^*_{t,n+1}$ là giá trị dự đoán tại một điểm kế tiếp cho chuỗi R_t theo mô hình ARMA.

Bước 4: Tổng hợp kết quả dự đoán từ bước 2 và bước 3. Để có được kết quả dự đoán cho chuỗi Y_t, ngoài việc dự đoán cho chuỗi T_t và R_t ta phải biết được giá trị chuỗi thời vụ S_t. Do S_t thể hiện tính lặp lại của các giá trị trong một chu kỳ, nên ta hoàn toàn tính được giá trị tương ứng trong chu kỳ của S_t tại điểm đang dự đoán, gọi giá trị này là $\overline{S}_{t,n+1}$. $Y^*_{t,n+1}$ là kết quả dự đoán tại một thời điểm kế tiếp cho chuỗi thời gian đầu vào được cho bởi công thức nhân sau:

$$Y^*_{t,n+1} = T^*_{t,n+1} * R^*_{t,n+1} * \overline{S}_{t,n+1}. \tag{2}$$



Hình 2. Quy trình thực hiện phương pháp GPR-ARMA.



Hình 3. Lịch sử giá đóng cửa của chỉ số VN-Index.

Sau khi đã dự đoán tại một điểm kế tiếp, ta bổ sung giá trị quan sát thực tế tại điểm đã được dự đoán này vào tập huấn luyện và lặp lại các bước từ bước 1 đến bước 4 ở trên cho bộ dữ liệu đầu vào mới được bổ sung này để dự đoán cho điểm kế tiếp tiếp theo trong tập kiểm thử. Hình 2 mô tả quy trình thực hiện phương pháp dự đoán GPR-ARMA để dự đoán xu thế giá đóng cửa chỉ số VN-Index cho một ngày kế tiếp.

1. Phân tích dữ liệu đầu vào

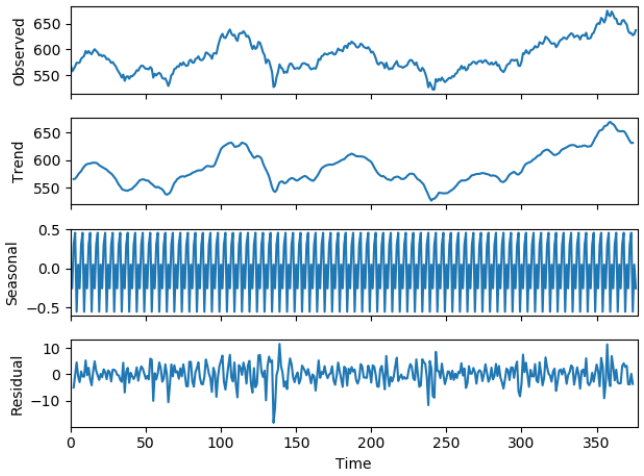
Bộ dữ liệu đầu vào là một chuỗi thời gian với các giá trị là giá đóng cửa của chỉ số VN-Index qua các ngày giao dịch. Hình 3 thể hiện biểu đồ lịch sử giá của chỉ số VN-Index từ ngày 02/02/2015 đến ngày 09/08/2016, tương ứng với 377 ngày giao dịch, được chúng tôi sử dụng là đầu vào cho phương pháp dự đoán GPR-ARMA.

Dữ liệu lịch sử giá của chỉ số VN-Index được phân tách thành ba chuỗi thành phần: xu thế, thời vụ và ngẫu nhiên. Chuỗi xu thế được tính theo phương pháp trung bình động từ một phía theo công thức sau:

$$X_{T_i} = \frac{X_{i-4} + X_{i-3} + X_{i-2} + X_{i-1} + X_i}{\sum_{j=i-4}^{i}(1 - \delta(X_j, 0))}, \tag{3}$$

trong đó, $\delta(x, x')$ là hàm Kronecker, các giá trị X_j ($j \leq 0$; $j > n$) không xác định và được coi bằng 0. Chu kỳ chuỗi thời vụ được chúng tôi sử dụng là 5 ngày tương ứng với một tuần giao dịch trên TTCK. Để tính chuỗi thời vụ, ta

lấy chuỗi thời gian ban đầu chia cho chuỗi xu thế, lấy trung bình giá trị tại các điểm có cùng ngày trong tuần từ thứ hai đến thứ sáu ta thu được mảng năm giá trị, lấy từng phần tử trong mảng này trừ đi giá trị trung bình của mảng, lặp theo chu kỳ 5 ngày các giá trị này ta sẽ thu được chuỗi thời vụ. Các giá trị trong chuỗi ngẫu nhiên thu được bằng cách lấy chuỗi thời gian đầu vào trừ đi chuỗi xu thế và chuỗi thời vụ tính toán ở trên. Hình 4 minh họa các chuỗi thành phần được phân tách từ chuỗi thời gian đầu vào. Đường trên cùng là chuỗi thời gian đầu vào. Đường thứ hai là chuỗi thời gian xu thế. Đường thứ ba là chuỗi thời gian thời vụ và đường cuối cùng là chuỗi thời gian ngẫu nhiên.



Hình 4. Các thành phần của chuỗi thời gian đầu vào.

Các phương pháp dự đoán dựa trên lý thuyết xác suất đều suy diễn dựa trên giả thiết độc lập giữa các giá trị của chuỗi, hay nói cách khác bộ dữ liệu đầu vào phải thỏa mãn điều kiện dừng. Qua phân tích biểu đồ hàm tự tương quan và phân phối các giá trị của chuỗi ngẫu nhiên, chúng tôi nhận thấy chuỗi ngẫu nhiên có tính dừng, còn chuỗi xu thế không có tính dừng. Chúng tôi biến đổi chuỗi xu thế bằng cách lấy sai phân bậc một của chuỗi xu thế để thu được chuỗi mới có tính dừng, gọi là chuỗi DT_t.

Tiếp đến, chúng tôi phân tập dữ liệu đầu vào thành tập huấn luyện và tập kiểm thử. Tập dữ liệu huấn luyện chứa các dữ liệu quan sát được và được dùng để huấn luyện mô hình giúp cho việc tìm ra các tham số mô hình theo cách suy diễn của mỗi phương pháp. Trong nghiên cứu này, tập huấn luyện là các giá trị nằm trong khoảng thời gian từ ngày 02/02/2015 tới ngày 13/04/2016 tương ứng với 296 ngày giao dịch trên TTCK.

Tập dữ liệu kiểm thử dùng để kiểm chứng phương pháp đã được huấn luyện trên tập dữ liệu huấn luyện. Tập kiểm thử chứa các dữ liệu quan sát được trên thực tế và được dùng để kiểm chứng mô hình dự đoán bằng cách so sánh giữa giá trị dự đoán và giá trị quan sát được để tính sai số dự đoán. Ở đây, chúng tôi sử dụng tập kiểm thử là các giá trị từ ngày 14/04/2016 đến 09/08/2016 tương ứng với 81 ngày giao dịch liên tiếp.

Ưu điểm của phân tích GPR là dựa trên toàn bộ dữ liệu huấn luyện đầu vào với độ dài lịch sử đủ lớn, mô hình có khả năng "học" để phát hiện các mẫu hình xuất hiện trong bộ dữ liệu huấn luyện [1–3]. Từ đó việc áp dụng phân tích GPR để dự đoán cho chuỗi xu thế nhằm tận dụng khả năng học của phương pháp này để tìm kiếm các mẫu hình lặp lại trong chuỗi xu thế T_t là khả thi. Mô hình ARMA thích hợp để dự đoán các chuỗi thời gian biến thiên ngẫu nhiên có tính dừng [4–6]. Như chỉ ra ở trên, với tính chất biến thiên ngẫu nhiên và có tính dừng của chuỗi ngẫu nhiên R_t, chuỗi ngẫu nhiên là đầu vào khả thi cho phương pháp dự đoán theo mô hình ARMA. Phần tiếp theo chúng tôi trình bày từng phương pháp dự đoán được sử dụng.

2. Phân tích hồi quy quá trình Gauss

Phân phối trong quá trình Gauss được biểu diễn bởi một hàm kỳ vọng $m(x)$ và một hàm hiệp phương sai $k(x, x')$. Trên thực tế ta thường coi biến ngẫu nhiên có kỳ vọng $m(x) = 0$ và chỉ quan tâm tới hàm hiệp phương sai [1], tức là

$$f(x) \sim GP(0, k(x, x')), \tag{4}$$

trong đó $k(x, x') = E[f(x)f(x')]$ biểu thị sự tương quan giữa các đầu ra $f(x)$ và $f(x')$ tương ứng với các biến đầu vào x và x', nói cách khác nó thể hiện sự phân phối giữa các hàm. Ma trận K biểu diễn mối tương quan giữa tất cả các biến đầu vào gọi là ma trận hiệp phương sai kích thước $n \times n$. Tham số của hàm hiệp phương sai được gọi là siêu tham số. Chúng tôi sử dụng hàm hiệp phương sai phổ biến và cũng được sử dụng trong [3], là hàm hiệp phương sai lũy thừa bình phương. Công thức hàm hiệp phương sai cho bởi

$$k(x, x') = \sigma^2 \exp\left[\frac{-(x - x')^2}{2l^2}\right]. \tag{5}$$

Hàm hiệp phương sai này có hai siêu tham số là $\theta = (\sigma^2, l)$. Để đánh giá các siêu tham số ta suy diễn sử dụng công thức xác suất Bayes như sau:

$$p(f|Y, X) = \frac{p(Y|X, f)p(f|X)}{p(Y|X)}, \tag{6}$$

trong đó $p(f|Y, X)$ gọi là xác suất hậu nghiệm (posterior), $p(Y|X, f)$ gọi là xác suất khả năng (likelihood), $p(f|X)$ gọi là xác suất tiên nghiệm, và $p(Y|X)$ gọi là xác suất biên (marginal likelihood). Các siêu tham số hàm hiệp phương sai tìm được sao cho hàm logarit của xác suất biên sau đây đạt giá trị lớn nhất [1]:

$$\log p(Y|X) = -\frac{1}{2}Y^T(K)Y - \frac{1}{2}\log|K| - \frac{n}{2}\log 2\pi. \tag{7}$$

Phương pháp được sử dụng để tìm điểm tối ưu cho hàm logarit xác suất biên ở trên là phương pháp Gradient liên hợp. Sau khi tìm được các siêu tham số cho hàm hiệp phương sai, ma trận hiệp phương sai K hoàn toàn xác định. Xác suất có điều kiện $p(f*|f)$ mang ý nghĩa là, đối với bộ dữ liệu huấn luyện tại các điểm f, việc dự đoán tại các điểm dữ liệu kiểm thử $f*$ sẽ cho độ chính xác với xác suất bao nhiêu. Phân phối của xác suất có điều kiện $p(f*|f)$ cũng là phân phối quá trình Gauss có dạng sau [1]:

$$f_*|X_*, X, f \ GP(\widehat{m}, \widehat{k}), \tag{8}$$

trong đó

$$\widehat{m} = K(X_*, X)K(X, X)^{-1}f$$
$$\widehat{k} = K(X_*, X_*) - K(X_*, X)K(X, X)^{-1}K(X, X_*).$$

Công thức (8) cho phép ta xác định kết quả dự đoán bằng việc lấy kỳ vọng f_* tại các điểm thử X_*.

Chuỗi xu thế biến đổi DT_t là đầu vào cho phương pháp phân tích GPR. Gọi $DT^*_{t,n+1}$ là kết quả dự đoán chuỗi DT_t tại một điểm kế tiếp cho bởi công thức (8). Gọi $T^*_{t,n+1}$ là kết quả dự đoán tại một điểm kế tiếp của chuỗi đầu vào T_t, first$(\cdot)$ là hàm lấy giá trị phần tử đầu tiên của chuỗi, sum$(\cdot)$ là hàm lấy tổng các giá trị của chuỗi. Đối với biến đổi lấy sai phân bậc một ta có kết quả sau:

$$T^*_{t,n+1} = \text{first}(T_t) + \text{sum}(DT_t) + DT^*_{t,n+1}. \tag{9}$$

Công thức (9) cho phép truy ngược kết quả dự đoán chuỗi xu thế T_t từ kết quả dự đoán biển đổi xu thế DT_t.

3. Mô hình tự hồi quy trung bình động

Mô hình ARMA là một quá trình được tạo ra bởi từ tổ hợp giữa các giá trị của chuỗi trong quá khứ và các giá trị của nhiễu trong quá khứ và hiện tại. Công thức sau thể hiện mối quan hệ giữa các đại lượng trong mô hình [4, 10]:

$$Y_t - \Phi_1 Y_{t-1} - \cdots - \Phi_p Y_{t-p} = X_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q},$$

hay tương ứng là

$$\Phi(B)Y_t = \theta(B)Z_t, \tag{10}$$

trong đó, $(Y_t, Y_{t-1}, \ldots)$ là các giá trị của chuỗi thời gian đầu vào; $(Z_t, Z_{t-1}, \ldots)$ là các sai số tương ứng với nhiễu trắng, kí hiệu là $Z_t \ WN(0, \sigma^2)$, B là toán tử dịch ngược thời gian ($B^j Y_t = Y_{t-1}$). Mô hình ARMA có các tham số là $\theta = (\Phi_1, \Phi_2, \ldots, \Phi_p, \theta_1, \theta_2, \ldots, \theta_q, \sigma^2)$. Các tham số của mô hình ARMA được xác định sao cho hàm logarit xác suất khả năng cho bởi công thức sau đây đạt giá trị lớn nhất [4, 5]:

$$\log p(\theta|Y) = -\frac{1}{2}\sum_{j=1}^{n}\frac{(Y_j - Y_j^*)^2}{\sigma^2 \upsilon_{j-1}} - \frac{1}{2}\sum_{j=0}^{n-1}\log(\sigma^2\upsilon_j) - \frac{n}{2}\log(2\pi). \tag{11}$$

Trong công thức (11), $\{v_i\}$ là kỳ vọng của sai số bình phương tại bước dự đoán tiếp theo. Sau khi xác định các tham số của mô hình, việc dự đoán tại một điểm kế tiếp thu được bằng các biến đổi chuỗi thời gian $\{Y_i\}$ thành chuỗi thời gian mới $\{W_i\}$ như sau:

$$W_t = \begin{cases} \sigma^{-1}Y_t, & 1 \le t \le m, \\ \sigma^{-1}\Phi(B)Y_t, & t > m. \end{cases} \tag{12}$$

Trong công thức (12), $m = \max(p, q)$. Giá trị dự đoán tại điểm kế tiếp của chuỗi $\{W_i\}$ được cho bởi công thức sau [4, 5]:

$$W^*_{n+1} = \begin{cases} \sum_{j=1}^{n}\theta_{nj}(W_{n+1-j} - W^*_{n+1-j}), & 1 \le n < m, \\ \sum_{j=1}^{q}\theta_{nj}(W_{n+1-j} - W^*_{n+1-j}), & n \ge m. \end{cases} \tag{13}$$

Các hệ số θ_{nj} trong công thức (13) được xác định từ giải thuật Innovations [4, 5] cho bởi công thức đệ quy

$$\begin{cases} \upsilon_0 = \kappa(1,1), \\ \theta_{n,n-k} = \upsilon_k^{-1}[\kappa(n+1,k+1) - \sum_{j=0}^{k-1}\theta_{k,k-j}\theta_{n,n-j}\upsilon_j], \\ \upsilon_n = \kappa(n+1,n+1) - \sum_{j=0}^{n-1}\theta^2_{n,n-j}\upsilon_j, \end{cases} \tag{14}$$

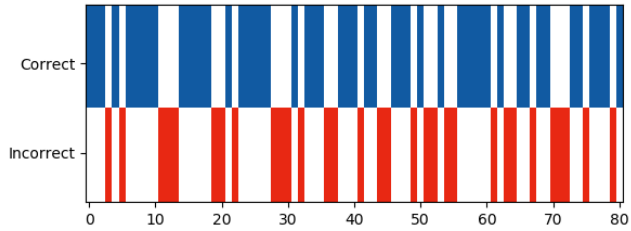
trong đó $\kappa(i, j)$ là hàm tự tương quan giữa hai điểm (i, j) của chuỗi $\{W_i\}$. Từ công thức (12) và (13), với nhận xét $W_t - W^*_t = \sigma^{-1}(Y_t - Y^*_t)$; $\forall t \ge 1$, ta có kết quả dự đoán tại một điểm kế tiếp $(t+1)$ theo mô hình ARMA cho bởi công thức (15) dưới đây.

$$Y^*_{n+1} = \begin{cases} \sum_{j=1}^{n}\theta_{nj}(Y_{n+1-j} - Y^*_{n+1-j}), & 1 \le n < m, \\ \sum_{j=1}^{q}Y_{n+1-j} + \sum_{j=1}^{p}\theta_{nj}(Y_{n+1-j} - Y^*_{n+1-j}), & n \ge m. \end{cases} \tag{15}$$

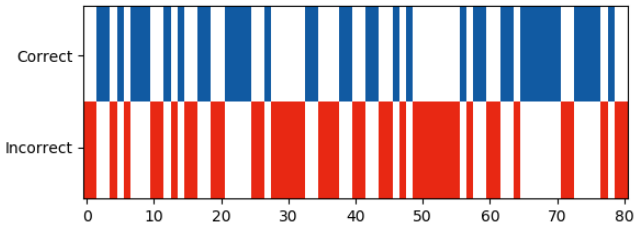
III. CÀI ĐẶT VÀ ĐÁNH GIÁ THỰC NGHIỆM

Chương trình thực hiện phương pháp dự đoán kết hợp GPR-ARMA được cài đặt bằng ngôn ngữ Python, chạy trên hệ điều hành Windows Server 64-bit, sử dụng các gói thư viện xử lý toán học và thống kê như Numpy, Scipy, Pandas, Statsmodels; gói thư viện xử lý đồ họa Matplotlib và gói thư viện xử lý phân tích GPR là PyGPs [15].

Đánh giá kết quả dự đoán ngoại suy của phương pháp kết hợp GPR-ARMA được thực hiện như sau: lặp qua từng điểm trong tập kiểm thử để dự đoán giá đóng cửa chỉ số chứng khoán VN-Index tại mỗi điểm kiểm thử đó theo công thức (2). Sau mỗi bước lặp ta bổ sung giá trị quan sát tại điểm được dự đoán vào tập huấn luyện và lặp lại các bước



Hình 5. Kết quả dự đoán xu thế chỉ số VN-Index theo phương pháp ARMA.



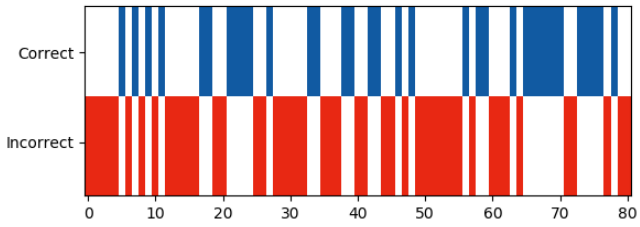
Hình 6. Kết quả dự đoán xu thế chỉ số VN-Index theo phương pháp GPR.

thực hiện quá trình dự đoán. Phương pháp tối ưu hóa trong phân tích GPR được chúng tôi sử dụng là phương pháp Gradient liên hợp tuyến tính với số bước lặp khởi tạo là 30. Sử dụng kết quả bài báo [6], đối với mô hình tự hồi quy trung bình động ARMA, chúng tôi sử dụng tiêu chuẩn thông tin Akaike để tìm bộ tham số (p, q) ở mỗi bước lặp sao cho tiêu chuẩn thông tin Akaike đạt giá trị nhỏ nhất.

Chúng tôi cài đặt thực nghiệm phương pháp dự đoán GPR-ARMA và thu được đồ thị biểu diễn trực quan kết quả dự đoán của phương pháp GPR-ARMA cho cho 81 ngày giao dịch trong tập kiểm thử từ 14/04/2016 đến 09/08/2016 như Hình 7.

Tiếp đến, chúng tôi tiến hành cài đặt từng phương pháp dự đoán riêng lẻ là phân tích GPR và mô hình ARMA sử dụng cùng bộ dữ liệu đầu vào và thực hiện dự đoán cùng tập dữ liệu kiểm thử với phương pháp kết hợp GPR-ARMA. Phương pháp phân tích GPR và mô hình ARMA được cài đặt bằng cách biến đổi dữ liệu đầu vào sử dụng lấy sai phân bậc một. Dữ liệu biến đổi này là đầu vào cho từng phương pháp và thực hiện truy ngược kết quả dự đoán cho chuỗi thời gian đầu vào tương tự công thức (9). Kết quả thực nghiệm từng phương pháp riêng lẻ, chúng tôi thu được đồ thị biểu diễn kết quả dự đoán như sau.

Từ các hình 5, 6 và 7, ta có thể nhận thấy phương pháp kết hợp GPR-ARMA cho kết quả dự đoán tốt hơn khi mật độ các ngày dự đoán đúng nhiều hơn so với từng phương pháp riêng lẻ. Để định lượng chính xác, chúng tôi coi bài toán dự đoán xu thế chỉ số chứng khoán VN-Index là một bài toán phân lớp, bao gồm lớp tăng và lớp giảm. Kết quả dự đoán được xếp vào lớp tăng khi dự đoán chỉ số VN-



Hình 7. Kết quả dự đoán xu thế chỉ số VN-Index theo phương pháp GPR-ARMA.

Bảng I
BẢNG NHẦM LẪN KẾT QUẢ DỰ ĐOÁN XU THẾ THEO PHƯƠNG PHÁP KẾT HỢP GPR-ARMA

Tất cả các lớp	Thuộc lớp	Không thuộc lớp
Dự đoán thuộc lớp	TP = 50	FP = 31
Dự đoán không thuộc lớp	FN = 31	TN = 50

Bảng II
CÁC ĐẠI LƯỢNG SAI SỐ DỰ ĐOÁN CỦA TỪNG PHƯƠNG PHÁP DỰ ĐOÁN ĐƯỢC NGHIÊN CỨU

Phương pháp	RMSE	MAD	MAPE
ARMA	6,034	4,717	0,0075
GPR	8,176	6,416	0,0102
GPR-ARMA	**6,015**	**4,564**	**0,0073**

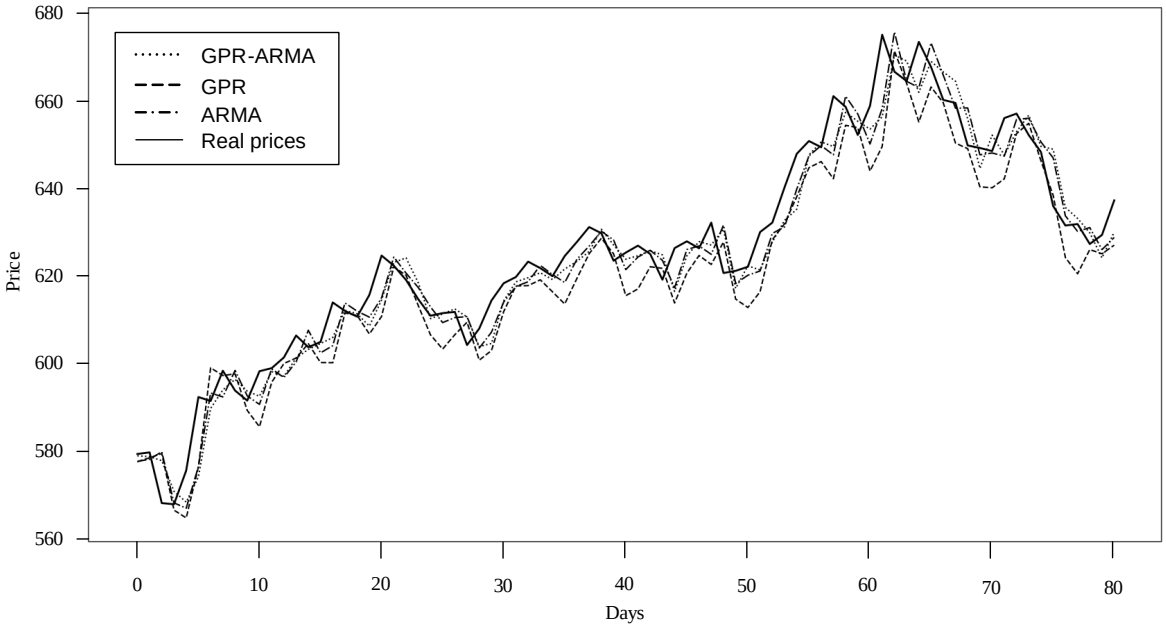
Index ngày giao dịch kế tiếp lớn hay bằng chỉ số VN-Index ngày giao dịch hiện tại. Kết quả dự đoán được xếp vào lớp giảm khi dự đoán chỉ số VN-Index ngày giao dịch kế tiếp nhỏ hơn chỉ số VN-Index ngày giao dịch hiện tại. Từ kết quả thực nghiệm phương pháp GPR-ARMA, chúng tôi thu được bảng nhầm lẫn dự đoán (Bảng I) [16].

Từ Bảng nhầm lẫn của kết quả dự đoán xu thế chỉ số VN-Index, chúng tôi tính độ chính xác kết quả dự đoán xu thế chỉ số VN-Index theo phương pháp kết hợp GPR-ARMA theo công thức sau [16]:

$$P_{\text{GPR-ARMA}} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{50}{50 + 31} = 61,73\%. \qquad (16)$$

Thực hiện tính toán tương tự, chúng tôi thu được độ chính xác dự đoán xu thế chỉ số VN-Index của phương pháp phân tích GPR là 48,15% và độ chính xác của phương pháp ARMA là 41,98%. Các đại lượng đánh giá sai số dự đoán bao gồm RMSE, độ lệch trị tuyệt đối trung bình (MAD: Mean absolute deviation) và phần trăm sai số trị tuyệt đối trung bình (MAPE: Mean absolute percentage error) của từng phương pháp được cho trong Bảng II.

Hình 8 biểu diễn trực quan đồ thị dự đoán chỉ số VN-Index của từng phương pháp. Như vậy kết quả thực nghiệm cho thấy so với từng phương pháp dự đoán riêng lẻ, phương pháp dự đoán kết hợp GPR-ARMA cho độ chính xác cao



Hình 8. Kết quả dự đoán giá chỉ số VN-Index của từng phương pháp dự đoán được nghiên cứu.

nhất là 61,73%. Đồng thời, các sai số dự đoán thấp hơn so với từng phương pháp dự đoán riêng lẻ.

IV. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Chúng tôi đã trình bày một phương pháp mới giải bài toán dự đoán xu thế VN-Index nhờ việc phân tách chuỗi thời gian đầu vào và sử dụng kết hợp phân tích GPR và mô hình ARMA để dự đoán các chuỗi thời gian thành phần một cách hợp lý, qua đó tận dụng ưu điểm của từng phương pháp dự đoán riêng lẻ. Thành phần xu thế thể hiện xu thế rõ ràng hơn nhờ việc loại bỏ nhiễu là thành phần ngẫu nhiên, nên việc áp dụng phân tích GPR làm tăng khả năng học để nhận biết các mẫu hình lặp lại trong chuỗi xu thế. Thành phần ngẫu nhiên có tính dừng, có giá trị biến thiên ngẫu nhiên, nên phù hợp để áp dụng mô hình ARMA dự đoán cho thành phần ngẫu nhiên này. Kết quả dự đoán các thành phần riêng lẻ được tổng hợp lại để đưa ra kết quả dự đoán cuối cùng cho phương pháp kết hợp GPR-ARMA. Kết quả thực nghiệm cho thấy, với cùng bộ dữ liệu đầu vào và cùng tập kiểm thử tiến hành dự đoán, phương pháp kết hợp GRP-ARMA cho độ chính xác cao nhất là $P_{\text{GPR-ARMA}} = 61,73\%$ (dự đoán đúng 50 ngày trong số 81 ngày tiến hành dự đoán). Các phương pháp dự đoán riêng lẻ là phân tích GPR và mô hình ARMA có độ chính xác dự đoán thấp hơn nhiều so với phương pháp kết hợp. Đồng thời, giá trị các sai số dự đoán RMSE, MAD và MAPE của phương pháp kết hợp GPR-ARMA đều thấp hơn so với từng phương pháp dự đoán riêng lẻ. Phương pháp của chúng tôi đã tận dụng được ưu điểm của từng phương pháp dự đoán riêng lẻ để có kết quả dự đoán tốt hơn. Từ đó khẳng định tính đúng đắn của phương pháp dự đoán kết hợp GPR-ARMA được đề xuất.

Mỗi mô hình định lượng được sử dụng trong bài báo này đều có thể được cải tiến nhằm tăng độ chính xác dự đoán của phương pháp kết hợp GPR-ARMA. Với mô hình ARMA, việc biến đổi dữ liệu đầu vào phù hợp để làm giảm khoảng cách biến thiên giữa các điểm có thể tăng độ chính xác của phương pháp này. Với phân tích GPR, việc lựa chọn các lớp hàm hiệp phương sai tốt có thể cải thiện đáng kể độ chính xác của phương pháp này. Một hướng phát triển tiếp theo là sử dụng các giải thuật xấp xỉ để cải thiện tốc độ tính toán cho phân tích GPR khi dữ liệu đầu vào lớn. Cuối cùng, phương pháp GPR-ARMA là phương pháp dự đoán tổng quát cho chuỗi thời gian bất kì nên phương pháp này có thể sử dụng để dự đoán các chuỗi thời gian khác như giá cổ phiếu, hay giá của các chỉ số chứng khoán khác như chỉ số S&P 500, Nasdaq, Dow Jones, FTSE 100, BSE SENSEX.

TÀI LIỆU THAM KHẢO

[1] C. E. Rasmussen and C. K. Williams, "Gaussian processes for machine learning. 2006," *The MIT Press, Cambridge, MA, USA*, vol. 38, pp. 715–719, 2006.

[2] B. Wang and T. Chen, "Gaussian process regression with multiple response variables," *Chemometrics and Intelligent Laboratory Systems*, vol. 142, pp. 159–165, 2015.

[3] M. T. Farrell and A. Correa, "Gaussian process regression models for predicting stock trends," *Relation*, vol. 10, pp. 1–9, 2007.

[4] P. J. Brockwell, R. A. Davis, and M. V. Calder, *Introduction to time series and forecasting*, 2nd ed. Springer, 2010.

[5] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*, 5th ed. John Wiley & Sons, 2015.
[6] P. Mondal, L. Shit, and S. Goswami, "Study of effectiveness of time series modeling (arima) in forecasting stock prices," *International Journal of Computer Science, Engineering and Applications*, vol. 4, no. 2, pp. 13–29, 2014.
[7] G. Dutta, P. Jha, A. K. Laha, and N. Mohan, "Artificial neural network models for forecasting stock price index in the bombay stock exchange," *Journal of Emerging Market Finance*, vol. 5, no. 3, pp. 283–295, 2006.
[8] Y. Zuo and E. Kita, "Up/down analysis of stock index by using bayesian network," *Engineering Management Research*, vol. 1, no. 2, pp. 46–52, 2012.
[9] S. S. Patil, K. Patidar, and M. Jain, "Stock market prediction using support vector machine," *International Journal of Current Trends in Engineering & Technology*, vol. 2, no. 1, pp. 18–25, 2016.
[10] T. Awokuse and T. Ilvento, "Using statistical data to make decisions-module 6: Introduction to time series forecasting," *University of Delaware, College of Agriculture and Natural Resources, Food and Resource Economics*, 2012. [Online]. Available: http://www1.udel.edu/FREC/ilvento/BUAD820/MOD604.pdf
[11] E. Haven, P. Molyneux, J. O. Wilson, S. Fedotov, and M. Duygun, *The Handbook of Post Crisis Financial Modelling*. Springer, 2016.
[12] Đỗ Văn Thành, Nguyễn Minh Hải, "Phân tích và dự báo chỉ số thị trường chứng khoán bằng sử dụng chỉ số báo trước," in *Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9)*, Cần Thơ, Việt Nam, Aug., pp. 559–565.
[13] Hồ Thủy Tiên, Hồ Thu Hoài, Ngô Văn Toàn, "Mô hình hóa biến động thị trường chứng khoán: Thực nghiệm từ việt nam," *Tạp chí Khoa học ĐHQGHN: Kinh tế và Kinh doanh*, vol. 33, no. 3, pp. 1–11, 2017.
[14] M. H. Nguyen and O. Darné, "Forecasting and risk management in the vietnam stock exchange," *Laboratoire d'Economie et de Management Nantes-Atlentique Université de Nantes*, 2018. [Online]. Available: https://halshs.archives-ouvertes.fr/halshs-01679456
[15] M. Neumann, S. Huang, D. E. Marthaler, and K. Kersting, "pygps: A python library for gaussian process regression and classification," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 2611–2616, 2015.
[16] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge University Press, 2012.

Huỳnh Quyết Thắng sinh năm 1967 tại Hà Nội. Ông tốt nghiệp Trường Đại học Điện-Máy Varna, Cộng hòa Bungary, năm 1990; nhận bằng Tiến sĩ tại Trường Tổng hợp kỹ thuật Varna (TU Varna), Cộng hòa Bungary, năm 1995; nhận học hàm PGS năm 2007. Hiện nay, ông đang công tác tại Viện Công nghệ Thông tin và Truyền thông, Trường Đại học Bách khoa Hà Nội.

Lĩnh vực nghiên cứu ông quan tâm là Techniques and Math models in Software Quality Prediction/Measurement; Secure Coding, Program Analysis; Methods in Software Development; Cost/Effort Evaluation.

Phùng Đình Vũ sinh năm 1989 tại Nam Định. Ông tốt nghiệp Đại học và Thạc sĩ Công nghệ thông tin tại Trường Đại học Bách khoa Hà Nội năm 2012 và 2017. Lĩnh vực nghiên cứu ông quan tâm là Các mô hình định lượng như Gaussian Process, mạng Nơ-ron, Giải thuật di truyền, mạng Bayes, Support Vector Machine.

Tống Văn Vinh sinh năm 1997 tại Hà Nội. Tác giả là sinh viên năm thứ tư, lớp Kỹ sư Tài năng, chuyên ngành Công nghệ Thông tin, Trường Đại học Bách khoa Hà Nội. Lĩnh vực nghiên cứu quan tâm của tác giả là Gaussian Process, mạng Nơ-ron, Support Vector Machine, mạng Bayes.