Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Đà Nẵng, ngày 17-18/08/2017
DOI: 10.15625/vap.2017.00051

MÔ HÌNH HÓA DỰ BÁO GIÁ CỔ PHIẾU TRONG NGỮ CẢNH DỮ LIỆU SỐ CHIỀU CAO

Đỗ Văn Thành

Khoa Công nghệ thông tin - Trường Đại học Nguyễn Tất Thành, dvthanh@ntt.edu.vn

TÓM TẮT *- Dự báo giá cổ phiếu luôn được quan tâm đặc biệt và luôn được xem là một trong những loại dự báo khó nhất trong lĩnh vực kinh tế - tài chính do tính dễ thay đổi và biến động khó lường của nó. Mục đích của bài báo này là trình bày việc mô hình hóa dự báo giá của một cổ phiếu nào đó theo tập tất cả các biến kinh tế - tài chính có ảnh hưởng đến sự biến động của giá cổ phiếu đó. Các biến này không hoàn toàn độc lập với nhau và số lượng các biến cũng như số lượng các quan sát theo mỗi biến nói chung là rất lớn.*

Phương pháp xây dựng mô hình dự báo giá cổ phiểu được đề xuất trong bài báo này sẽ sử dụng kết hợp kỹ thuật lựa chọn thuộc tính và học thuộc tính để chuyển tập dữ liệu số chiều cao về tập dữ liệu số chiều thấp nhưng cơ bản vẫn giữ được khá đầy đủ thông tin trong tập dữ liệu số chiều cao và bảo toàn được quan hệ giữa biến giá cổ phiếu với các biến kinh tế - tài chính khác nhiều như có thể. Bài báo cũng sử dụng mô hình trễ phân bố tự hồi quy để xây dựng mô hình dự báo trung bình của giá cổ phiếu và sử dụng một trong các mô hình thuộc họ các mô hình phương sai thay đổi điều kiện tự hồi quy để dự báo tính không chắc chắn của phương sai phần dư của mô hình dự báo. Kết quả dự báo bằng mô hình được xây dựng theo phương pháp được đề xuất cho thấy triển vọng tốt của phương pháp này và nó có thể được xem là hướng dẫn cụ thể cho việc thực hành mô hình hóa dự báo giá của các hàng hóa và dịch vụ khác.

Từ khóa *- dữ liệu số chiều cao, giảm chiều dữ liệu, giá cổ phiếu, mô hình ARCH, mô hình hóa dự báo tài chính.*

I. GIỚI THIỆU VẤN ĐỀ

Dự báo thị trường chứng khoán gồm 2 nội dung quan trọng nhất là dự báo giá trị của chỉ số chứng khoán và giá của các cổ phiếu được niêm yết trên thị trường [17]. So với dự báo chỉ số chứng khoán thì dự báo giá cổ phiếu nhìn chung là khó khăn hơn bởi sự dễ thay đổi của nó.

Do có quá nhiều yếu tố tác động đến giá hàng hóa và giá dịch vụ nói chung, chỉ số chứng khoán và giá cổ phiếu nói riêng nên có một thời gian rất dài người ta cho rằng không thể dự báo được giá. Đến năm 1978, người ta nhận thấy khẳng định trên là đúng khi thị trường hoạt động hiệu quả, trong thị trường hoạt động không hiệu quả thì có thể dự báo được giá một phần do các yếu tố tâm lý của những người tham gia thị trường cùng với khả năng thị trưởng không thể phản ứng được ngay với những thông tin mới được công bố [16].

Hiện tại đã có khá nhiều kỹ thuật được ứng dụng trong xây dựng mô hình dự báo giá cố phiếu của thị trường chứng khoán [6, 17, 22]. Các kỹ thuật dự báo chỉ số giá cổ phiếu có thể được phân thành 2 nhóm theo 2 cách tiếp cận khác nhau [22] là nhóm các kỹ thuật thống kê và nhóm các kỹ thuật trí tuệ nhân tạo.

- Các kỹ thuật dự báo thống kê nói chung thường đòi hỏi các biến phải được đưa về chuỗi dừng trước khi ứng dụng nó và các kỹ thuật này yêu cầu phải thực hiện rất nhiều kiểm định thống kê khác nhau nhằm chẩn đoán, khắc phục và đánh giá chất lượng của mô hình trước khi tiến hành dự báo tương lai. Ưu điểm chính của các kỹ thuật dự báo thống kê là đưa ra được giá trị dự báo tương lai một cách cụ thể và nếu tương lai không có những biến động bất thường so với hiện tại và quá khứ thì độ chính xác của dự báo được thực hiện bằng những kỹ thuật này thường khá cao. Các kỹ thuật dự báo thống kê có thể xem xét và phân tích hành vi, phát hiện và xử lý tốt các dữ liệu ngoại lai, cung cấp một cách tường minh về hàm dự báo và cho biết một cách rõ ràng các quan hệ giữa các yếu tố đầu vào và biến đích đầu ra. Trong lĩnh vực kinh tế - xã hội các mối quan hệ giữa các yếu tố đầu vào và biến đích đầu ra là hàm ý những quy luật kinh tế đặc thù, chúng gợi ý những phản ứng chính sách cần có để tận dụng hoặc giảm nhẹ tác động của những quy luật ấy. Trong điều hành và quản lý nền kinh tế, việc phát hiện được những quy luật kinh tế đặc thù nói chung được xem trọng hơn so với việc đưa ra những kết quả dự báo cụ thể. Nhược điểm chính của các kỹ thuật dự báo thống kê là khó tự động hóa được toàn bộ quá trình dự báo và không thể thực hiện được trên các tập dữ liệu số chiều cao. Để xây dựng được mô hình dự báo trên tập dữ liệu có số chiều cao trước hết phải chuyển tập dữ liệu số chiều cao về tập dữ liệu số chiều thấp nhưng cơ bản phải giữ được khá đầy đủ thông tin trong tập dữ liệu số chiều cao và bảo toàn được quan hệ giữa biến đích đầu ra với các biến gốc đầu vào nhiều như có thể.

- Các kỹ thuật dự báo trí tuệ nhân tạo (như mạng nơtron, hệ suy luận nơtron-mờ, giải thuật di truyền, luật kết hợp, khai phá mẫu chuỗi, k- người láng giềng gần nhất, mạng Bayes,...) là những kỹ thuật phi tuyến, chủ yếu được sử dụng để dự báo phân lớp dữ liệu. Các kỹ thuật này không đòi hỏi các biến dữ liệu đầu vào phải dừng và nói chung không cần thực hiện các kiểm định thống kê. Ưu điểm chính của các kỹ thuật trí tuệ nhân tạo là có thể thực hiện được trên các tập dữ liệu đầu vào rất lớn, có thể tự động được toàn bộ quá trình dự báo, kết quả dự báo phân lớp nói chung cũng có độ chính xác tương đối cao. Nhược điểm chính là chỉ thích hợp với dự báo xu thế, khó đưa ra được những giá trị dự báo cụ thể hoặc nếu có thì hoặc độ chính xác dự báo là không cao hoặc phải thêm rất nhiều phí tổn (nhất là thời gian) để nâng cao độ chính xác dự báo. Các kỹ thuật dự báo trí tuệ nhân tạo hạn chế trong việc phân tích và xử lý hành

vi, phát hiện và xử lý dữ liệu ngoại lai và đặc biệt là chúng là những kỹ thuật hộp đen, hàm dự báo chưa được chỉ ra một cách tường minh và chưa cung cấp được các mối quan hệ cụ thể giữa các biến gốc đầu vào với biến đích đầu ra nên không biết được từng yếu tố đầu vào đã tác động mạnh, yếu thế nào đến sự thay đổi của biến đích. Đến thời điểm này, cho dù đã có rất nhiều nghiên cứu và thực nghiệm nhưng có thể nói các kỹ thuật dự báo trí tuệ nhân tạo mới phù hợp để phân tích, dự báo dữ liệu khoa học, chưa phù hợp để phân tích, dự báo dữ liệu kinh tế - xã hội nói chung, dữ liệu tài chính - kinh tế nói riêng, ở đó hành vi của các tác nhân kinh tế có ảnh hưởng rất lớn đến kết quả dự báo. Mặc dù các kỹ thuật trí tuệ nhân tạo có thể xử lý được tập dữ liệu rất lớn, nhưng một phần tập dữ liệu đầu vào có thể có lỗi, có thể chứa dữ liệu ngoại lai, dữ liệu không liên quan cũng như dữ liệu dư thừa và phần khác nhằm để tăng hiệu quả xử lý và nâng cao chất lượng phân lớp dữ liệu, việc thực hiện giảm chiều dữ liệu trước khi thực hiện các kỹ thuật trí tuệ nhân tạo để phân lớp vẫn là rất cần thiết.

Những phân tích ở trên cho thấy để đưa ra được giá trị dự báo cụ thể có độ chính xác cao, có thể ứng dụng được trong thế giới thực thì cần sử dụng kỹ thuật dự báo thống kê. Trong dự báo bằng kỹ thuật thống kê cũng như bằng kỹ thuật trí tuệ nhân tạo, điểm mấu chốt nhất để nâng cao độ chính xác của dự báo là xử lý tốt sai số (hay phần dư) của mô hình dự báo. Để xử lý phần dư của mô hình dự báo thống kê người ta thường xem nó như là mô hình trung bình trượt tự hồi quy (ARMA), song như thế vẫn chưa đủ vì thế trong rất nhiều trường hợp người ta phải thực hiện nhiều kỹ thuật xử lý khác nữa [12]. Năm 1982, Engle, R. F. đã phát hiện ra một nguyên nhân rất quan trọng có tác động đến sự dễ thay đổi (hay tính không chắc chắn) của phần dư, đó là hiện tượng phần dư có phương sai thay đổi điều kiện (gọi tắt là hiện tượng ARCH). Bài báo [9] đã đề xuất mô hình phương sai thay đổi điều kiện tự hồi quy ARCH(p) để dự báo phương sai phần dư của mô hình dự báo. Hiện đã hình thành một họ các mô hình ARCH và tùy theo vấn đề cụ thể cần thực hiện một số kiểm định thống kê, để so sánh lựa chọn một mô hình họ ARCH phù hợp nhất.

Trong trường hợp dự báo giá cổ phiếu thì phần dư chính là kỳ vọng lợi nhuận của đầu tư cổ phiếu nên các mô hình họ ARCH được xem là những mô hình để dự báo tính không chắc chắn của lợi nhuận (hay lợi nhuận kỳ vọng) của đầu tư. Họ mô hình ARCH đã được ứng dụng trong việc dự báo lợi nhuận đầu tư vào thị trường Mỹ và tác giả của bài báo [9] đã được trao giải Nobel kinh tế năm 2003 về những đóng góp này. Hiện nay trong lĩnh vực kinh tế - tài chính họ mô hình ARCH rất được quan tâm ứng dụng. Điều đó gợi ý rằng trong thế giới thực nên lựa chọn họ mô hình ARCH để dự báo tính không chắc chắn của phần dư (hay sai số) khi dự báo biến đích trong ngữ cảnh số lượng các biến gốc tiềm năng có tác động đến biến đích cũng như số lượng quan sát của các biến là rất lớn.

Kỹ thuật giảm chiều dữ liệu là làm giảm số lượng các biến gốc (gọi là giảm chiều biến) và/hoặc giảm số lượng quan sát (gọi là giảm chiều quan sát). Hiện đã có khá nhiều kỹ thuật giảm chiều dữ liệu, trong đó nhất là các kỹ thuật giảm chiều biến. Kỹ thuật giảm chiều biến bao gồm 2 loại: Lựa chọn biến (hay Lựa chọn thuộc tính) và Chiết xuất biến (Chiết xuất thuộc tính hay Học thuộc tính). *Lựa chọn thuộc tính* là trích xuất một vài thuộc tính để đại diện cho tập dữ liệu ban đầu [3, 13, 15] trong khi *Học thuộc tính* là kết hợp một số thuộc tính ban đầu để tạo ra các thuộc tính mới nhưng không làm thay đổi các biểu diễn ban đầu của các biến dữ liệu [5, 15].

Lựa chọn thuộc tính được phân theo 3 phương pháp tiếp cận [3, 18]: Phương pháp tiếp cận bộ lọc (Filter): Trước tiên lựa chọn tập con thuộc tính và sau đó sử dụng tập con này để thực hiện thuật toán phân lớp hoặc dự báo. Phương pháp tiếp cận nhúng (Embeded): Việc lựa chọn thuộc tính xuất hiện như là một phần của thuật toán phân lớp/dự báo mà không chia tách tập dữ liệu dầu vào thành tập dữ liệu huấn luyện và thử nghiệm. Phương pháp tiếp cận bọc (Wrapper): thuật toán phân lớp/dự báo được áp dụng trên toàn thể tập dữ liệu ban đầu nhằm xác định các thuộc tính khi đó tiêu chí lựa chọn thuộc tính là thành tích của thuật toán phân lớp/dự báo [3].

Trong rất nhiều kỹ thuật giảm chiều dữ liệu được biết, các kỹ thuật thuộc họ phân tích thành phần chính (PCA) như: Phân rã phương sai đơn (SVD), Phân tích thành phần chính tuyến tính (PCA), Phân tích thành phần chính mờ mạnh (RFPCA), Phân tích thành phần chính hạt nhân (KPCA),… vẫn được xem là hiệu quả nhất [3, 14, 20]. Cụ thể trong bài báo [20] các tác giả đã so sánh kỹ thuật Phân tích thành phần chính tuyến tính (PCA) với 12 kỹ thuật giảm chiều phi tuyến hàng đầu như: Multidimensional Scaling, Isomap, Maximum Variance Unfolding, kernel *PCA*, Diffusion Maps, Multilayer Autoencoders, Locally Linear Embedding, Laplacian Eigenmaps, Hessian LLE, Local Tangent Space Analysis, Locally Linear Coordination và Manifold Charting bằng cách thực nghiệm chúng trên các tập dữ liệu nhân tạo và tập dữ liệu thực. Kết quả cho thấy mặc dù 12 kỹ thuật phi tuyến có thể giảm chiều tốt trên các tập dữ liệu nhân tạo được chọn, nhưng với các tập dữ liệu trong thế giới thực thì không có kỹ thuật nào trong số 12 kỹ thuật đã nêu làm giảm chiều tốt hơn so với PCA truyền thống.

Hiện đã có tới hàng trăm kỹ thuật dự báo thị trường chứng khoán nói chung và dự báo giá cổ phiếu nói riêng [6-7, 18], nhưng những nghiên cứu liên quan đến dự báo giá cố phiếu trong ngữ cảnh dữ liệu có số chiều cao còn khá ít.

Bài báo mới đây [22] về dự báo lợi nhuận của thị trường chứng khoán theo ngày bằng cách sử dụng kỹ thuật PCA và 02 kỹ thuật PCA phi tuyến khác là phân tích thành phần chính mờ mạnh (RFPCA) và phân tích thành phần chính hạt nhân (KPCA) để giảm chiều của tập dữ liệu gồm 60 biến và sử dụng kỹ thuật mạng nơtron nhân tạo (ANN) để phân lớp. Bài báo đó đã chỉ ra rằng PCA+ANN cho kết quả dự báo phân lớp tốt hơn so với RFPCA+ANN và KPCA+ANN. Mặc dù kết quả dự báo phân lớp được đánh giá là độ chính xác khá cao nhưng vẫn hạn chế vì nó chỉ cho biết xu hướng lợi nhuận của thị trường mà không đưa ra được giá trị cụ thể. Phương pháp giảm chiều ở bài báo này cũng có 02 hạn chế đó là: khi các điểm dữ liệu của các biến gốc không xấp xỉ thuộc về một siêu phẳng và tổng quát

hơn là xấp xỉ thuộc về một đa tạp (manifold), hoặc khi số lượng các biến gốc là rất lớn thì việc sử dụng phương pháp giảm chiều PCA là không hiệu quả hoặc gặp nhiều khó khăn.

Bài báo [21] đã dự báo chỉ số giá cổ phiếu tổng hợp của trị trường chứng khoán Hàn Quốc (KOSPI) và chỉ số chứng khoán Hangseng (HSI) bằng cách sử dụng kỹ thuật phân tích thành phần chính (PCA) và học máy véctơ hỗ trợ (SVM) để giảm các điểm dữ liệu và để phân chúng thành hai lớp. Phân tích hai lớp này bài báo nhận thấy rằng có thể hình thành một cụm các cổ phiếu cùng thay đổi bằng việc sử dụng các thành phần chính được tạo ra từ PCA. Bài báo này cũng có nhược điểm chính tương tự như [22].

Bài báo [6] đã đề xuất sử dụng quan hệ nhân quả để giảm chiều biến của tập dữ liệu gồm 277 biến kinh tế - tài chính và sử dụng mô hình trễ phân bố tự hồi quy (ADL) được ước lượng theo phương pháp hồi quy nhiều biến để dự báo chỉ số chứng khoán VNINDEX theo ngày. Độ chính xác dự báo là khá cao. Ưu điểm chính của phương pháp này là có thể nhận được giá trị dự báo của VNINDEX mà không cần phải dự báo các biến ngoại sinh có trong mô hình. Nhược điểm chính của bài báo này là chỉ có một số ít biến gốc được đưa vào mô hình dự báo, điều đó cũng có nghĩa là chất lượng dự báo bằng mô hình có thể bị suy giảm do còn nhiều yếu tố ảnh hưởng đến sự thay đổi của VNINDEX chưa được đưa vào mô hình. Nhược điểm khác của bài báo này là các quan hệ nhân quả thường là quan hệ ngắn hạn, dễ thay đổi khi số quan sát của các biến được tăng thêm, nên việc xác định lại quan hệ nhân quả và xây dựng lại mô hình dự báo phải được thực hiện thường xuyên.

Bài báo [7] đã sử dụng kỹ thuật xếp hạng các biến gốc theo hệ số tương quan của chúng với biến đích để giảm số biến lần đầu và sau đó sử dụng kỹ thuật PCA để giảm tiếp chiều biến của tập dữ liệu sau lần giảm đầu và cuối cùng sử dụng mô hình ADL được ước lượng theo phương pháp hồi quy nhiều biến để dự báo chỉ số VNINDEX theo ngày. Độ chính xác dự báo theo phương pháp này tốt hơn so với phương pháp được đề xuất trong [6]. Tuy nhiên Bài báo này vẫn còn 2 nhược điểm chính. Thứ nhất là chưa thực hiện kiểm định để biết phần dư có phương sai thay đổi điều kiện hay không? Nếu có thì khi có những cú sốc tác động đến thị trường chứng khoán (như tình hình thị trường tài chính thế giới thay đổi, chính sách tiền tệ, lãi suất của chính phủ thay đổi,…) phần dư của mô hình sẽ thay đổi đột ngột trong khi mô hình dự báo trung bình không nắm bắt được, dẫn đến hạn chế độ chính xác dự báo. Nhược điểm thứ 2 là: trong số các biến gốc có hệ số tương quan cao với biến đích được lựa chọn lần đầu để sau đó áp dụng kỹ thuật PCA có thể có một số biến có tương quan cao với nhau, khi đó xảy ra hiện tượng một số biến gốc có thể được xác định thông qua một số biến gốc khác. Điều này có nghĩa là có sự dư thừa các biến được lựa chọn lần đầu và có thể đã bỏ sót một số biến thích đáng khác cung cấp thông tin có ích cho dự báo biến đích mặc dù hệ số tương quan của nó với biến đích là không lớn lắm.

Bài báo [19] đã sử dụng kỹ thuật xếp hạng các biến gốc là nguyên nhân có ý nghĩa thống kê cao trong quan hệ nhân quả giữa biến gốc và biến đích để giảm số biến lần đầu và sau đó sử dụng kỹ thuật PCA để giảm tiếp chiều biến của tập dữ liệu gốc gồm hơn 310 biến và cuối cùng sử dụng mô hình ADL được ước lượng theo phương pháp hồi quy nhiều biến để dự báo chỉ số VNINDEX theo tháng. Ưu điểm của phương pháp trong bài báo này bao gồm ưu điểm của cá 2 bài báo vừa nêu trên và nhược điểm chính cũng tương tự như nhược điểm của bài báo [7].

Bài báo này sẽ khắc phục các nhược điểm chính của tất cả các bài báo đã nêu ở trên. Cụ thể bài báo sẽ đề xuất khung lý thuyết để dự báo giá cổ phiếu trong ngữ cảnh số chiều biến là rất lớn và ứng dụng khung lý thuyết này trong việc dự báo giá cổ phiếu trên tập dữ liệu thực của nền kinh tế.

Khác với các phương pháp giảm chiều biến trong các nghiên cứu trước đó là được thực hiện theo một trong hai cách khác nhau đó là: sử dụng kỹ thuật Lựa chọn thuộc tính hoặc Học thuộc tính để tạo ra một nhóm các biến nhỏ hơn thay thế cho các biến gốc đầu vào [3], bài báo này đề xuất kết hợp cả hai phương pháp: Lựa chọn thuộc tính và Học thuộc tính trong việc làm giảm chiều dữ liệu trong bối cảnh phải đảm bảo yêu cầu giữ được nhiều nhất có thể quan hệ giữa biến đích và các biến gốc.

Để dự báo giá cổ phiếu, 02 mô hình dự báo thống kê sẽ được sử dụng. Mô hình trễ phân bố tự hồi quy (ADL) [12] được ước lượng bằng sử dụng kỹ thuật hồi quy nhiều biến để dự báo giá trung bình của cổ phiếu. Mô hình đó được gọi là mô hình dự báo trung bình. Trong mô hình này các biến giải thích và các biến trễ của chúng cũng như các biến trễ của biến đích đều được đưa vào. Điều đó hàm ý rằng sự thay đổi của biến đích không chỉ phụ thuộc vào các biến giải thích mà còn phụ thuộc vào quá khứ của chính nó và quá khứ của các biến giải thích. Mô hình phương sai thay đổi điều kiện tự hồi quy *GARCH(p,q)* [2, 8] *mở rộng* để dự báo phương sai phần dư của Mô hình dự báo trung bình nếu như phần dư có hiện tượng ARCH. Mô hình GARCH là một trong những mô hình thuộc họ ARCH được sử dụng phổ biến nhất. Các mô hình dự báo trung bình và mô hình dự báo phương sai được kết nối với nhau và được ước lượng đồng thời.

Kết quả dự báo giá cổ phiếu bằng mô hình được xây dựng theo phương pháp được đề xuất một mặt khẳng định ý nghĩa của khung lý thuyết này trong việc dự báo giá cổ phiếu, và mặt khác quan trọng hơn, nó có thể được xem là hướng dẫn cho việc mô hình hóa dự báo giá của rất nhiều loại hàng hóa và dịch vụ khác.

Bài báo này được cấu trúc như sau: tiếp theo phần này, phần II tiếp theo sẽ trình bày rõ hơn về vấn đề đặt ra và đề xuất phương pháp giải quyết. Phần III sẽ ứng dụng phương pháp được đề xuất để dự báo giá cổ phiếu trên tập dữ liệu thực của nền kinh tế và cuối cùng là một vài kết luận.

II. XÁC ĐỊNH VẤN ĐỀ VÀ PHƯƠNG PHÁP GIẢI QUYẾT

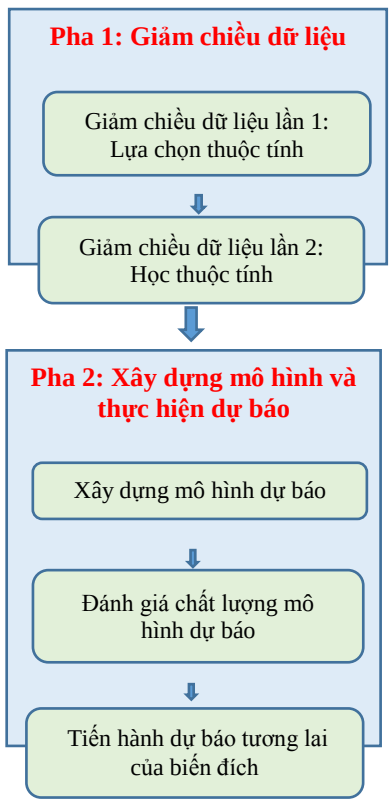
2.1. Xác định vấn đề

Ký hiệu Y là biến giá của một cổ phiếu nào đó (gọi là biến đích), X_i (i =1, 2,…, n) là các biến phản ánh các yếu tố có tác động đến sự thay đổi của Y (gọi là biến gốc); Y và các X_i (i =1, 2,…, n) đều thuộc không gian R^m. Nói cách khác Y, X_j là các một biến véctơ, $Y^T = (y_j)$, $X_i^T = (x_{ji})$, j =1, 2,…, m; (y_j, x_{1j}, x_{2j}, …, x_{nj}) được gọi là quan sát thứ j (hay trường hợp thứ j) của các biến Y, X_i. Một số biến gốc X_i có thể không có hoặc có tác động rất ít đến sự thay đổi của Y; một số biến gốc khác có thể có tương quan với nhau. Giả sử số biến gốc n là rất lớn.

Vấn đề đặt ra: xây dựng mô hình dự báo giá cổ phiếu (biến đích Y) theo tập các biến gốc X_i (i =1, 2, …, n).

2.2. Khung lý thuyết dự báo

Hình 1 ở dưới trình bày một cách tóm tắt khung lý thuyết của quá trình dự báo biến đích trong ngữ cảnh dữ liệu có số chiều biến cao. Theo đó quá trình này gồm 2 giai đoạn cơ bản: giảm chiều dữ liệu của tập dữ liệu đầu vào và xây dựng mô hình dự báo trên tập dữ liệu mới và thực hiện dự báo.



Hình 1. Khung lý thuyết dự báo trong ngữ cảnh dữ liệu chiều cao

Dưới đây trình bày chi tiết hơn Khung lý thuyết này.

2.3. Pha 1: Giảm chiều dữ liệu

2.3.1. Giảm chiều lần 1: Sử dụng kỹ thuật lọc (Filter)

Mục đích của pha này là giảm được số biến trong khi vẫn giữ được các quan hệ giữa biến đích và các biến gốc nhiều như có thể. Trong tập các biến gốc có thể có những biến không hoặc tác động rất không đáng kể đến sự thay đổi của biến đích cũng như có thể hiện tượng dư thừa biến. Quan trọng nhất của giảm chiều lần 1 nhằm giảm các biến như vậy. Khi đó kỹ thuật để giảm chiều Lần 1 cần thuộc phương pháp tiếp cận lọc (filter) không thể là cách tiếp cận nhúng (embedded) hoặc bọc (wapper). Thuật toán **ChonTapcon** ở dưới sẽ thực hiện giảm chiều lần 1.

Theo cách tiếp cận lọc, người ta thường sử dụng một trong loại 2 độ đo:

- Độ đo sự tương quan giữa 2 biến X, Y là |R(X,Y)| ở đây R(X,Y) được xác định bởi công thức (1):

$$R(X,Y) = \frac{\sum_{i=1}^{m}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{m}(y_i-\bar{y})^2}}, \text{ ở đây } \bar{x} = \frac{\sum_{i=1}^{m} x_i}{m}, \quad (1)$$

$$\bar{y} = \frac{\sum_{i=1}^{m} y_i}{m}, \text{ X=}(x_i), \text{ Y= } (y_i), \text{ i=1, m.}$$

$|R(X,Y)| \leq 1$ và càng gần 1 thì khả năng một biến được xác định thông qua biến còn lại càng cao. Độ đo này được gọi là độ đo tương quan Pearson [3, 8].

- Độ đo thông tin tương hỗ giữa 2 biến. Thuộc loại này có một số độ đo, trong đó độ đo thông tin tương hỗ dựa vào Entropy thường được sử dụng [13]. Với hai biến X, Y nêu trên, độ đo thông tin tương hỗ giữa 2 biến này được xác định bởi:

$$I(X,Y) = \sum_{x_i}\sum_{y_j} P(X=x_i, Y=y_j) log \frac{P(X=x_i,Y=y_j)}{P(X=x_i).P(Y=y_j)} \quad (2)$$

Các bài báo [6, 20] đã đề xuất một cách đo khác để đo mức độ quan hệ nguyên nhân - kết quả giữa 2 biến. Đó chính là giá trị của xác suất thống kê T trong kiểm định quan hệ nhân quả được đề xuất bởi nhà toán học - giải Nobenl kinh tế Granger C. W. J. [11]. Quan hệ nhân quả được xác định dựa trên việc xây dựng một mô hình toán học như sau:

Giả sử X và Y đều là những chuỗi dừng [10, 12], xét 2 phương trình [11]:

$$Y = \sum_{i=1}^{n} a_i X(-i) + \sum_{j=1}^{m} b_j Y(-j) + u_{1t} \quad (3)$$

$$X_t = \sum_{i=1}^{p} c_i X(-i) + \sum_{j=1}^{q} d_j Y(-j) \ + u_{2t} \quad (4)$$

ở đây các a_i, c_i, b_j, d_j, là các tham số; X(-i), Y(-j) tương ứng là các X trễ i bước và Y trễ j bước; vì trễ của các biến cũng là biến nên trong các phương trình (3) và (4): n, m, p, q chính là số các biến giải thích; nó chính là độ dài trễ lớn nhất của các biến X, Y trong mỗi phương trình; u_{it} (i =1, 2) là sai số được giả định là nhiễu trắng. Các tham số trên được xác định bằng sử dụng phương pháp hồi quy nhiều biến trên tập dữ liệu đầu vào.

Nếu $\sum_{i=1}^{n} a_i^2 \neq 0$ và $\sum_{j=1}^{m} d_j^2 = 0$ thì ta nói tồn tại mối quan hệ nhân quả duy nhất chiều từ X đến Y và tương tự, nếu $\sum_{i=1}^{n} a_i^2 = 0$ và $\sum_{j=1}^{m} d_j^2 \neq 0$ thì tồn tại mối quan hệ nhân quả duy nhất chiều từ Y đến X. Nếu $\sum_{i=1}^{n} a_i^2 \neq 0$ và $\sum_{j=1}^{m} d_j^2 \neq 0$ giữa hai biến X và Y còn được gọi là có quan hệ nhân quả hai chiều (hay quan hệ phản hồi). Các biến X và Y là độc lập nhau nếu $\sum_{i=1}^{n} a_i^2 = 0$ và $\sum_{j=1}^{m} d_j^2 = 0$. Khi có quan hệ nhân quả chiều từ X đến Y thì X chính là nguyên nhân gây ra Y và nó cũng chính là chỉ số báo trước của Y. Kiểm định nhân quả Granger có phân phối Student T. Dựa vào xác suất của thống kê T ta biết được mức độ của quan hệ nhân quả giữa X và Y.

Ký hiệu d(X,Y) là chung cho các độ đo tương quan Pearson, độ đo thông tin tương hỗ và xác suất thống kê T về quan hệ nhân quả. Khi đó một biến gốc X_i được coi là không hoặc ít liên quan đến biến đích Y nếu $d(X_i,Y) \leq \alpha$; Biến gốc $\mathbf{X}_j$ được coi là dư thừa nếu tồn tại biến gốc khác X_i sao cho $d(\mathbf{X}_j, X_i) > \beta$ và $d(X_i,Y) > d(X_j,Y)$, ở đây α, β là những số dương nhỏ hơn 1 do người sử dụng xác định.

Ký hiệu G ={X_i, i=1,2, …,n}là tập tất cả các biến gốc đầu vào. Khi đó thuật toán giảm chiều dữ liệu bằng sử dụng kỹ thuật Lọc thuộc tính với sử dụng độ đo tương quan d(X,Y) được viết dưới dạng giả code như sau:

Thuật toán ChonTapcon

Đầu vào: Y biến đích, tập biến gốc G = {X_i}**,** 2 ngưỡng do người sử dụng xác định: ngưỡng α cho độ đo $d(X_i, Y)$ để xác định X_i không hoặc có tác động rất ít đến Y? ngưỡng β cho độ đo $d(X_j, X_i)$ để xác nhận X_j có thể xác định được từ X_i?

Đầu ra: Tập dữ liệu con của G, ở đó không còn biến không hoặc có tác động rất ít đến biến đích Y và không còn biến dư thừa.

1. **for** i ←1 to So_bien_trong_G **đo** // *loại bỏ các biến không hoặc ít liên quan với biến đích Y*
2. **Tinhdodo** $d(X_i, Y)$;
3. **If** $d(X_i,Y) \leq \alpha$ **then** G ← G\{X_i}
4. **end for**
5. **Order(G)** // *sắp xếp các biến trong G theo thứ tự giảm dần của* $|d(X_i,Y)|$
6. **for** i ← 2 to So_bien_trong_G **do**
7. **for** j ← 1 to i-1 **do**
8. **Tinhdodo** $d(X_j, X_i)$

9. **If** $|d(X_j, X_i)| > \beta$ **then** $G \leftarrow G\backslash\{X_i\}$ // *loại bỏ các biến dư thừa*

10. **end for**

11. **end for**

12. **Return** G

2.3.2. Giảm chiều lần 2: sử dụng kỹ thuật PCA

Việc giảm chiều dữ liệu bằng sử dụng kỹ thuật PCA có thể được thực hiện trong môi trường của các ngôn ngữ, công cụ thống kê như MATLAB, R, SAS, EVIEW, SPSS, STATA. Bài báo này sử dụng công cụ EVIEW [25].

Việc sinh ra các biến mới thay thế cho tập các biến gốc G bằng sử dụng kỹ thuật PCA gồm những nội dung chinh sau [1, 7]:

1) Tính ma trận tương quan **R** của tập gồm n biến gốc X_i;

2) Tìm các giá trị riêng và véctơ riêng của ma trận **R**. Giả sử có h giá trị riêng ($h \le n$).

3) Sắp xếp các giá trị riêng theo thứ tự giảm dần;

4) Phân tích tỷ lệ tích lũy của các giá trị riêng, chọn số thành phần chính có ứng với các giá trị riêng cao nhất và có tổng tích lũy giá trị riêng từ 70% đến 90%, hàm ý các thành phần chính được chọn khi đó giải thích được tương ứng từ 70% đến 90% của tập dữ liệu gốc;

5) Giả sử có k thành phần chính được giữ lại, ký hiệu là $PC_1, PC_2, \ldots, PC_k$. khi đó ($k \le h$). Sử dụng các véctơ riêng làm trọng số để tạo ra các thành phần chính theo số lượng được chọn. Ký hiệu $V_1, V_2, \ldots, V_k$ là các véctơ riêng ứng với các thành phần chính $PC_1, PC_2, \ldots, PC_k$. mỗi V_i là một véctơ n chiều cụ thể $V_i^T = (v_{i1}, v_{i2}, \ldots, v_{in})$, khi đó thành phần chính PC_i là véctơ m chiều ứng với véctơ riêng V_i được xác định như sau [7]:

$$PC_i = v_{i1} * \hat{X}_1 + v_{i2} * \hat{X}_2 + \ldots + v_{in} * \hat{X}_n, \quad (5)$$

ở đây $\hat{X}_i = \frac{X_i - \bar{\bar{X}}_i}{S_i}$, (6)

trong đó $\bar{\bar{X}}_i$, S_i tương ứng là giá trị trung bình và độ lệch chuẩn của véctơ X_i. Các véctơ $\hat{X}_i$ được gọi là véctơ chuẩn hóa của véctơ X_i.

Các thành phần chính nhận được khi đó sẽ tự nhiên được sắp theo thứ tự theo độ lớn của các giá trị riêng tương ứng của nó.

2.4. Xây dựng mô hình và thực hiện dự báo

2.4.1. Xây dựng mô hình dự báo

Sau giảm chiều lần thứ nhất, giả sử tập thu gọn của tập biến gốc ban đầu là $G = \{X_1, X_2, \ldots, X_g\}$. Dựa vào lý thuyết tài chính - kinh tế, tập G được chia thành 2 tập, $G_1 = \{ X_1, X_2, \ldots, X_k \}$ với $k \le g$ và $G_2 = G\backslash G_1$ trong đó G_2 gồm các biến thường khó dự báo và sự thay đổi của nó thường gây ra hiện tượng "sốc" cho các hoạt động tài chính - kinh tế hoặc có thể là tập rỗng. Số biến trong G_2 thường nhỏ, không gây lo ngại về thách thức số chiều cao của dữ liệu. Khi đó việc giảm chiều lần 2 chủ yếu được thực hiện đối với tập G_1.

Không giảm tổng quát ta có thể coi rằng các tập $G_1 \cup G_2 = \{X_1, X_2, \ldots, X_k\} \cup \{ X_{k+1}, X_{k+2}, \ldots, X_g\}$ là tập biến mới sau 2 lần được giảm chiều và được sử dụng để thay thế cho tập biến gốc ban đầu.

Mô hình dự báo giá cổ phiếu gồm 2 mô hình:

a. Mô hình dự báo trung bình

Là mô hình ADL có dạng:

$$Y = c + \sum_{i=0}^{r1} a_{1i} X_1(-i) + \sum_{i=0}^{r2} a_{2i} X_2(-i) + \ldots + \sum_{i=0}^{rk} a_{ki} X_k(-i) + \sum_{q=1}^{r} b_q Y(-q) + u(t) \quad (7)$$

ở đây $X(-i) \equiv X(t-i)$ là ký hiệu biến trễ i bước của X. Trong công thức (7) trễ của các biến khác nhau nói chung là khác nhau, nhưng có thể dễ dàng xác định độ dài trễ này nếu kiểu tần suất thu thập dữ liệu của chúng như tuần, tháng, quý hay năm,… Về bản chất, độ dài trễ thường là độ dài mùa vụ của chuỗi dữ liệu. Do các biến X_i, i =1,2, …, k là các thành phần chính nên chúng trực giao với nhau do đó mô hình được xác định bởi phương trình (7) không có hiện tượng đa cộng tuyến [1].

Thực hiện kiểm định hiện tượng ARCH của **u(t)**, nếu không có hiện tượng này thì cần kiểm định và xử lý để **u(t)** không là nội sinh, **u(t)** có phân phối chuẩn, có kỳ vọng bằng 0, phương sai thay đổi và không tự tương quan theo

những cách đã được nêu trong [8, 12], trường hợp trái lại thì cần phải sử dụng mô hình GARCH(p,q) [8] để dự báo phương sai phần dư của mô hình như được xác định theo phương trình (8).

b. Mô hình dự báo phương sai phần dư

Ký hiệu h(t) là phương sai của **u(t)** trong phương trình (7), khi đó phương trình dự báo phương sai phần dư GARCH(p,q) có dạng:

$$H = \propto + \sum_{i=1}^{p} a_i H(-i) + \sum_{i=1}^{q} b_i . u(-i)^2 + \sum_{i=0}^{r1} c_{1i.} X_{k+1}(-i) + \sum_{i=0}^{r2} c_{2i.} X_{k+2}(-i) + \cdots + \sum_{i=0}^{r(g-k)} c_{(g-k)i.} X_g(-i) + \epsilon(t) \quad (8)$$

Trong phương trình (8) H là phương sai của phần dư u(t); $\epsilon(t)$ là phần dư của mô hình dự báo phương sai được giả định là biến ngẫu nhiên có kỳ vọng bằng 0, có phân phối chuẩn, không có hiện tượng ARCH và không tự tương quan chuỗi; các biến ngoại sinh X_{k+1},…, X_g là thuộc tập G_2 như đã nêu, phần: $\propto + \sum_{i=1}^{p} a_i h(-i) + \sum_{i=1}^{q} b_i . u(-i)^2$ là thuộc về GARCH(p, q) và:

$$\sum_{i=0}^{r1} c_{1i.} X_{k+1}(-i) + \sum_{i=0}^{r2} c_{2i.} X_{k+2}(-i) + \cdots + \sum_{i=0}^{r(g-k)} c_{(g-k)i.} X_g(-i)$$

được gọi là phần mở rộng của mô hình GARCH, nó được sử dụng để nghiên cứu, đánh giá tác động của các "sốc" đến tính không chắc chắn của lợi nhuận [2] cũng như đánh giá hiệu quả hoạt động của thị trường [4].

c. *Ước lượng và chẩn đoán mô hình dự báo*

Chia tập dữ liệu thành 2 tập: tập thứ nhất dùng để huấn luyện mô hình, tập thứ 2 để dự báo kiểm định, đánh giá chất lượng mô hình. Do các dữ liệu kinh tế, tài chính thường biến đổi rất nhanh, nên thời gian dự báo xa nhất của dự báo nên là trung hạn. Khái niệm dự báo ngắn hạn, trung hạn hay dài hạn trong lĩnh vực kinh tế - tài chính được xác định cụ thể như sau: dự báo cho 1-2 kỳ dữ liệu tiếp theo là dự báo ngắn hạn, 3-5 kỳ dữ liệu tiếp theo là trung hạn, từ 6 kỳ dữ liệu tiếp theo trở lên là dài hạn [10].

Thực hiện ước lượng mô hình dự báo trung bình theo phương trình (7) bằng sử dụng phương pháp hồi quy nhiều biến trên tập dữ liệu thứ nhất. Thực hiện xử lý phần dư như đã nêu ở trên ứng với 2 trường hợp phần dư có hoặc không hiện tượng ARCH.

2.4.2. Đánh giá chất lượng mô hình

Thực chất của nội dung này là kiểm thử khả năng dự báo ngoài mẫu của mô hình, bằng cách sử dụng mô hình được xây dựng trên tập dữ liệu thứ nhất để dự báo cho tập dữ liệu thứ 2, sau đó so sánh tập thứ 2 thực tế và tập thứ 2 dự báo. Có 2 độ đo được sử dụng nhiều nhất để đo độ chính xác của dự báo là: phần trăm sai số giữa giá trị dự báo và giá trị thực và trung bình hoặc căn bậc 2 của trung bình bình phương sai số của dự báo. Nếu độ đo sai số dự báo nhỏ như mong muốn của người sử dụng thì có thể sử dụng mô hình để dự báo tương lai của biến đích.

2.4.3. Thực hiện dự báo tương lai

Để dự báo tương lai của biến đích, cần:

- Thực hiện dự báo các biến ngoại sinh X_1, X_2, …, X_k cho mô hình dự báo trung bình theo phương trình (7) bằng việc sử dụng mô hình tự hồi quy AR(p) có xu thế như được xác định bởi phương trình (9).

$$\Delta Y(t) = \alpha + \rho Y(-1) + \gamma_1 \Delta Y(-1) + \ldots + \gamma_p \Delta Y(-p) + \delta t + e_t, \quad (9)$$

ở đây: ΔY là ký hiệu sai phân bậc 1 của Y, t là biến để đo số lượng các quan sát.

- Thực hiện dự báo các biến X_{k+1}, X_{k+2}, …, X_g theo mô hình AR(p) có xu thế nếu người dự báo cảm nhận rằng những yếu tố tác động đến các biến này trong tương lai là tương tự như hiện tại và quá khứ, nhưng nói chung không phải như vậy, nên các biến này thường được dự báo giả định.

- Thực hiện ước lượng lại mô hình dự báo trung bình và dự báo phương sai trên toàn bộ tập dữ liêu, sau đó sử dụng mô hình này cùng các biến ngoại sinh đã được dự báo trước đó để dự báo biến đích Y.

III. ỨNG DỤNG KHUNG LÝ THUYẾT DỰ BÁO

3.1. Bài toán cụ thể và tập dữ liệu được sử dụng để dự báo

Giả sử biến đích Y là biến giá cổ phiếu của Công ty FPT và tập các biến gốc có tác động đến sự biến động của giá cổ phiếu FPT được xác định theo cách tiếp cận như trong [22]. Giả sử các biến gốc và nguồn thu thập dữ liệu cho các biến này được mô tả trong Bảng 1 ở dưới.

Bảng 1. Các biến gốc đầu vào cho việc xây dựng mô hình dự báo giá cổ phiếu FPT

Data	Variables properties	Source
06 biến gốc: Chỉ số phát triển Công nghiệp: IIP, dư nợ tín dụng: DUNO, lãi suất tiền gửi ngắn hạn: INT, tỷ giá hối đoái VNĐ và USD: ER, kim ngạch xuất khẩu và nhập khẩu của Việt Nam theo tháng: EX và IMP,	Phản ánh các điều kiện phát triển chung của nền kinh tế.	www.gso.gov.vn; www.vietcombank.com.vn Tần suất dữ liệu theo tháng
41 biến gốc: Giá của 29 mã cổ phiếu cổ phiếu BULUECHIP (mã cổ phiếu được xem là tên biến); Các chỉ số chứng khoán: VNINDEX, HNX, chỉ số chứng khoán của 30 cổ phiếu BLUECHIP: VN30; chỉ số: UPCOM; Các chỉ số chứng khoán theo các ngành, Công nghiệp: CNINDEX, Khoáng sản: KSINDEX, Ngân hàng: NHINDEX, Năng lượng: NLINDEX; Chỉ số giá tiêu dùng: CPI, chỉ số giá vàng: GOLDINDEX, chỉ số giá đô la ở Việt Nam: USINDEX	Các biến số cụ thể liên quan đến biến động giá cổ phiếu FBT	www.cophieu68.org.vn www.gso.gov.vn; Dữ liệu theo tháng là trung bình của dữ liệu theo các ngày sàn giao dịch chứng khoán hoạt động trong tháng
01 biến: Giá cổ phiếu của Công ty FPT	Biến cụ thể của công ty	www.cophieu68.org.vn Dữ liệu theo tháng là trung bình của dữ liệu theo các ngày sàn giao dịch chứng khoán hoạt động trong tháng
0 biến: chưa thu thập được những thông tin như vậy.	Các biến số tâm lý của các nhà đầu tư vào Công ty FPT, như kỳ vọng và lựa chọn mức giá để mua cổ phiếu của các nhà đầu tư trong đó nhất là của những nhà đầu tư có tổ chức	
04 biến: Các chỉ số chứng khoán NASDAD 100 và tổng hợp: NDX100 và NDX_COM, chỉ số chứng khoán S&P500: SP500, giá thế giới về dầu thô ở thị trường TEXAS- Mỹ: OIL;	Các biến số phản ánh kinh tế- chính trị thế giới và của những nước lớn. Các biến cố chính trị: sự xuất hiện và sự ra đời của các sự kiện chính trị quan trọng.	Federal Reserve Bank of ST. Louis, https://fred.stlouisfed.org/ Dữ liệu theo tháng là trung bình của dữ liệu theo các ngày trong tháng

Như vậy tập dữ liệu bao gồm 65 quan sát từ tháng 1 năm 2012 đến tháng 5 năm 2017 cho 01 biến đích và 51 biến gốc đầu vào.

3.2. Giảm chiều dữ liệu

3.2.1. Giảm chiều lần 1

Trong thuật toán **ChonTapcon** ở trên nếu chọn d(X,Y) là độ đo tương quan Pearson, tức d9X,Y) = |R(X,Y| ở đây R(X,Y) là hệ số tương quan Pearson giữa 2 biến này. Nếu chọn $\alpha = 0.26$ hàm ý rằng những biến gốc có hệ số tương quan với biến FPT nhỏ hơn 0.26 thì ta coi những biến này là không hoặc ít có tác động đến sự thay đổi của FPT và nếu chọn $\beta = 0.91$ hàm ý rằng 2 biến có hệ số tương quan lớn hơn 0.91, thì biến này có thể được xem là được xác định thông qua biến kia (một cách chính xác biến này có thể giải thích được trên 91% sự thay đổi của biến kia).

Thực hiện thuật toán **ChonTapcon** trên tập dữ liệu đầu vào của biến đích và các biến gốc cũng 2 tham số α, β nêu trên ta sẽ nhận được tập con thuộc tính sau Lần giảm chiều lần thứ nhất như được chỉ ra trong Bảng 2 ở dưới:

Bảng 2. Tập con thuộc tính (biến) có ý nghĩa với biến FPT và không dư thừa từ 51 biến gốc

Số TT	Biến gốc	Hệ số tương quan	Số TT	Biến gốc	Hệ số tương quan	Số TT	Biến gốc	Hệ số tương quan
1	CNINDEX	0.996	11	EX	0.778	21	NHINDEX	0.584
2	SP500	0.922	12	VNM	0.763	22	HAG	-0.575
3	VNINDEX	0.916	13	CII	0.752	23	UPCOM	0.498
4	VSH	0.907	14	CPI	-0.742	24	FLC	0.494
5	DUNO	0.890	15	OIL	-0.738	25	GOLDINDEX	0.490
6	HCM	0.869	16	VN30	0.737	26	CTG	0.441
7	GMD	0.825	17	OGC	-0.735	27	USDINDEX	0.414
8	MSN	-0.807	18	KDC	0.699	28	IJC	0.319
9	EIB	-0.784	19	CSM	0.675			
10	PVT	0.781	20	BVH	0.598			

Thực ra 4 biến là: PVD, ITA, DPM và PPC là các biến ít liên quan với FPT do có hệ số tương quan với FPT tương ứng là: -0.0007, -0.175, -0.211 và 0.254 nhỏ hơn $\alpha = 0.26$, còn 19 biến khác không được chọn bởi chúng là các biến dư thừa.

3.2.2. Giảm chiều lần 2

Phân tích 28 biến trong tập con thuộc tính được lựa chọn để xây dựng mô hình dự báo giá cổ phiếu FPT, ta thấy có 2 biến là chỉ số chứng khoán S&P500 (SP500) và giá thế giới về dầu thô (OIL) là những biến phản ánh tình hình kinh tế - chính trị thế giới. Những biến này là khó lường trong bối cảnh ở Việt Nam nên có thể xác định nó là những biến thường gây nên tính chắc chắn của kết quả dự báo. Bài báo này đề xuất đưa 2 biến này làm biến ngoại sinh trong mô hình dự báo phương sai được xây dựng theo phương trình (8), 26 biến còn lại được sử dụng để xây dựng mô hình dự báo trung bình theo phương trình (7).

Thực hiện phân tích PCA của 26 biến gốc phản ánh kinh tế - tài chính trong nước với dữ liệu của 60 quan sát từ tháng 1/2012 đến tháng 12/2016 theo các bước được nêu trong mục 3.3.2, ta thấy ma trận các hệ số tương quan **R** giữa 26 biến này có 26 giá trị riêng, ở đó có 5 giá trị riêng lớn hơn 1 (**Hình 2**) và tổng tích lũy của 5 giá trị riêng đó là 0.909 (**Bảng 3**). Điều đó ngầm ý rằng chỉ cần chọn 5 thành phần chính ứng với 5 giá trị riêng lớn nhất để làm đại diện cho tập gồm 26 biến gốc (cũng có nghĩa là cho 49 biến gốc ban đầu) và 5 thành phần chính đó giải thích được đến 90.9% sự thay đổi trong tập dữ liệu của 49 biến gốc.

Để xác định được 5 thành phần chính, trước hết cần tìm 5 véctơ riêng ứng với 5 giá trị riêng lớn nhất và nó sẽ véctơ trọng số tương ứng của 5 thành phần chính, đồng thời thực hiện chuẩn hóa tập dữ liệu đầu vào của 26 biến gốc theo công thức (6). Các thành phần chính sẽ được tính theo công thức (5). Các thành phần chính là không tương quan với nhau, nói cách khác chúng luôn là các biến độc lập. Ký hiệu 5 thành phần chính tương ứng với 5 giá riêng lớn nhất và giảm dần là PC_1, PC_2, PC_3, PC_4, PC_5.

3.3. Xây dựng mô hình dự báo và thực hiện dự báo

3.3.1. Ước lượng mô hình

Các tham số của các mô hình dự báo được xây dựng theo phương trình (7) và (8) đều được ước lượng bằng sử dụng phương pháp hồi quy OLS. Để tránh hiện tượng hồi quy sai thì các biến trong mô hình phải là chuỗi thời gian dừng.

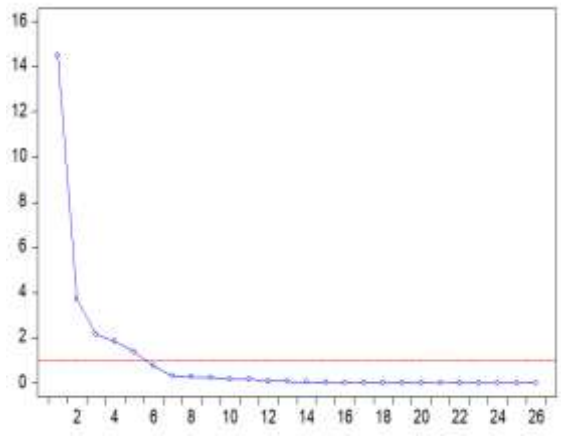
Thực hiện các kiểm định Dickey-Fuller tăng cường và kiểm định Phillips - Perron [8, 11] trên các biến, có thể thấy các biến chuỗi thời gian FPT, OIL, SP500 không dừng nhưng lôga tự nhiên của các biến này đều dừng sai phân bậc 1, trong khi đó 5 thành phần chính PC_1, PC_2, PC_3, PC_4, PC_5 đều không dừng nhưng sai phân bậc 1 của chúng đều dừng.

Vì vậy các phương trình (7), (8) trong trường hợp này tương ứng trở thành:

$$d\log(FPT) = c + \sum_{i=0}^{r1} a_{1i} d(PC_1(-i)) + \sum_{i=0}^{r2} a_{2i} d(PC_2(-i)) + \ldots + \sum_{i=0}^{rk} a_{5i} d(PC_5(-i)) + \sum_{q=1}^{r} b_q d\log(FPT(-q)) + u(t) \quad (10)$$

$$H = \propto + \sum_{i=1}^{p} a_i H(-i) + \sum_{i=1}^{q} b_i . u(-i)^2 + \sum_{i=0}^{r1} c_{1i.} dlog(SP500(-i)) + \sum_{i=0}^{r2} c_{2i.} dlog\big(OIL(-i)\big) + \epsilon(t) \quad (11)$$

ở đây d(X) là ký hiệu sai phân bậc 1 của X.



Hình 2. Các giá trị riêng được sắp thứ tự giảm dần

Số thứ tự	Giá trị riêng	Chênh lệch	Tỷ lệ	Giá trị tích lũy	Tỷ lệ tích lũy
1	14.511	10.789	0.558	14.511	0.558
2	3.722	1.552	0.143	18.233	0.701
3	2.170	0.327	0.083	20.403	0.785
4	1.843	0.461	0.071	22.246	0.856
5	**1.382**	**0.622**	**0.053**	**23.628**	**0.909**
6	0.760	0.439	0.029	24.388	0.938
7	0.321	0.042	0.012	24.709	0.950
8	0.279	0.027	0.011	24.989	0.961
9	0.252	0.065	0.010	25.241	0.971
10	0.187	0.006	0.007	25.428	0.978
...	...	.....	.....	....	....

Bảng 3. Các giá trị riêng và tỷ lệ tích lũy của chúng

Sử dụng tiêu chuẩn AIC [12] để xác định các độ dài trễ tối ưu trong phương trình (10) ta sẽ nhận được độ trễ tối ưu cho tất cả các biến trong mô hình này đều là 1. Sử dụng kiểm định WALD [11] để kiểm tra các biến trong mô hình có thực sự cần thiết nằm trong mô hình này hay không và sử dụng phương pháp hồi quy nhiều biến để ước lượng các tham số trong mô hình được xây dựng theo phương trình (10). Xem xét đồ thị phần dư u(t) sẽ nhận thấy rằng không có hiện tượng giá trị phần dư vượt quá 1,5 lần độ lệch chuẩn của nó trong những tháng gần đây. Nói cách khác không có hiện tượng dữ liệu bất thường/dữ liệu ngoại lai trong tập dữ liệu đầu vào được sử dụng. Kiểm định hiện tượng ARCH [8,11] của phần dư **u(t)** trong mô hình dự báo trung bình vừa được ước lượng ta nhận được phân dư có hiện tượng ARCH. Khi đó cần phải ước lượng mô hình dự báo phương sai theo phương trình (11) và được tiến hành đồng thời với quá trình ước lượng mô hình dự báo trung bình theo phương trình (10). Kết quả ước lượng là các mô hình dự báo trung bình và phương sai như sau:

$$\textbf{dlog(FPT)} = 0.233^{*}\,\mathrm{dlog(FPT(-1))} + 0.065^{*}\,\mathrm{d(PC_1)} + 0.026^{*}\mathrm{d(PC_5)} + \mathrm{u(t)}; \quad (12)$$

Std: (0.052) (0.009) (0.005)

$$\mathbf{H} = 0.001 - 0.246^{*}u^{2}(-1) + 0.831^{*}H(-1) + 0.002^{*}\mathrm{dlog(OIL)} + 0.006\,\mathrm{dlog(SP500)} + \epsilon(t) \quad (13)$$

Std: (0.0003) (0.067) (0.129) (0.0008) (0.004)

R^2 = 0.59; DW: 1.93; SMPL: 58 quan sát sau khi được điều chỉnh.

Kiểm định phần dư $\epsilon(t)$ trong mô hình (13) ta thấy $\epsilon(t)$ có kỳ vọng bằng 0, có phân phối chuẩn, không còn hiện tượng ARCH và không tự tương quan chuỗi. Phương trình (13) cho thấy các biến OIL và SP500 đều có ảnh hưởng đến sự thay đổi của phương sai phần dư mô hình dự báo trung bình; tác động của biến OIL đến sự thay đổi của phương sai H có ý nghĩa thống kê cao, song với biến SP500 thì tác động này có ý nghĩa thống kê không cao. Phương trình (13) cho thấy tốc độ thay đổi của OIL và SP500 có ảnh hưởng thuận chiều đến sự thay đổi của phương sai H và mặt khác quan trọng hơn đó là nó thể hiện thị trường chứng khoán Việt Nam hoạt động chưa hiệu quả đối với sự thay đổi của các biến giá cổ phiếu công ty FPT, trong đó nhất là đối với biến OIL. Nói cách khác các thông tin liên quan đến sự thay đổi của OIL chưa được các nhà đầu tư vào cổ phiếu FPT phản ứng kịp thời, trong khi phản ứng như vậy có khá hơn đối với sự thay đổi của chỉ số SP500. Khi phản ứng của thị trường liên quan đến cổ phiếu FPT chưa kịp thời với sự thay đổi của OIL và SP500 thì cũng có nghĩa là sự thay đổi của các biến này không gây ra những biến động của lợi nhuận đầu tư vào cổ phiếu FPT.

3.3.2. Dự báo kiểm định

Sử dụng mô hình dự báo giá cổ phiếu FPT được xây dựng trên tập dữ liệu đầu vào của các quan sát từ tháng 1/2012 đến tháng 12/2016 để dự báo giá cổ phiếu FPT trên tập dữ liệu đầu vào của 5 quan sát từ tháng 1/2017 đến tháng 5/2017 sẽ nhận được kết quả trong Bảng 4.

Bảng 4. Kết quả dự báo kiểm thử chấp nhận mô hình

Quan sát	Giá trị thực tế	Giá trị dự báo	% sai số dự báo
Tháng 1/2017	38.280	38.124	-0.407
Tháng 2/2017	38.590	40.538	5.047
Tháng 3/2017	39.550	41.004	3.117
Tháng 4/2017	39.580	39.399	-0.357
Tháng 5/2017	41.290	42.440	2.784

Bảng này cho thấy, xu hướng tăng giá của mã cổ phiếu trong các tháng 3 và 4/2017 được dự báo trái với xu hướng thực tế mặc dù % sai số dự báo so với thực tế là không cao, nói chung không vượt quá 5%.

3.3.3. Thực hiện dự báo

Để thực hiện dự báo chẳng hạn cho 3 tháng tiếp theo: từ tháng 6 đến tháng 8/2017 về giá cổ phiếu FPT, ta cần dự báo các biến ngoại sinh trong các phương trình (12) và (13) là PC_1, PC_5, SP500 và OIL ở 3 tháng tiếp theo. Bài báo này sử dụng kết quả dự báo chỉ số chứng khoán SP500, giá dầu thế giới của Trung tâm dự báo tài chính Hoa Kỳ [24], còn mô hình dự báo các thành phần chính PC_1, PC_5 được cho tương ứng bởi các phương trình (14) và (15) ở dưới.

$$\mathbf{d(PC_1,2)} = -2.06^{*}\mathrm{d(PC_1(-1))} + 1.05^{*}\mathrm{d(PC_1(-1),2)} + 0.84^{*}\mathrm{d(PC_1(-2),2)} + 0.59^{*}\mathrm{D(PC_1(-3),2)} + 0.52^{*}\mathrm{d(PC_1(-4),2)}$$

Std: (0.40) (0.34) (0.29) (0.24) (0.18)

$$+ 0.31^{*}\mathrm{d(PC_1(-5),2)} + 0.43 \quad (14)$$

(0.14) (0.19)

R^2: 0.59; DW: 2.09; SMPL: 58, sau khi đã được điều chỉnh bởi trễ.

$$\mathbf{d(PC_5,2)} = -0.698*d(PC_5(-1)) - 0.227*d(PC5(-2),2) - 0.192*d(PC_5(-3),2) - 0.207*d(PC_5(-9),2) \quad (15)$$

Std: (0.144) (0.118) (0.100) (0.099)

R^2: 0.49; DW: 1.88; SMPL: 54, sau khi đã được điều chỉnh bởi trễ.

ở đây d(X,2) là ký hiệu sai phân bậc 2 của biến X [8].

Kết quả dự báo được cho trong Bảng 5.

Bảng 5. Kết quả dự báo giá của cổ phiếu FPT 3 tháng tiếp theo

Biến	Tháng 6/2017	Tháng 7/2017	Tháng 8/2017	Nguồn
OIL	48.2	51.8	53.2	www.forecast.org
SP500	2426	2478	2512	www.forecast.org
PC1	7.559	7.420	7.406	Tác giả
PC5	0.567	0.567	0.569	Tác giả
FPT	41.451	41.147	41.068	Tác giả
Sai số trung bình	+/- 1.582	+/- 1.565	+/- 1.561	Tác giả
Sai số trung bình (%)	+/- 3.82	+/- 3.80	+/- 3.80	

IV. KẾT LUẬN

Bài báo đã đề xuất khung lý thuyết để dự báo giá hàng hóa và dịch vụ nói chung, giá cổ phiếu nói riêng trong ngữ cảnh tập dữ liệu đầu vào cho dự báo chúng là rất lớn. Kỹ thuật giảm chiều được đề xuất trong bài báo là sự kết hợp của 2 kỹ thuật lựa chọn thuộc tính và học thuộc tính theo cách sao cho loại bỏ được những thuộc tính không có ích, những thuộc tính dư thừa trong khi vẫn đảm bảo được tối đa nhất có thể quan hệ giữa các biến gốc và biến đích.

Bài báo đã chỉ ra rằng trong lĩnh vực kinh tế - xã hội cho đến thời điểm này việc sử dụng các kỹ thuật dự báo thống kê vẫn là lựa chọn được ưu tiên, đồng thời với bài toán dự báo giá, trong rất nhiều trường hợp phải sử dụng một trong những mô hình thuộc họ ARCH để dự báo sự thay đổi của phương sai phần dư trong mô hình dự báo chính.

Việc ứng dụng khung lý thuyết để dự báo giá cổ phiếu FPT trên tập số liệu thống kê thực của nền kinh tế cho thấy độ chính xác dự báo là rất khả quan trong khi mới chỉ có 51 biến kinh tế - tài chính trong và ngoài nước được xem xét đưa vào mô hình dự báo và chắc chắn còn thiếu rất nhiều biến kinh tế - tài chính khác, trong đó nhất là thiếu những biến đo lường tâm lý, kỳ vọng của các nhà đầu tư đang đầu tư vào cổ phiếu này cũng như những biến đo lường các cú sốc chính trị, kinh tế thế giới và trong nước.

Mô hình dự báo phương sai phần dư trong bài báo này được phát triển dựa trên mô hình GARCH(p, q). Trong trường hợp bài toán này, việc lựa chọn mô hình GARCH đã thực sự phù hợp chưa so với các mô hình thuộc họ ARCH khác như ARCH(p), EARCH, PARCH, ARCH-M cần phải được xác định thông qua thực hiện một số kiểm định thống kê. Bài báo này đã bỏ qua không thực hiện những nội dung đó.

Phân tích quan hệ giữa hệ số của các biến và sai số chuẩn tương ứng trong phương trình (13) sẽ nhận thấy rằng các biến dlog(SP500) tham gia vào mô hình dự báo phương sai với ý nghĩa thống kê không cao? Phần trăm sai số của dự báo kiểm thử chấp nhận mô hình nói chung không quá 5% trong khi trong mô hình dự báo được xây dựng vẫn chưa tính đến các biến số đo lường tâm lý của các nhà đầu tư vào công ty FPT như kỳ vọng lợi tức và lựa chọn mức giá để mua cổ phiếu FPT của các nhà đầu tư trong đó nhất là của những nhà đầu tư có tổ chức cũng như chưa tính đến những thông tin liên quan đến các cú "sốc" chính trị, "sốc giá" diễn ra trong 5 tháng đầu năm 2017 vào mô hình dự báo đã gợi ý rằng cần kết hợp sử dụng mô hình dự báo mã cổ phiếu FPT được xây dựng với những phân tích định tính khác (nếu chưa lượng hóa được các yếu tố này để đưa vào mô hình) khi dự báo giá cổ phiếu FPT.

TÀI LIỆU THAM KHẢO

[1]. A¨ıt-Sahalia, Y., Dacheng Xiu, D. (2015), *Principal Component Analysis of High Frequency Data*, Working paper, Princeton University and University of Chicago, 47 pages, March 2015.

[2]. Bollerslev, T., Chou, R.Y. and Kroner, K.F. (1992), *ARCH Modeling in Finance*, Journal of Econometrics, 52, 5-59.

[3]. Chandrashekar, G., Sahin, F. (2014), *A survey on feature selection methods*, Computers and Electrical Engineering 40, 16-28.

[4]. Claessen, H., Mittnik, S. (2002), *Forecasting Stock Market Volatility and the Informational Efficiency of the DAX-index Options Market*, Working Paper No 2002/4, Center for Financial Studies, Germany.

[5]. Diamantini, C., Potena, D. (2008), *Chapter 6: A Study of Feature Extraction Techniques Based on Decision Border Estimate*, in Book: Computational Methods of Feature Selection, editors: Huan Liu and Hiroshi Motoda, Chapman & Hall/CRC.

[6]. 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*, Kỷ yếu Hội nghị FAIR9, 2016, Cần Thơ, 04-05/8/2016, 299-308. DOI: 10.15625/vap.2016.00069.

[7]. Đỗ Văn Thành và Nguyễn Minh Hải (2016), *Mô hình dự báo tần suất cao đối với chỉ số thị trường chứng khoán, Kỷ yếu Hội nghị FAIR9*, 2016, Cần Thơ, 04-05/8/2016, 559-566. DOI: 10.15625/vap.2016.00037.

[8]. Enders, W. (2014), *Applied Econometric Time Series*, 4th Edition, Wiley: USA, 2014.

[9]. Engle, R.F. (1982), *Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of the U.K. Inflation*, Econometrica, 50, 987-1008.

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

[11]. Granger, C. W. J. (1969), *Investigating Causal Relations by Econometric Models and Cross-Spectral Methods*, Econometrica, 37, 424-438.

[12]. Greene W. H. (2012), *Econometric Analysis*, New York University, Seventh Edition, Prentice Hall, 2012.

[13]. Guyon, I., Elisseeff, A. (2003), *An Introduction to Variable and Feature Selection*, Journal of Machine Learning Research 3 (2003) 1157-1182.

[14]. Hargreaves, C. A., Mani, C. K. (2015), *The Selection of Winning Stocks Using Principal Component Analysis*, American Journal of Marketing Research, Vol. 1, No. 3, 2015, pp. 183-188.

[15]. Hou, C., Nie, F., Yi, D. and Wu, Y., *Feature Selection via Joint Embedding Learning and Sparse Regression*, Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, 1325-1329.

[16]. Jensen, M. (1978), *Some anomalous evidence regarding market efficiency*, Journal of Financial Economics, 6 (2/3), 95-101.

[17]. Preethi, G. and Santhi, B. (2012), *Stock Market Forecasting Techniques*: A Survey, Journal of Theoretical and Applied Information Technology, Vol 46, No 1, 2012, pp. 24-30.

[18]. Sorzano, C. O. S. , Vargas, J. , & Pascual-Montano, A. (2014). *A survey of dimensionality reduction techniques*, Cornell University Library Abstracts (1-35).

[19]. Thanh D. V, Hai N. M. and Hieu D. D., (2016): *Building unconditional forecast model of Stock Market Indexes using combined leading indicators and principal components: application to Vietnamese Stock Market* (submited for publication).

[20]. Van Der Maaten, L. , Postma, E. , & Van den Herik, J. (2009), *Dimensionality reduction: A comparative*, Journal of Machine Learning Research, 10 (1-41), 66-71.

[21]. Yanshan, W., Choi, I. C. (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.

[22]. Zhong, X., Enke, D. (2017), *Forecasting daily stock market return using dimensionality reduction*, Expert Systems With Applications 67 (2017) 126-139. DOI: 10.1016/j.eswa. 2016.09.027.

[23]. Weinberger, K. Q., & Saul, L. K. (2006), *An Introduction to Nonlinear Dimensionality Reduction by Maximum Variance Unfoldin*g, http://www.aaai.org/Papers/AAAI/2006/AAAI06-280.pdf 1683-1686;

[24]. www.forecast.org.

[25]. www.eviews.com.

MODELLING OF A STOCK'S PRICE FORECAST IN THE CONTEXT OF HIGH DIMENSIONAL DATA SET

Thanh Do Van

ABSTRACT: *Forecasting stock prices has always been of particular interest and are always considered one of the most difficult forecasts in the socio-economic field due to volatility and its unpredictable fluctuations. The purpose of this paper is to present the modeling of a stock's price forecast based on the set of all economic- financial variables affecting the fluctuations of this stock price. These variables in general are not completely independent of each other, and the number of variables as well as the number of observations for every variable are generally very large.*

The methodology of building the forecast model of a stock's price proposed in the paper will use a combination of attribute selection and learn techniques to transform high - dimensional data sets to low-dimensional data ones so that where the information in the high dimensional data sets as well as relationships between this stock's price with other economic - financial variables are retained as much as possible. The paper uses the autoregressive distributed lag model to build the forecast model of average of the stock's price and uses one of the models in the family of autoregressive conditional heteroscedasticity models to build the forecast model of uncertainty of the residual variance. The forecasted results using the built models show good prospects of the methodology and the methodology proposed in this paper can be considered as the guidlines to practique modelling of price forecast of other goods and services.

Keyword: *dimentionality reduction, high dimensional data, stock price, the ARCH model, modeling financial forecasts.*