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, ngày 4-5/8/2016
DOI: 10.15625/vap.2016.00037

MÔ HÌNH DỰ BÁO TẦN SUẤT CAO ĐỐI VỚI CHỈ SỐ THỊ TRƯỜNG CHỨNG KHOÁN

Đỗ Văn Thành[1] và Nguyễn Minh Hải[2]

[1]Khoa Công nghệ thông tin, Trường Đại học Nguyễn Tất Thành
[2]Khoa cơ bản, Trường Đại học Công nghiệp, Thành phố Hồ Chí Minh
dvthanh@ntt.edu.vn, nguyenminhhaidhcn@gmail.com.

TÓM TẮT — *Bài báo đề xuất phương pháp kết hợp kỹ thuật hồi quy nhiều biến và kỹ thuật phân tích thành phần chính (PCA) trên tập các biến dữ liệu giao dịch cổ phiếu có tương quan mẫu cao với chỉ số thị trường chứng khoán để xây dựng mô hình dự báo chỉ số thị trường chứng khoán. Việc thực hành xây dựng mô hình dự báo được thực hiện trên dữ liệu thực theo ngày của sàn giao dịch chứng khoán thành phố Hồ Chí Minh từ đầu năm 2010 đến nay. Kết quả dự báo bằng mô hình được xây dựng cho thấy triển vọng tốt của phương pháp được đề xuất trong dự báo chỉ số thị trường chứng khoán cũng như nhiều chỉ số kinh tế - xã hội khác.*

Từ khóa— *Mô hình dự báo, chỉ số thị trường chứng khoán, phân tích thành phần chính, tần suất cao.*

I. GIỚI THIỆU

Dự báo thị trường chứng khoán là vấn đề luôn được cộng đồng quốc tế quan tâm. Dự báo thị trường chứng khoán bao gồm 2 nội dung chính là dự báo chỉ số thị trường chứng khoán (nói gọn là chỉ số thị trường chứng khoán) và dự báo giá của các biến dữ liệu giao dịch cổ phiếu trên thị trường [9].

Hiện đã có hàng trăm bài báo nghiên cứu đề xuất các cách tiếp cận kỹ thuật dự báo thị trường chứng khoán. Đã có một số nghiên cứu điều tra và tổng quan tương đối có hệ thống về các kỹ thuật dự báo thị trưởng chứng khoán [9]. Theo đó các cách tiếp cận kỹ thuật dự báo thị trường chứng khoán đã được phân loại và được giới thiệu trong [3, 9]. Một số kỹ thuật dự báo thị trường chứng khoán được nghiên cứu nhiều trong vài năm gần đây cũng đã được giới thiệu trong [3], trong đó có kỹ thuật phân tích thành phần chính (PCA) [1-2, 5, 7, 10-14]. Những phương pháp kết hợp sử dụng kỹ thuật PCA với một hoặc một vài kỹ thuật khai phá dữ liệu khác như phương pháp học máy véc tơ hỗ trợ, giải thuật di truyền, mạng nơtron,… trong phân tích và dự báo thị trường chứng khoán đang được quan tâm nghiên cứu ứng dụng [1, 4, 11-12, 14].

PCA là kỹ thuật được sử dụng để làm đơn giản hóa một tập dữ liệu phức tạp thành một tập dữ liệu nhỏ hơn và có thể hiểu dễ dàng hơn [6]. PCA khác với các kỹ thuật hồi qui nhiều biến ở chỗ các kỹ thuật hồi qui nhiều biến tập trung vào các suy luận thống kê hoặc xây dựng mô hình để giải thích hoặc dự báo mối quan hệ giữa các biến. PCA tóm lược các tập dữ liệu lớn và phức tạp bằng cách tạo ra các biến mới là tổ hợp tuyến tính có trọng số của các biến ban đầu. Mỗi biến mới được gọi là thành phần chính sẽ không tương quan với các thành phần chính khác.

Với việc làm giảm số chiều các biến, PCA rất hữu ích trong phân tích thống kê nhiều biến. Các thành phần chính là dữ liệu tần suất cao nên mô hình dự báo có các biến giải thích là các thành phần chính được gọi là mô hình dự báo tần suất cao [4]. Hầu hết các mô hình dự báo kinh tế được giới thiệu trong [5] là được phát triển ở một số nước có trình độ khoa học phát triển trong những năm gần đây đều là những mô hình dự báo được xây dựng bằng phương pháp hồi quy nhiều biến trên tập các biến giải thích là các thành phần chính của tập biến gốc và chúng đều là những mô hình dự báo tần suất cao [1, 4, 11-12].

Các công trình nghiên cứu trong [5] là những gợi ý quan trọng cho nghiên cứu bài báo này. Mục đích của bài báo này là đề xuất phương pháp xây dựng mô hình dự báo chỉ số thị trường chứng khoán bằng sử dụng phương pháp hồi qui nhiều biến, ở đó các biến giải thích là các thành phần chính của tập biến dữ liệu ban đầu. Các thành phần chính được lựa chọn sao cho nắm bắt được những thay đổi nhiều nhất như có thể của tập dữ liệu gốc. Do các biến mới là các thành phần chính nên trong mô hình dự báo sẽ không còn hiện tượng đa cộng tuyến và việc giải thích các yếu tố tác động thực sự đến thay đổi của chỉ số thị trường chứng khoán sẽ dễ dàng và chuẩn xác hơn. Nói ngắn gọn bài báo này sẽ kết hợp kỹ thuật phân tích hồi qui nhiều biến và kỹ thuật PCA nhiều biến trong việc dự báo chỉ số thị trường chứng khoán. Khác với hầu hết các nghiên cứu dự báo về thị trường chứng khoán bằng sử dụng các kỹ thuật khai phá dữ liệu trước đây, phương pháp dự báo này sẽ chú trọng đưa ra kết quả dự báo cụ thể chứ không đơn thuần chỉ là dự báo xu thế của thị trường.

Việc thực hành xây dựng mô hình dự báo chỉ số thị trường chứng khoán theo phương pháp được đề xuất sẽ được thực hiện trên tập dữ liệu thực của sàn giao dịch chứng khoán Thành phố Hồ Chí Minh.

Bài báo được cấu trúc thành 5 mục. Mục 2 tiếp theo sẽ giới thiệu một số công trình nghiên cứu liên quan. Mục 3 sẽ trình bày phương pháp xây dựng mô hình dự báo tần suất cao đối với chỉ số thị trường chứng khoán. Mục 4 sẽ trình bày kết quả xây dựng mô hình dự báo chỉ số thị trường chứng khoán theo phương pháp được đề xuất trên tập dữ liệu thực tế theo tần suất ngày của sàn giao dịch chứng khoán Thành phố Hồ Chí Minh. Việc đánh giá chất lượng dự báo của mô hình cũng được thực hiện trong Mục này và Mục 5, cuối cùng, sẽ trình bày một số kết luận.

II. NHỮNG NGHIÊN CỨU LIÊN QUAN

Trong mục này sẽ tổng quan một số công trình nghiên cứu có sử dụng kỹ thuật PCA trong phân tích và xây dựng mô hình dự báo thị trường chứng khoán trong những năm gần đây.

Wang Yansshan [14] đã sử dụng kết hợp kỹ thuật học máy véc tơ hỗ trợ và kỹ thuật PCA để làm giảm các điểm dữ liệu thành hai thành phần và quan sát thấy rằng có thể hình thành một cụm các cổ phiếu cùng thay đổi bằng việc sử dụng các thành phần được tạo ra từ PCA khi nghiên cứu dự báo chỉ số giá cổ phiếu tổng hợp KOSPI của Hàn Quốc và chỉ số thị trường chứng khoán Hangseng (HSI).

Phân tích thành phần chính dạng hàm (FPCA) là tương tự như PCA nhiều biến thông thường nhưng ở đây trọng số của các thành phần chính (hay véc tơ riêng) là các hàm số phụ thuộc thời gian. Wang [13], đã sử dụng kỹ thuật FPCA để nghiên cứu về giá của 50 cổ phiếu giao dịch chính trên sàn giao dịch chứng khoán Thượng Hải. Wang [13] đã làm giảm số chiều và rút ra được những thành phần giàu ý nghĩa nhất. So sánh với PCA thông thường thì FPCA có thể giải quyết được vấn đề các chiều trong mẫu là khác nhau và nó là cách tiếp cận hợp lý để rút ra các yếu tố thay đổi chính.

Mbeledogu [7], cũng đã sử dụng kỹ thuật PCA để giảm các biến dữ liệu giao dịch cổ phiếu thành 9 biến mới nhằm phục vụ cho hệ thống dự báo giao dịch cổ phiếu Nigiêria. Nghiên cứu này đã lựa chọn được các thành phần mang nhiều thông tin nhất và loại bỏ các thành phần dư thừa để nâng cao hiệu quả cho việc phân lớp, phân loại tiếp theo. Xa hơn kết quả bài báo sẽ hỗ trợ việc sử dụng PCA để nhận diện các yếu tố quan trọng nhất trong quá trình làm giảm đáng kể số lượng các biến đầu vào nhưng vẫn đảm bảo hiệu quả và đầy đủ thông tin cho việc phân lớp, phân loại.

Carol Anne Hargreaves, Chandrika Kadirvel Mani [2] đã rút ngắn quá trình và thời gian xác định một cổ phiếu tốt bằng cách giảm số lượng các biến để phân tích và chỉ phân tích các biến quan trọng. Trong thực hành các tác giả đã chọn 22 biến chứa thông tin cơ bản về kinh tế vĩ mô hoặc thông tin về tài chính và thực hiện kỹ thuật PCA để giảm 22 biến thành một số ít hơn đáng kể các biến.

Có thể thấy rằng việc sử dụng đơn lẻ kỹ thuật PCA trong nghiên cứu thị trường chứng khoán chỉ có thể hỗ trợ làm giảm số lượng các biến đầu vào cần được phân tích, làm cho việc phân tích dữ liệu thị trường chứng khoán trở lên thuận lợi, dễ dàng hơn [2, 7, 13]. Việc kết hợp sử dụng kỹ thuật PCA và kỹ thuật học máy véc tơ hỗ trợ trong nghiên cứu thị trường chứng khoán [14] giúp cho có thể phân cụm, phân lớp dữ liệu, chỉ có thể hỗ trợ dự báo xu thế chưa hỗ trợ trực tiếp cho việc xây dựng mô hình dự báo thị trường chứng khoán nói chung, chỉ số thị trường chứng khoán nói riêng.

Phân tích mô hình dự báo kinh tế ở mức quốc gia ở một số nước có trình độ khoa học phát triển được xây dựng trong những năm gần đây [5] có thể nhận thấy:

- Việc lựa chọn các biến gốc đầu vào cho các biến phụ thuộc là dựa vào lý thuyết kinh tế. Các biến giải thích thường là các chỉ số thuộc tài khoản quốc gia, tài khoản thu nhập và tài khoản sản xuất nên nói chung chúng là khá lớn [4].

- Hầu hết các mô hình dự báo kinh tế trong [5] đều sử dụng kỹ thuật PCA để lựa chọn các thành phần chính thay thế các biến giải thích gốc ban đầu trong mô hình dự báo [1, 4, 11-12].

- Dạng phương trình hồi qui của biến phụ thuộc theo các thành phần chính có thể ở dạng tổng quát nhất là mô hình trễ phân bố tự hồi qui [1] hoặc ở dạng rút gọn hơn là hàm loga tuyến tính nhiều biến [4, 11-12]. Quá trình hồi qui nhiều biến có thể sử dụng thêm biến giả để loại bỏ những điểm dữ liệu bất thường và xem phần dư như là một mô hình ARMA để khắc phục hiện tượng tự tương quan chuỗi của phần dư cũng như kỳ vọng phần dư khác 0; xem xét đưa vào các yếu tố phi tuyến (tích chéo của các biến giải thích) khi xử lý phần dư của phương trình ước lượng có phương sai không đổi.

Những nhận xét trên là những gợi ý cho thực hiện nghiên cứu xây dựng mô hình dự báo chỉ số thị trường chứng khoán, trong đó nhất là [1] trong việc chỉ định mô hình trễ phân bố tự hồi qui làm mô hình lý thuyết cho mô hình dự báo cần xây dựng.

III. PHƯƠNG PHÁP XÂY DỰNG MÔ HÌNH DỰ BÁO

Ký hiệu Y là biến chỉ số thị trường chứng khoán. Y là một biến véctơ, $Y^T = (y_1, y_2, \ldots, y_m)$, y_i , i=1, 2, …, m, là giá trị của biến Y ở ngày thứ i, m là số ngày giao dịch được thực hiện trên sàn.

Ký hiệu X_j (j =1, 2, …, n), là biến dữ liệu giao dịch cổ phiếu thứ j và n là số các biến dữ liệu giao dịch cổ phiếu được niêm yết trên sàn, X_j cũng là biến véctơ, $X_j^T = (x_{j1}, x_{j2}, \ldots, x_{jm})$, ở đây x_{jk} là giá trị của X_j trong ngày thứ k, nó là tích của giá trung bình của biến dữ liệu giao dịch cổ phiếu X_j nhân với số lượng cổ phiếu này được giao dịch trong ngày thứ k. Ta có thể sử dụng mã cổ phiếu làm tên của biến dữ liệu giao dịch cổ phiếu.

Ký hiệu A = (X_1, X_2, …, X_n) là véc tơ mà mỗi phần tử của nó là véctơ cột m chiều. Nói cách khác A biểu diễn một ma trận mxn chiều: A = (x_{ij}), i = 1, 2, …,m , j = 1, 2, …, n ở đây x_{ij} là giá trị giao dịch trong ngày giao dịch thứ i của biến dữ liệu giao dịch cổ phiếu X_j.

Phương pháp xây dựng mô hình dự báo tần suất cao đối với chỉ số thị trường chứng khoán được trình bày trong Hình 1 ở dưới và được giải thích tóm tắt như sau:

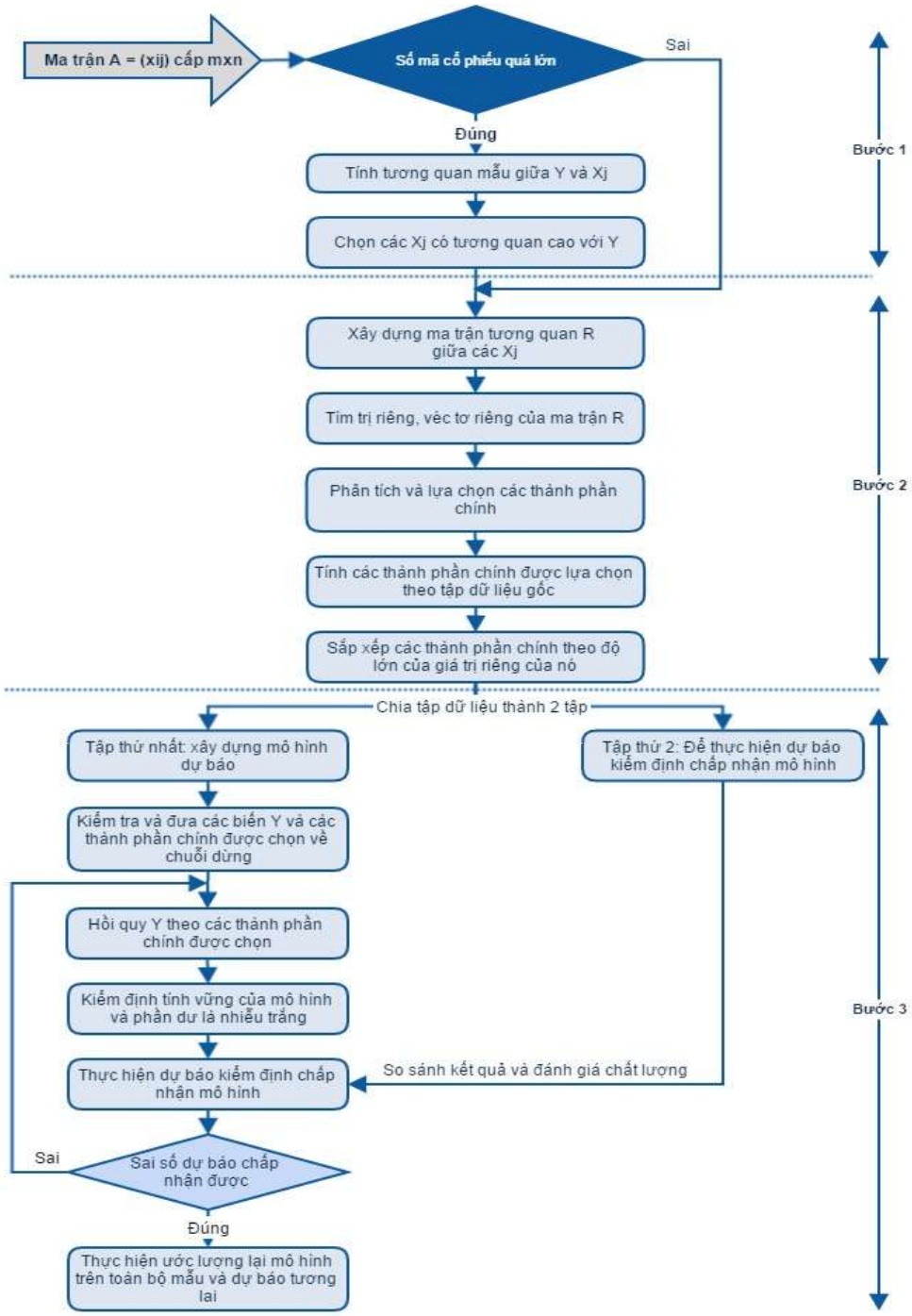
3.1. Bước 1: *Lựa chọn các biến dữ liệu giao dịch cổ phiếu có tương quan mẫu cao với chỉ số thị trường chứng khoán*

1) Nếu số lượng các biến dữ liệu giao dịch cổ phiếu sàn không quá lớn (khuyến nghị là dưới 100) thì chuyển sang Bước 3;

2) Tính tương quan mẫu giữa chỉ số thị trường chứng khoán Y và các biến dữ liệu giao dịch cổ phiếu X_j;

3) Lựa chọn các biến dữ liệu giao dịch cổ phiếu có tương quan mẫu cao với chỉ số thị trường chứng khoán Y (mức tương quan mẫu cao do người dùng quyết định).

Không giảm tổng quát, giả sử tất cả các biến dữ liệu giao dịch cổ phiếu X_j (j =1, 2, …., n) đều là các biến dữ liệu giao dịch cổ phiếu có tương quan mẫu cao với Y và chúng đều được chọn để thực hiện PCA.



Hình 1. Phương pháp xây dựng mô hình dự báo tần suất cao đối với chỉ số thị trường chứng khoán

3.2. Bước 2*: Tính toán và lựa chọn thành phần chính đại diện cho tập dữ liệu gốc*

4) Tính ma trận tương quan của tập các biến dữ liệu giao dịch cổ phiếu đã được lựa chọn ở **Bước 1**, ký hiệu là ma trận vuông **R** bậc h (h≤n).

5) Tính giá trị riêng và véc tơ riêng của ma trận **R**.

6) Xác định số thành phần chính bằng cách sắp xếp giá trị riêng theo giá trị giảm dần và phân tích tỷ lệ tích lũy của các giá trị riêng. Tỷ lệ tích lũy của các giá trị riêng cũng chính là tỷ lệ nắm bắt những thay đổi của tập dữ liệu gốc của thành phần chính tương ứng với các giá trị riêng đó. Một số nhà nghiên cứu gợi ý nên giữ lại tất cả các thành phần chính có giá trị riêng lớn hơn 1 trong khi nhiều nhà nghiên cứu khác gợi ý nên giữ lại các thành phần chính có thể giải thích được từ 70% - 90% sự thay đổi trong tập dữ liệu gốc.

Giả sử có k thành phần chính được giữ lại, ký hiệu là PC_1, PC_2, …, PC_k. khi đó (k ≤ h).

7) Xác định các thành phần chính, nó là tổ hợp tuyến tính của tập dữ liệu gốc được chuẩn hóa với trọng số là các véc tơ riêng tương ứng. Cụ thể:

Giả sử V_1, V_2, …., V_k là các véc tơ riêng ứng với các thành phần chính PC_1, PC_2, …, PC_k. mỗi V_i là một véc tơ h chiều cụ thể $V_i^T = (v_{1i}, v_{2i}, …, v_{hi})$, khi đó thành phần chính PC_i là véc tơ m chiều ứng với véc tơ riêng V_i được xác định như sau [6]:

$$PC_i = v_{1i}*\hat{X}_1 + v_{2i}*\hat{X}_2 + … + v_{hi}*\hat{X}_h, \quad (1)$$

ở đây $\hat{X}_j = \frac{X_j - \bar{\bar{X}}_j}{s_j}$, (2)

trong đó $\bar{\bar{X}}_j$, S_j tương ứng là kỳ vọng (giá trị trung bình) và độ lệch chuẩn của véc tơ X_j. Các véc tơ $\hat{X}_j$ được gọi là véc tơ chuẩn hóa của véc tơ X_j. Nói cách khác các thành phần chính được tính theo h véc tơ cột của ma trận A được chuẩn hóa và lấy các véc tơ giá trị riêng làm trọng số. Véc tơ riêng V_i là trọng số của thành phần chính PC_i tương ứng.

3.3. Bước 3: *Xây dựng mô hình dự báo chỉ số thị trường chứng khoán tần suất cao*

8) Chia tập dữ liệu thành 2 tập. Tập dữ liệu thứ nhất để xây dựng mô hình dự báo và tập dữ liệu thứ 2 để kiểm định mô hình dự báo.

9) Kiểm tra tính dừng và đưa về chuỗi dừng biến Y và các thành phần chính PC_1, PC_2, ⋯, PC_k nhằm đảm bảo phép hồi qui trên các biến này là hồi qui đúng. Giả sử Y và các thành phần chính PC_1, PC_2, ⋯, PC_k đều là chuỗi dừng.

10) Hồi quy biến Y theo các thành phần chính theo công thức:

$$Y = \sum_{i=0}^{5} a_{1i} PC_1(-i) + \sum_{i=0}^{5} a_{2i} PC_2(-i) + … + \sum_{i=0}^{5} a_{ki} PC_k(-i) + \sum_{q=1}^{5} b_q Y(-q) + u_t \quad (3)$$

ở đây u_t là phần dư của phương trình ước lượng được giả thiết là nhiễu trắng. Quá trình hồi qui được lặp đi lặp lại sao cho tất cả các tham số ước lượng α_i (i= 1, 2, …, k) trong mô hình dự báo đều có ý nghĩa thống kê, theo thông lệ thường ở mức dưới 10%.

Phương trình (3) thực chất thuộc dạng mô hình trễ phân bố tự hồi qui [15], nó hàm ý rằng giá trị của chỉ số thị trường chứng khoán Y không chỉ phụ thuộc các biến PC_1, PC_2, ⋯, PC_k và đến trễ 5 của các biến này mà còn phụ thuộc vào trễ của chính nó. Độ dài trễ là 5 do các giao dịch trên thị trường chứng khoán được thực hiện theo 5 ngày làm việc của tuần.

11) Phương trình ước lượng cuối cùng không chỉ phải đảm bảo các tham số ước lượng đều có ý nghĩa thống kê mà còn phải ổn định (hay vững), phần dư là nhiễu trắng và không có hiện tượng nội sinh phần dư. Mô hình dự báo khi đó được xem là ổn định, có ước lượng không chệch và là tốt nhất theo phương pháp hồi qui bình phương tối thiểu [8, 15].

12) Thực hiện kiểm định mô hình dự báo bằng cách: sử dụng mô hình được xây dựng trên tập dữ liệu thứ nhất để dự báo tập dữ liệu thứ 2 và so sánh kết quả dự báo với số liệu thực tế trong tập thứ 2 để đánh giá chất lượng dự báo của mô hình.

13) Nếu sai số của dự báo là chấp nhận được theo quan điểm của người dùng thì ước lượng lại mô hình trên toàn bộ tập dữ liệu gốc và sử dụng nó để dự báo tương lai.

Để xây dựng mô hình dự báo chỉ số thị trường chứng khoán theo phương pháp vừa nêu có thể sử dụng các phần mềm công cụ sau: SAS, STATA, EVIEW hay R,… Bài báo này sử dụng phần mềm công cụ EVIEW [16].

IV. XÂY DỰNG MÔ HÌNH DỰ BÁO CHỈ SỐ VNINDEX

4.1. Tập dữ liệu được sử dụng

Dữ liệu về chỉ số thị trường chứng khoán VNINDEX và các biến dữ liệu giao dịch cổ phiếu được niêm yết trên sàn giao dịch Thành phố Hồ Chí Minh được thu thập theo tần suất ngày từ ngày 4/01/2010 đến ngày 5/5/2016, bao

gồm 278 biến dữ liệu giao dịch cổ phiếu kể cả chỉ số thị trường chứng khoán và 1574 ngày giao dịch (hay quan sát). Giá trị giao dịch của cổ phiếu bằng giá trung bình * khối lượng khớp lệnh của cổ phiếu này trong ngày.

4.2. Xây dựng mô hình dự báo

Trong phần này chỉ trình bày một cách rất tóm tắt việc xây dựng mô hình dự báo phương pháp được trình bày ở Mục 3 ở trên.

Bước 1: *Lựa chọn các biến dữ liệu giao dịch cổ phiếu có tương quan cao với chỉ số VNINDEX*

Do số lượng biến dữ liệu giao dịch cổ phiếu là khá lớn (277 mã cổ phiếu) nên ở Bước này phải tính tương quan mẫu giữa chỉ số VNINDEX với 277 biến dữ liệu giao dịch cổ phiếu để lựa chọn ra các biến dữ liệu giao dịch cổ phiếu có tương quan mẫu với chỉ số thị trường chứng khoán là cao.

Như đã biết nếu hệ số tương quan mẫu giữa 2 biến là dương thì hai biến này là biến đổi cùng chiều và ngược lại là biến đổi trái chiều bởi vậy việc lựa chọn ngưỡng cho hệ số tương quan mẫu cho các tương quan mẫu dương và âm nên là khác nhau.

Nếu chọn ngưỡng hệ số tương quan mẫu dương là 0.49 và ngưỡng hệ số tương quan mẫu âm là -0.20 thì sẽ có 25 các biến dữ liệu giao dịch cổ phiếu được chỉ ra trong Bảng 1. Tập dữ liệu về giá trị giao dịch của các biến này từ ngày 04/01/2010 đến ngày 05/5/2016 trở thành tập dữ liệu gốc.

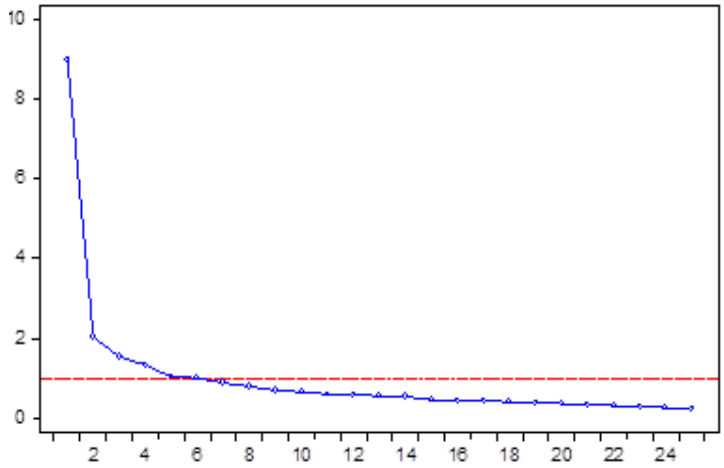
STT	Mã cổ phiếu	Hệ số tương quan mẫu	STT	Mã cổ phiếu	Hệ số tương quan mẫu
1	DLG	0.5996	14	MHC	0.5446
2	DXG	0.6279	15	PVD	0.5810
3	FLC	0.6353	16	PXS	0.5435
4	FPT	0.5612	17	TCM	0.6483
5	HAG	0.5208	18	TTF	0.6209
6	HAR	0.5298	19	VHG	0.5870
7	HCM	0.5963	20	BMC	-0.2312
8	HHS	0.4931	21	CLG	-0.2226
9	HPG	0.5581	22	LCM	-0.2023
10	HQC	0.5645	23	PNJ	-0.2650
11	HT1	0.5093	24	VMD	-0.2893
12	HVG	0.5385	25	VPK	-0.3180
13	IDI	0.4990			

Bảng 1. Các biến dữ liệu giao dịch cổ phiếu có tương quan mẫu cao với VNINDEX

Bước 2: *Tính toán và lựa chọn thành phần chính đại diện cho tập dữ liệu gốc*

Trước hết xây dựng ma trận tương quan **R** của 25 biến dữ liệu được nêu trong Bảng 1. Mỗi phần tử của **R** là hệ số tương quan mẫu giữa 2 biến bất kỳ trong 25 biến dữ liệu giao dịch cổ phiếu được chọn. Do đó **R** là ma trận vuông đối xứng cấp 25x25 và $\mathbf{RR}^T$ là dạng toàn phương xác định dương nên giá trị riêng của ma trận R sẽ là số dương.

Đồ thị về các giá trị riêng của ma trận R được trình bày trong Hình 2 ở dưới. Theo đó có 5 giá trị riêng có giá trị lớn hơn 1, còn lại nhỏ hơn 1. Theo [6], điều đó hàm ý rằng có thể chọn ít nhất 5 thành phần chính để nghiên cứu cấu trúc của tập dữ liệu gốc cũng như làm các biến giải thích trong phân tích hồi qui nhiều biến trên tập dữ liệu gốc.



Hình 2. Đồ thị các giá trị riêng được sắp thứ tự của ma trận tương quan **R**

Phân tích cụ thể và chi tiết hơn về các giá trị riêng ta sẽ thấy, ma trận R có 25 giá trị riêng. Tổng các giá trị riêng là 25. Giá trị riêng lớn nhất là 8.9921, giá trị riêng thứ 2 là: 2.0265. Giá trị riêng thứ nhất lớn hơn giá trị riêng thứ 2 là 6.9656. Tỷ lệ của giá trị riêng thứ nhất trên tổng các giá trị riêng là 0.3597 hàm ý rằng thành phần chính ứng với giá trị riêng này phản ánh được 35.97% sự thay đổi của tập dữ liệu gốc. Trong khi đó tỷ lệ giá trị riêng thứ 2 trên tổng các giá trị riêng là 0.0811 và thành phần chính ứng với giá trị riêng thứ 2 phản ánh được 8.11% sự thay đổi của tập dữ liệu gốc. Tỷ lệ của tổng hai giá trị riêng đầu tiên trên tổng các giá trị riêng (được gọi là giá trị tích lũy) là 11.0187 và tỷ lệ tích lũy của nó thể hiện hai thành phần chính đầu tiên phản ánh được bao nhiêu % sự thay đổi của tập dữ liệu gốc. Trong trường hợp này là 44.07%.

Như vậy ta có thể thấy 8 thành phần chính tương ứng với 8 giá trị riêng đầu tiên phản ánh được 70.28% sự thay đổi của tập dữ liệu gốc. Một số nhà nghiên cứu khác cho rằng [6] việc lựa chọn số thành phần chính cần sao cho nó phản ánh được từ 70% đến 90% những thay đổi của tập dữ liệu gốc.

Kết hợp hai gợi ý về lựa chọn số thành phần chính thay thế cho tập dữ liệu gốc trong [6], bài báo này đề xuất sử dụng 8 thành phần chính làm biến giải thích trong mô hình dự báo chỉ số thị trường chứng khoán VNINDEX.

Số giá trị riêng: 25.

Số thứ tự	Giá trị riêng	Chênh lệch giữa 2 giá trị riêng	Tỷ lệ trên tổng giá trị riêng	Giá trị riêng tích lũy	Tỷ lệ tích lũy
1	8.992132	6.965577	0.3597	8.992132	0.3597
2	2.026554	0.488001	0.0811	11.01869	0.4407
3	1.538553	0.212911	0.0615	12.55724	0.5023
4	1.325642	0.300307	0.0530	13.88288	0.5553
5	1.025336	0.027213	0.0410	14.90822	0.5963
6	0.998123	0.122096	0.0399	15.90634	0.6363
7	0.876027	0.088688	0.0350	16.78237	0.6713
8	**0.787339**	**0.098074**	**0.0315**	**17.56971**	**0.7028**
9	0.689266	0.040528	0.0276	18.25897	0.7304
10	0.648738	0.056668	0.0259	18.90771	0.7563
11	0.592070	0.021553	0.0237	19.49978	0.7800
12	0.570516	0.019259	0.0228	20.07030	0.8028
13	0.551257	0.021945	0.0221	20.62155	0.8249
14	0.529312	0.075979	0.0212	21.15087	0.8460
15	0.453333	0.017841	0.0181	21.60420	0.8642
16	0.435492	0.010644	0.0174	22.03969	0.8816
17	0.424848	0.024578	0.0170	22.46454	0.8986
18	0.400270	0.015855	0.0160	22.86481	0.9146
19	0.384414	0.028449	0.0154	23.24922	0.9300
20	0.355966	0.032409	0.0142	23.60519	0.9442
21	0.323556	0.013374	0.0129	23.92874	0.9571
22	0.310182	0.032247	0.0124	24.23893	0.9696
23	0.277935	0.024296	0.0111	24.51686	0.9807
24	0.253640	0.024141	0.0101	24.77050	0.9908
25	0.229499	---	0.0092	25.00000	1.0000

Bảng 2. Các giá trị riêng và mức độ phản ánh sự thay đổi của tập dữ liệu gốc

Bảng 3 bao gồm 8 véc tơ riêng tương ứng với 8 giá trị riêng đầu tiên. Mỗi véc tơ riêng trở thành trọng số của thành phần chính tương ứng.

Thực hiện chuẩn hóa 25 biến dữ liệu giao dịch cổ phiếu (25 cột trong ma trận A) theo công thức (2) và thực hiện tính toán 8 thành phần chính PC_1, PC_2, …, PC_8 theo công thức (1) với các véc tơ trọng số được cho tương ứng với các thành phần chính (Bảng 3) ta sẽ nhận được 8 thành phần chính đầu tiên. Các thành phần khi đó đã được sắp thứ tự theo độ lớn của các giá trị riêng của chúng.

Mã cổ phiếu	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
DLG	0.23	-0.10	-0.18	-0.06	-0.23	0.30	-0.05	-0.02
DXG	0.24	-0.04	-0.05	0.10	-0.15	0.09	0.00	0.28
FLC	0.26	0.05	-0.04	-0.14	-0.03	-0.04	0.07	-0.31
FPT	0.21	-0.01	0.19	0.32	-0.22	-0.02	0.19	-0.02
HAG	0.21	0.15	0.22	0.09	0.09	-0.11	-0.33	0.17
HAR	0.21	0.15	-0.03	-0.30	-0.14	-0.08	-0.47	-0.06

Mã cổ phiếu	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
HCM	0.23	0.05	0.11	0.36	0.03	-0.18	0.07	0.08
HHS	0.18	-0.23	-0.28	0.06	-0.12	0.36	0.11	0.14
HPG	0.20	-0.02	0.02	0.42	0.12	0.08	-0.11	0.02
HQC	0.23	0.15	0.15	0.00	-0.30	0.07	-0.15	-0.28
HT1	0.21	0.12	0.13	0.08	-0.13	-0.11	0.12	-0.53
HVG	0.24	0.07	0.09	-0.18	0.36	-0.13	0.20	0.15
IDI	0.23	0.13	0.00	-0.26	0.15	0.20	0.21	0.01
MHC	0.21	-0.15	-0.06	-0.04	0.15	0.21	0.08	0.06
PVD	0.24	-0.01	-0.03	-0.02	0.16	-0.01	0.06	0.00
PXS	0.25	0.10	0.07	-0.17	0.21	-0.07	0.38	-0.06
TCM	0.24	-0.06	0.11	0.20	-0.03	-0.24	0.11	0.23
TTF	0.22	-0.07	-0.14	0.14	0.00	0.18	-0.30	0.16
VHG	0.24	-0.03	-0.08	-0.36	0.17	0.04	-0.17	0.12
BMC	-0.08	0.52	-0.16	0.14	0.04	0.24	0.09	0.11
CLG	-0.04	0.18	0.51	-0.23	-0.25	-0.01	0.01	0.46
LCM	-0.04	0.49	0.17	0.06	0.02	0.43	-0.06	-0.08
PNJ	-0.09	-0.13	0.33	0.17	0.59	0.28	-0.25	-0.17
VMD	-0.09	-0.26	0.40	-0.08	-0.18	0.42	0.27	-0.01
VPK	-0.09	0.40	-0.34	0.11	0.04	0.00	0.19	0.13

Bảng 3. Các véc tơ riêng tương ứng với 8 giá trị riêng đầu tiên

Bước 3: *Xây dựng mô hình dự báo tần suất cao*

Thực hiện kiểm định Dickey – Fuller tăng cường về nghiệm đơn vị của biến VNINDEX và các thành phần chính ta thấy VNINDEX dừng sai phân bậc 1, còn tất các các thành phần chính PC1, PC2,…, PC8 đều là chuỗi dừng. Do đó ta có thể hồi qui sai phân bậc 1 của VNINDEX theo 8 thành phần chính đầu tiên.

Chia tập số gốc thành 2 tập, tập thứ nhất gồm dữ liệu giao dịch từ các ngày 04/1/2010 đến 22/4/2016, tức có 1567 quan sát, tập thứ 2 gồm dữ liệu giao dịch của các ngày từ 25/4/2016 đến các ngày 05/5/2016, gồm 7 quan sát.

Thực hiện quá trình lặp phép hồi qui VNINDEX theo các biến PC1, PC2, …, PC8 trên tập dữ liệu thứ nhất dựa theo công thức (3) bằng cách loại trừ dần các biến mà hệ số ước lượng của nó không có ý nghĩa thông kê, ta nhận được phương trình ước lượng sau:

Biến phụ thuộc D(VNINDEX)

Số quan sát: 1556 (sau khi điều chỉnh)

Biến	Hệ số	Sai số chuẩn	Mức ý nghĩa
D(VNINDEX(-1))	0.270726	0.025290	***
D(VNINDEX(-2))	-0.077034	0.026149	***
D(VNINDEX(-3))	0.070167	0.026060	***
D(VNINDEX(-4))	-0.073430	0.024966	***
PC1(-2)	-4.00E-09	2.49E-09	*
PC2	5.62E-08	1.99E-08	***
PC2(-2)	-3.79E-08	1.74E-08	**
PC3	6.21E-08	1.66E-08	***
PC3(-5)	-5.22E-08	1.50E-08	***
PC6	7.06E-08	1.67E-08	***
PC6(-5)	-5.10E-08	1.66E-08	***

R^2 = 0.11; Thống kê Durbin – Watson: 1.99477;

Bảng 4. Phương trình ước lượng (hay mô hình dự báo) VNINDEX theo các thành phần chính
Các ký hiệu: *, ** và *** trong Cột 4 (Bảng 4) tương ứng là các mức ý nghĩa thống kê dưới 10%, 5% và 1%.

Phương trình ước lượng (hay mô hình dự báo) dưới dạng tường minh có dạng:

D(VNINDEX) = 0.271*D(VNINDEX(-1)) - 0.077*D(VNINDEX(-2)) + 0.070*D(VNINDEX(-3)) - 0.073*D(VNINDEX(-4)) - 3.995*PC1(-2) + 5.625e-08*PC2 - 3.793e-08*PC2(-2) + 6.208e-08*PC3 - 5.218e-08*PC3(-5) + 7.059e-08*PC6 - 5.101e-08*PC6(-5)

Thực hiện các kiểm định Ramsey về tính ổn định của mô hình, kiểm định Jarque-Bera về phần dư có phân phối chuẩn và có kỳ vọng bằng 0, kiểm định Breusch-Godfrey về không có hiện tượng nội sinh phần dư và phần dư không tự tương quan, kiểm định Breusch-Pagan [14] về phương sai phần dư không đổi ta nhận được các kiểm định đều thỏa mãn.

Đánh giá chất lượng dự báo của mô hình bằng cách sử dụng mô hình được xây dựng để dự báo chỉ số VNINDEX ở 7 phiên giao dịch tiếp theo.

Giá trị của chỉ số VNINDEX thực tế và kết quả dự báo chỉ số này bằng sử dụng mô hình dự báo được trình bày trong Bảng 5. Phần trăm sai số tuyệt đối nói chung là khá nhỏ dưới 1%, trừ thứ 2 ngày 25/04/2016, phần trăm sai số tuyệt đối là trên 1%, nó thể hiện trong những ngày nghỉ cuối tuần có thể đã có những thông tin không tích cực đến thị trường chứng khoán và đã tác động xấu tới tâm lý, niềm tin thị trường của các nhà đầu tư.

Thứ	Ngày	VNINDEX	VNINDEXF	% sai số
Thứ 2	25/04/2016	596.60	586.84	-1.63
Thứ 3	26/04/2016	594.35	599.70	0.90
Thứ 4	27/04/2016	596.60	594.02	-0.43
Thứ 5	28/04/2016	593.65	597.10	0.58
Thứ 6	29/04/2016	595.45	592.16	-0.55
Thứ 4	04/05/2016	597.75	597.52	-0.04
Thứ 5	05/05/2016	602.15	600.47	-0.28

Bảng 5. So sánh kết quả dự báo bằng mô hình với số liệu thực tế

V. KẾT LUẬN

Bài báo này trình bày một cách khá tóm tắt phương pháp xây dựng mô hình dự báo chỉ số thị trường chứng khoán bằng kết hợp sử dụng phương pháp hồi qui nhiều biến và kỹ thuật PCA trên tập dữ liệu gốc để tạo ra một số biến mới thay thế các biến trong tập dữ liệu này làm biến giải thích trong mô hình hồi qui nhiều biến.

Thực hành phương pháp trên tập số liệu thực tế của sàn giao dịch chứng khoán Thành phố Hồ Chí Minh cho thấy mô hình dự báo có độ chính xác khá cao, mặc dù các thành phần chính được sử dụng trong xây dựng mô hình chỉ phản ánh 70,2% sự thay đổi của tập dữ liệu gốc gồm 25 biến dữ liệu giao dịch cổ phiếu có hệ số tương quan mẫu cao với chỉ số thị trường chứng khoán của sàn giao dịch.

Nếu tăng số lượng biến dữ liệu giao dịch cổ phiếu trong tập dữ liệu thay thế tập dữ liệu ban đầu hoặc tăng số thành phần chính để nó phản ánh được tỷ lệ % cao hơn sự thay đổi của tập dữ liệu gốc thì chất lượng dự báo bằng sử dụng mô hình sẽ được cải thiện hơn.

Phương pháp dự báo chỉ số thị trường chứng khoán trong bài báo này có thể được ứng dụng để dự báo nhiều sự kiện kinh tế - xã hội khác trong đó nhất là dự báo các chỉ tiêu kinh tế vĩ mô.

TÀI LIỆU THAM KHẢO

[1] Andrei Roudoi (2009), "Short-term forecasting of key indicators of the German economy", in The Making of National Economic Forecasts, Edited by Lawrence R. Klein, Published by Edward Elgar, Cheltenham, UK • Northampton, MA, USA, pp. 121-148.

[2] Carol Anne Hargreaves, Chandrika Kadirvel Mani (2015), "The Selection of Winning Stocks Using Principal Component Analysis", American Journal of Marketing Research, Vol. 1, No. 3, 2015, pp. 183-188.

[3] Đỗ Văn Thành và Nguyễn Minh Hải (2016), "Phân tích và dự báo chỉ số thị trường chứng khoán bằng sử dụng chỉ số báo trước ", Báo cáo Hội nghị FAIR, 2015, Cần Thơ, 04/8 – 5/8/2016, 8 trang.

[4] Lawrence R. Klein (2009), "Background to national economic forecasts and the high-frequency model of the USA", in The Making of National Economic Forecasts, Edited by Lawrence R. Klein, Published by Edward Elgar, Cheltenham, UK • Northampton, MA, USA, pp.1-26.

[5] Lawrence R. Klein, "Background to national economic forecasts and the high-frequency model of the USA", in The Making of National Economic Forecasts, Edited by Lawrence R. Klein, Published by Edward Elgar, Cheltenham, UK • Northampton, MA, USA, 2009, 403 pages.

[6] Lindsay I Smith (2002), A tutorial on Principal Components Analysis.

[7] Mbeledogu, N. N., Odoh, M., Umeh, M.N. (2012). "Stock feature extraction using Principle Component Analysis". International Conference on Computer Technology and Science. IACSIT Press, Singapore DOI: 10.7763/IPCSIT, 2012, V47, 44.

[8] Graham E., Granger C.W.J., Timmerman A. (2006), *Hanbook of Economic Forecasting*, Volume 1, Elsevier BV, 2006, 933 p.

[9] Preethi, G. and Santhi, B.: "Stock Market Forecasting Techniques: A Survey", Journal of theoretical and Applied Information technology, Vol 46, No 1, 2012, pp. 24-30.

[10] Yacine A¨ıt-Sahalia and Dacheng Xiu, "Principal Component Analysis of High Frequency Data", Working paper, Princeton University and University of Chicago, 47 pages, March 2015.

[11] Yoshihisa Inada (2009), "A high-frequency forecasting model and its application to the Japanese economy", in The Making of National Economic Forecasts, Edited by Lawrence R. Klein, Published by Edward Elgar, Cheltenham, UK • Northampton, MA, USA, 172-197.

[12] Vladimir Eskin and Mikhail Gusev (2009), "High-frequency forecasting model for the Russian economy, in The Making of National Economic Forecasts, Edited by Lawrence R. Klein, Published by Edward Elgar, Cheltenham, UK • Northampton, MA, USA, pp. 93-120.

[13] Wang, Z., Sun, Y., Stockli, P. (2014). "Functional Principal Components Analysis of Shanghai Stock Exchange 50 Index". Discrete Dynamics in Nature and Society Volume 2014 (2014), Article ID 365204, 7 pages.

[14] Wang Yanshan, In-Chan Choi. (2013)." Market Index and stock price direction prediction using Machine Learning Techniques: An empirical study on the KOSPI and HSI". Science Direct, pp. 1-13.

[15] William H. Greene, "Economic Analysis", New York University, Seventh Edition, Prentice Hall, 2012.

[16] www.eviews.com.

HIGH-FREQUENCY FORECAST MODEL FOR STOCK MARKET INDEX

Thanh Do Van and Hai Nguyen Minh

ABSTRACT — *The paper proposed a methodology of combining techniques of multivariate regression and principal component analysis (PCA) on the stocks listed on a stock exchange and had high pattern correlations with the stock market index of this stock exchange to build a forecast model of stock market index. The practising to build forecast model of stock market index is implemented on real data from early 2010 to the present of Ho Chi Minh City's Stock Exchange. The forecasted results using the built model show good prospects of the proposed methodology for building forecast models of stock market index as well as others socio-economic indicators.*