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.2017.00056

MÔ PHỎNG VÀ PHÂN TÍCH RỦI RO DỰ BÁO TRÊN TẬP DỮ LIỆU SỐ CHIỀU CAO

Đỗ Văn Thành[1] và Đỗ Đức Hiếu[2]

[1]Khoa Công nghệ thông tin, Trường Đại học Nguyễn Tất Thành

[2]Khoa Công nghệ thông tin và Truyền thông, Trường Đại học Khoa học và Công nghệ Hà Nội, Viện Hàn lâm Khoa học và Công nghệ Việt Nam,

dvthanh@ntt.edu.vn, vincentdo2310@gmail.com

TÓM TẮT: *Sử dụng mô hình định lượng để dự báo là đã thừa nhận rằng tương lai diễn ra gần giống như hiện tại và quá khứ. Nhưng thực tế không phải như vậy bởi vậy việc dự báo luôn là khó khăn và có nhiều trường hợp kết quả dự báo là khác xa hoặc thậm chí là trái ngược với thực tế mặc dù mô hình dự báo được chẩn đoán và kiểm định cẩn thận và được đánh giá là rất phù hợp. Hiện tượng này được gọi là rủi ro dự báo. Rủi ro dự báo thường được đo bằng xác suất xuất hiện của nó.*

Mục đích của bài báo này là đề xuất khung lý thuyết để mô phỏng và phân tích rủi ro dự báo trong ngữ cảnh các yếu tố ảnh hưởng đến vấn đề (hay biến đích) cần được dự báo là rất lớn. Bài báo sử dụng kỹ thuật giảm chiều dữ liệu để chuyển tập dữ liệu số chiều cao (số biến gốc và/hoặc số quan sát là rất lớn) thành tập dữ liệu có số chiều thấp sao cho quan hệ giữa biến đích và các biến gốc thay đổi ít nhất có thể và về cơ bản tập dữ liệu có số chiều thấp phản ánh được khá đầy đủ thông tin trong tập dữ liệu số chiều cao. Đồng thời bài báo cũng ứng dụng mô hình hồi quy lôgit hoặc mô hình hồi quy có thứ tự để xây dựng mô hình tính xác suất dự báo theo xu thế hoặc theo mức độ khác biệt của kết quả dự báo. Bài báo ứng dụng phương pháp luận được đề xuất trong lĩnh vực kinh tế - tài chính.

Từ khóa: *giảm chiều dữ liệu, dữ liệu số chiều cao, mô phỏng rủi ro dự báo, phân tích rủi ro, hồi quy lôgit và hồi quy có thứ tự.*

I. GIỚI THIỆU

Rủi ro là một từ được sử dụng thường xuyên trong đời sống thực. Theo [11], rủi ro có thể được định nghĩa là "khả năng xảy ra một sự kiện ngẫu nhiên, không thể đoán trước và có thể ảnh hưởng xấu đến kết quả dự kiến". Rủi ro được gây ra bởi những lỗ hổng bên ngoài hoặc bên trong và có thể tránh được thông qua hành động phòng ngừa.

Phân tích rủi ro có thể được định nghĩa theo nhiều cách khác nhau, nhưng định nghĩa được sử dụng rộng rãi cho rằng phân tích rủi ro bao gồm: đánh giá rủi ro, tìm ra đặc trưng của rủi ro, truyền thông rủi ro, quản lý rủi ro và các chính sách liên quan đến rủi ro. Phân tích rủi ro có thể là định tính hoặc định lượng. Phân tích rủi ro định tính sử dụng các từ hoặc màu sắc để xác định và đánh giá rủi ro hoặc đưa ra mô tả bằng văn bản về rủi ro, trong khi phân tích rủi ro định lượng là tính toán *xác suất đối với những hậu quả* có thể xảy ra [4, 9].

Rủi ro dự báo được hiểu là khả năng kết quả dự báo về một sự kiện nào đó là khác xa, thậm chí là trái ngược với thực tế. Trong [2, 4, 12] để đo lường định lượng, đánh giá mức độ không chắc chắn của dự báo (hay rủi ro dự báo) là lớn hay nhỏ, người ta thường sử dụng các metric khác nhau, trong đó 3 metric sau là phổ biến: (1) Kết hợp xác suất rủi ro với độ lớn/mức độ nghiêm trọng của hậu quả; (2) Bộ ba (s_i, p_i, c_i), trong đó s_i là kịch bản thứ i, p_i là xác suất của kịch bản đó và c_i là hệ quả của kịch bản thứ i, i = 1,2, ... N; (3) Bộ ba (C ', Q, K), trong đó C' là một số hậu quả cụ thể, Q là độ đo không chắc chắn (thường là xác suất) của C' và K kiến thức nền hỗ trợ C' và Q. Nói cách khác dù sử dụng metric nào thì xác suất xảy ra của kết quả dự báo được sử dụng để đánh giá mức độ rủi ro dự báo.

Ký hiệu Y là biến định lượng đo lường một sự kiện kinh tế - xã hội nào đó (gọi là biến đích), X_i (i =1, 2, …, n) là các biến đo lường các yếu tố kinh tế - xã hội khác có thể có ảnh hưởng đến sự thay đổi của Y, được gọi là các biến gốc. Việc nhận biết giá trị và/hoặc xu hướng thay đổi trong tương lai của biến đích Y có thể được dự báo thông qua các biến X_i. Dự báo bằng sử dụng các kỹ thuật dự báo định lượng (kỹ thuật thống kê hoặc kỹ thuật trí tuệ nhân tạo) [13] thực chất là đã thừa nhận rằng các biến gốc cũng như quan hệ giữa biến đích và các biến gốc trong tương lai sẽ được thay đổi gần giống với quy luật của chúng trong hiện tại và quá khứ. Thừa nhận này cơ bản là đúng đối với những hiện tượng thuộc lĩnh vực tự nhiên, nhưng phần lớn là không đúng đối với những hiện tượng thuộc lĩnh vực xã hội, trong đó nhất là lĩnh vực kinh tế - tài chính, ở đó có sự tham gia hoạt động của con người. Hành vi của con người phụ thuộc nhiều vào những yếu tố rất khó có thể định lượng được như văn hóa, tâm lý dân tộc, đặc điểm tôn giáo, đặc điểm vùng miền,…, và nhất là tâm trạng của họ ở mỗi hoàn cảnh, thời điểm cụ thể. Điều đó ngầm ý dự báo kinh tế - xã hội là rất khó khăn. Thực tế cho thấy chỉ sử dụng các công cụ định lượng để dự báo kinh tế - xã hội là chưa đủ để có được kết quả dự báo chính xác cho dù các công cụ định lượng này có cơ sở khoa học tốt đến đâu.

Vấn đề đánh giá mức độ rủi ro dự báo đã được đặt ra từ những năm 50 của thế kỷ XX, và cho đến nay nó vẫn là vấn đề thời sự và luôn được quan tâm khi thực hiện dự báo trong thế giới thực. Việc đánh giá rủi ro dự báo biến đích theo các biến gốc được thực hiện theo cách như sau: dựa vào biến đích chia tập dữ liệu của các biến gốc thành một số tập dữ liệu con và gán nhãn cho mỗi tập con này (còn được gọi là xác định giá trị phân loại cho biến đích), xây dựng thuật toán phân lớp tập dữ liệu đầu vào theo các nhãn được xác định và với mỗi bộ dữ liệu đầu vào (hay một quan sát) mới nào đó của các biến gốc, thuật toán sẽ cho biết giá trị xác suất dự báo biến đích Y không thuộc hoặc thuộc vào mỗi lớp có nhãn được xác định.

Hiện tại đã có khá nhiều kỹ thuật đánh giá rủi ro dự báo và nó có thể chia thành 2 nhóm: kỹ thuật trí tuệ nhân tạo và kỹ thuật thống kê. Các kỹ thuật trí tuệ nhân tạo chủ yếu được sử dụng để đánh giá rủi ro dự báo bao gồm: cây quyết định xác suất, mạng nơron xác suất và học mạng Bayes,… trong đó 2 kỹ thuật đầu được sử dụng phổ biến nhất [15]. Trong cấu trúc của cây quyết định, các lá biểu diễn các nhãn lớp, các nhánh tượng trưng cho các kết nối của các thuộc tính (hay biến gốc) dẫn đến các nhãn lớp đó. Trong cây quyết định xác suất [3], mỗi kết nối 2 thuộc tính được gán một giá trị xác suất có điều kiện và như vậy theo quy tắc Bayes có thể tính được xác suất của các lá ứng với bộ giá trị đầu vào cụ thể của các thuộc tính (hay biến gốc). Mạng nơtron xác suất (PNN) là một mạng nơron truyền thẳng, được sử dụng rộng rãi trong các vấn đề phân lớp và nhận dạng mẫu. Trong thuật toán PNN, hàm phân bố xác suất cha (PDF) của mỗi lớp được xấp xỉ bằng một cửa sổ Parzen và một hàm phi tham số [15]. Xác suất lớp của một bộ dữ liệu đầu vào mới được tính dựa vào sử dụng PDF của mỗi lớp và quy tắc Bayes. Xác suất lớp cao nhất được sử dụng để phân bổ lớp cho dữ liệu đầu vào mới [15]. Trong đánh giá rủi ro dự báo, các kỹ thuật trí tuệ nhân tạo thường cung cấp xác suất rủi ro dự báo với độ chính xác khá cao và các kỹ thuật này về cơ bản cũng có ưu, nhược điểm chính như được trình bầy chi tiết trong [13], theo đó nhược điểm đáng lưu ý nhất của các kỹ thuật trí tuệ nhân tạo là không đưa ra được công thức tính xác suất rủi ro dự báo một cách tường mình và không đánh giá được tác động của từng yếu tố đầu vào đến xác suất dự báo do đó nó sẽ hạn chế trong việc mô phỏng, phân tích và đề xuất các giải pháp phòng ngừa rủi ro.

Các kỹ thuật thống kê được sử dụng trong đánh giá rủi ro dự báo có thể được kể đến gồm: phân lớp Bayes, phân tích phân biệt (tuyến tính cũng như phi tuyến), hồi quy lôgit, …trong đó hai kỹ thuật được xem là hiệu quả nhất và được sử dụng rộng rãi nhất là phân tích phân biệt và hồi quy lôgit [6-7, 10, 14]. Phân tích phân biệt nhằm phân các quan sát được mô tả bằng giá trị về các biến liên tục (hay biến nhận giá trị số thực) thành các lớp. Thành viên của lớp, được xác định bởi một biến phân lớp và được dự báo bởi các biến liên tục bằng cách ước tính khoảng cách Mahalanobis từ mỗi quan sát đến trung tâm (centroid) của các lớp. Quan sát được phân vào lớp gần nhất và xác suất tiên nghiệm của thành viên lớp được tính trong khi tính khoảng cách [10]. Phân tích hồi quy lôgit nghiên cứu sự kết hợp giữa biến đích (hay biến phụ thuộc) nhận giá trị phân loại và một tập các biến gốc (hay biến giải thích) liên tục và độc lập với nhau. Xác suất đo sự kết hợp ấy được sử dụng để đánh giá rủi ro dự báo. Hồi quy lôgit bao gồm hồi quy lôgit đơn tên và hồi quy lôgit đa tên. Phương pháp hồi quy lôgit đơn tên (gọi tắt là hồi quy lôgit) được sử dụng khi biến phụ thuộc chỉ có hai giá trị phân loại. Phương pháp hồi quy lôgit đa tên (gọi là hồi quy có thứ tự), được phát triển mở rộng của phương pháp hồi quy lôgit, được sử dụng khi biến phụ thuộc nhận nhiều hơn 2 giá trị phân loại. Về bản chất hồi quy lôgit cũng như hồi quy thứ tự là một loại phân tích hồi quy được sử dụng để dự đoán một biến có thứ tự. Nó có thể được coi là một vấn đề trung gian giữa hồi quy và lớp. Cây quyết định xác suất ở đó các thuộc tính (biến gốc) là liên tục (tức nhận giá trị số thực) là rất gần với mô hình hồi quy lôgit.

So sánh các phương pháp phân tích phân biệt và họ hồi quy lôgit có thể thấy trong phân tích phân biệt, các lớp đã được xác định trước và được cố định trong khi trong họ hồi quy lôgit, biến phân lớp là ngẫu nhiên, tuy nhiên trong cả hai kỹ thuật này, giá trị phân loại đều được dự báo bởi các biến liên tục [8]. Họ hồi quy lôgit được đánh giá là linh hoạt và phù hợp hơn cho việc mô hình hóa hầu hết các tình huống so với phân tích phân biệt vì hồi quy lôgit đòi hỏi ít giả thiết hơn về tập dữ liệu gốc. Chẳng hạn trái với họ hồi quy lôgit, phân tích phân biệt hạn chế trong việc xử lý dữ liệu ngoại lai; phân tích phân biệt đòi hỏi các biến gốc phải độc lập và có phân phối chuẩn trong khi họ hồi qui lôgit chỉ đòi hỏi tính độc lập; phân tích phân biệt đòi hỏi kích cỡ mỗi lớp phải lớn hơn số biến gốc trong khi họ hồi quy lôgit không có đòi hỏi này, …. Khi tập dữ liệu đầu vào đều thỏa mãn các đòi hỏi cần thiết để thực hiện phân tích phân biệt thì xác suất dự báo của 2 kỹ thuật này là xấp xỉ nhau [8].

Các kỹ thuật phân tích phân biệt và họ hồi quy lôgit đều đòi hỏi các biến gốc đầu vào phải độc lập với nhau, trong khi kỹ thuật phân lớp Bayes không đòi hỏi điều kiện này. Nói cách khác khi các biến gốc là độc lập việc lựa chọn họ mô hình hồi quy lôgit để xây dựng mô hình mô phỏng và phân tích rủi ro dự báo là lựa chọn chính xác và hiệu quả nhất, trong trường hợp trái lại thì sử dụng phương pháp phân lớp Bayes. Ngoài ra các kỹ thuật phân tích phân biệt và họ hồi quy lôgit trong mô phỏng và đánh giá rủi ro cũng có các ưu, nhược điểm chính của kỹ thuật dự báo thống kê đã được trình bầy chi tiết trong [13], trong đó hạn chế lớn nhất của các kỹ thuật này là không thực hiện được trên tập dữ liệu số chiều cao và không tự động hóa được toàn bộ quá trình đánh giá rủi ro dự báo.

Những phân tích về ưu, nhược điểm chính của các kỹ thuật đánh giá rủi ro dự báo ở trên đã gợi ý rằng việc lựa chọn các mô hình hồi quy lôgit và hồi quy có thứ tự để tính xác suất dự báo, mô phỏng và phân tích rủi ro dự báo trên tập dữ liệu số chiều cao trong thế giới thực dường như là giải pháp phù hợp nhất cho đến thời điểm hiện tại nếu như có thể khắc phục được 02 hạn chế của lớp mô hình hồi quy lôgit là không thực hiện được trên tập dữ liệu lớn và các biến gốc đòi hỏi phải là các biến độc lập. Rất may, cả 2 yêu cầu này đều được khắc phục bằng ứng dụng phương pháp giảm chiều dữ liệu được tác giả bài báo này đề xuất trong [13].

Mục đích của bài báo này là đề xuất khung lý thuyết để xây dựng mô hình mô phỏng và phân tích rủi ro dự báo trên tập dữ liệu số chiều cao ở đó các biến gốc có thể không có hoặc có tác động rất ít đến sự thay đổi của biến đích; các biến gốc có thể không độc lập với nhau, thậm chí có thể còn tương quan mạnh với nhau; trong tập dữ liệu đầu vào có thể có dữ liệu ngoại lai (dữ liệu bất thường).

Điểm mấu chốt của khung lý thuyết này trước hết là phải thực hiện những kỹ thuật giảm chiều dữ liệu để chuyển đổi tập dữ liệu đầu vào có số chiều cao thành tập dữ liệu số chiều thấp sao cho quan hệ giữa biến đích và các

biến gốc đầu vào được bảo toàn nhiều nhất như có thể, đồng thời các biến mới thay thế cho tập các biến gốc đầu vào đều có tác động thực sự đến sự thay đổi của biến đích và chúng độc lập với nhau. Kỹ thuật giảm chiều dữ liệu trong bài báo này là sự kết hợp của kỹ thuật lựa chọn thuộc tính và học thuộc tính theo cách tương tự như trong [13]. Điểm mấu chốt tiếp theo là sử dụng kỹ thuật hồi quy lôgit hoặc hồi quy có thứ tự để xây dựng mô hình dự báo phân lớp để thực hiện mô phỏng và phân tích rủi ro dự báo. Mô hình dự báo như vậy cho biết xác suất của việc dự báo sai xu hướng cũng như xác suất của sự khác biệt về kết quả dự báo biến đích theo tập các biến gốc đầu vào khi ứng dụng mô hình dự báo được đề xuất trong [13]. Việc phát hiện và xử lý dữ liệu ngoại lai trong tập dữ liệu mới sẽ được thực hiện trong quá trình xây dựng mô hình dự báo phân lớp.

Trong bài báo này khung lý thuyết sẽ được ứng dụng để xây dựng mô hình dự báo, thực hiện mô phỏng và phân tích rủi ro dự báo giá của cổ phiếu nào đó dựa trên dữ liệu thực của nền kinh tế.

Phần còn lại của bài báo được cấu trúc như sau, tiếp theo phần này, Phần II trình bầy cụ thể vấn đề đặt ra và đề xuất phương pháp giải quyết. Phần III sẽ ứng dụng phương pháp được đề xuất để mô phỏng và phân tích rủi ro 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 đích, X_i (i =1, 2, …, n) là các 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 biến véc tơ, $Y^T = (y_j)$, $X_i^T = (x_{ji})$, j =1, 2, …, m; bộ (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, i =1, 2, …, n). 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 không độc lập với nhau. Giả sử số biến gốc n là rất lớn.

Vấn đề đặt ra: xây dựng Mô hình dự báo xác suất dự báo, mô phỏng và phân tích rủi ro dự báo 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 bầy một cách tóm tắt khung lý thuyết của quá trình xây dựng Mô hình dự báo xác suất dự báo, mô phỏng và phân tích rủi ro dự báo biến đích trên tập dữ liệu số chiều biến cao của các biến gốc. Quá trình này gồm 2 giai đoạn cơ bản là giảm chiều dữ liệu của tập dữ liệu đầu vào và xây dựng mô hình dự báo để tính xác suất dự báo, thực hiện mô phỏng và phân tích rủi ro dự báo trên tập dữ liệu mới.

2.3. Pha 1: Giảm chiều dữ liệu

Phương pháp giảm chiều dữ liệu trong bài báo này là hoàn toàn tương tự như trong [13] nên sẽ không được nhắc lại ở đây.

2.4. Pha 2: Xây dựng Mô hình dự báo xác suất, mô phỏng và phân tích rủi ro

Rủi ro dự báo có thể xẩy ra khi kết quả dự báo thuộc hướng trái ngược với dữ liệu thực tế hoặc khi có sự khác biệt khá lớn giữa kết quả dự báo với thực tế. Ở trường hợp đầu có thể hiểu giá trị dự báo của biến đích không cùng trong một xu thế biến động trong quá khứ của nó. Ở trường hợp sau được hiểu là giá trị dự báo và giá trị ở quan sát ngay trước đó của biến đích có những khác biệt lớn và được phân thành k > 2 mức độ khác biệt khác nhau. Khung lý thuyết xây dựng mô hình mô phỏng và phân tích rủi ro dự báo cho 2 trường hợp này cơ bản là giống nhau, chúng chỉ khác nhau ở chỗ thực hiện gán 2 nhãn (hay giá trị phân loại) hoặc k nhãn (k>2) cho các bộ giá trị của các biến gốc và khi đó sẽ sử dụng mô hình hồi quy lôgit hoặc mô hình hồi quy có thứ tự để xây dựng Mô hình dự báo xác suất, mô phỏng và phân tích rủi ro dự báo. Cụ thể của Khung lý thuyết được minh họa trong Hình 1 ở dưới.

2.4.1. Gán nhãn cho tập dữ liệu đầu vào của các biến gốc

Ký hiệu Z = {z_i}, i =1, …, m là biến nhãn lớp của tập dữ liệu đầu vào. Ở trường hợp thứ nhất ta chọn 2 nhãn, chẳng hạn là 0 và 1 là miền giá trị của biến Z. Khi đó Z ={z_i} có thể được xác định theo biến đích Y = {y_i}như sau:

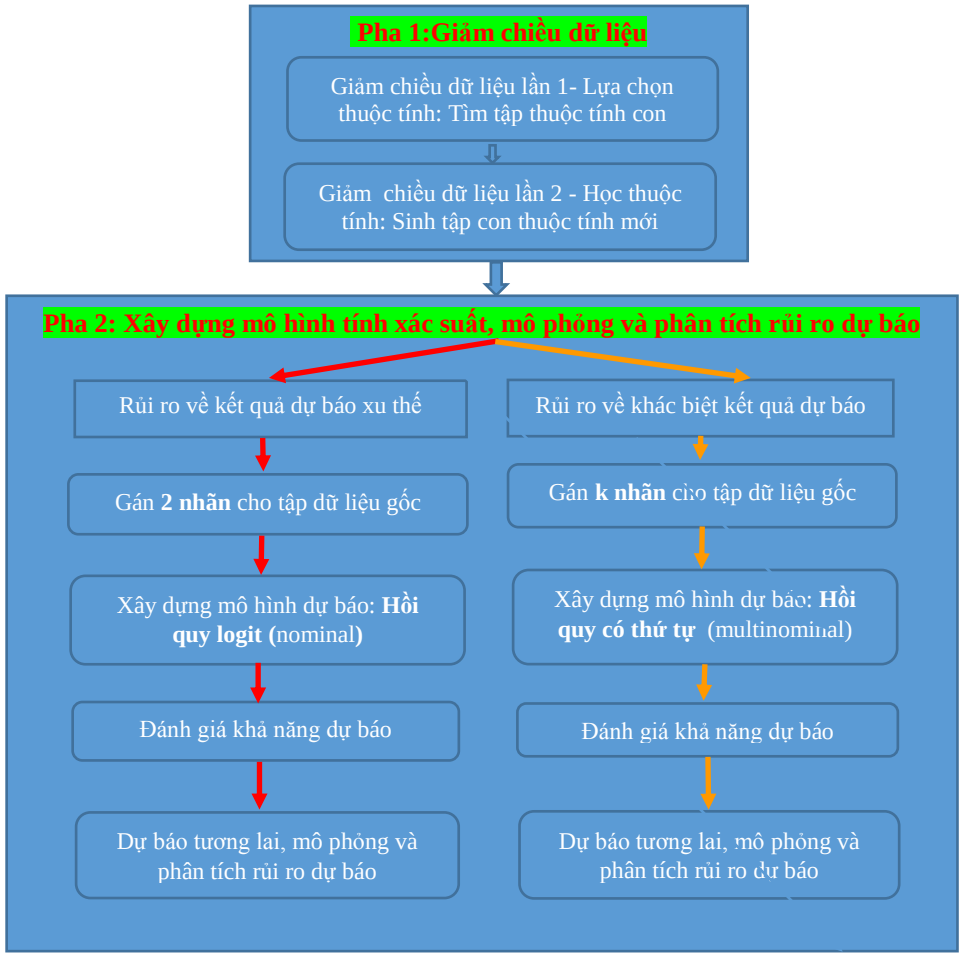
$$z_i = \begin{cases} 1 \ nếu \ y_i \geq y_{i-1} \\ 0, \ ngược \ lại. \end{cases} \text{ với mọi i >1.} \quad (1)$$

Và vấn đề đặt ra ban đầu của bài báo trở thành: xây dựng Mô hình dự báo xác suất kết quả dự báo biến nhãn lớp Z nhận giá trị 1 theo bộ dữ liệu x_{ji} của các biến gốc X_j, ở đây j =1, 2, …, n và i là quan sát thứ i nào đó.

Với trường hợp thứ 2, ta có thể chọn các nhãn 1, 2, …, K để gán cho biến nhãn lớp Z = {z_i}, i = 1, …, m. Khi đó miền giá trị của Z là 1, 2, …, K và các z_i được xác định dựa vào giá trị y_i của biến đích Y theo một cách nào đó. Chẳng hạn:

$$z_i = \begin{cases} 1, & nếu \ \frac{(y_i - y_{i-1})}{y_{i-1}} < \alpha_1 \\ h, & nếu \ \alpha_{h-1} \leq \frac{(y_i - y_{i-1})}{y_{i-1}} < \alpha_h, \ 1 < h \leq k-1 \\ K, & nếu \ \alpha_{k-1} \leq \frac{(y_i - y_{i-1})}{y_{i-1}} \end{cases} \quad (2)$$

ở đây 1< i ≤ m, và $\alpha_1 < \alpha_2 < \cdots < \alpha_{k-1}, \alpha_i \in R$. Khi đó vấn đề đặt ra của bài báo này trở thành xây dựng Mô hình dự báo xác suất kết quả dự báo biến nhãn lớp Z có giá trị là h (h=1, 2, …, K) trong điều kiện các biến gốc X_j nhận giá trị tương ứng là x_{ji}, với j =1, 2, …, n và i là quan sát thứ i nào đó.



Hình 1. Khung lý thuyết mô phỏng và phân tích rủi ro dự báo trên tập dữ liệu số chiều cao

2.4.3. Mô hình hồi quy lôgit và hồi quy có thứ tự

a). Mô hình hồi quy lôgit

Giả sử các biến gốc $X_1, X_2, \ldots, X_n$ là các *biến độc lập*. Xác suất dự báo biến Z nhận giá trị nhãn là 1 ứng với bộ dữ liệu quan sát của các biến gốc X_i, được ký hiệu là $P_r(Z=1| X_1, X_2, \ldots X_n)$, được xác định như sau [5-6]:

$$P_r(Z = 1|X_1, X_2, \ldots, X_n) = \frac{1}{1+e^{-(\alpha+\sum_{i=1}^{n}\beta_i X_i)}} \quad (3)$$

Các tham số α và β_i được xác định bằng cách ước lượng theo phương pháp cực đại hợp lý (maximum likelihood) [5] trên tập dữ liệu đầu vào.

Ký hiệu p = $P_r(Z = 1|X_1, X_2, \ldots, X_n)$ và $odds = \frac{p}{1-p}$, thế thì:

$$\ln(\text{odds}) = \alpha + \sum_{i=1}^{n}\beta_i X_i \quad (4)$$

Phương trình (4) cho biết rằng khi X_i tăng hoặc giảm 1 đơn vị còn các biến khác giữ nguyên thì odds sẽ tăng hoặc giảm β_i%, và ln(odds) = α khi tất cả các biến độc lập X_i có giá trị bằng 0, nhưng trong thực tế hiếm có những trường hợp như vậy nên giải thích này thường ít được nhắc tới. Tỷ số odds cho biết sự kiện Z =1 (hay biến đích tăng so với quan sát ngay trước đó) xảy ra bằng bao nhiêu lần so với sự kiến Z=0 (hay biến đích giảm so với quan sát trước đó).

Rủi ro của tỷ số odds: giả sử p_0, p_1 là xác suất để Z =1 tương ứng tại các quan sát k_0 và k_1. Từ công thức (4) ta có:

$$\ln(\text{odds}(k_0)) = \alpha + \sum_{i=1}^{n}\beta_i X_i\,(k_0) \text{ và } \ln(\text{odds}(k_1)) = \alpha + \sum_{i=1}^{n}\beta_i X_i\,(k_1)$$

Do đó $\ln(\text{odds}(k_1)) - \ln(\text{odds}(k_0)) = \sum_{i=1}^{n}\beta_i(X_i\,(k_1) - X_i(k_0))$ hay

$$\frac{odds(k_1)}{odds(k_0)} = e^{\sum_{i=1}^{n} \beta_i (X_i(k_1) - X_i(k_0))} \quad (5)$$

Giả sử $X_i(k_1) - X_i(k_0) = 1$ và $X_j(k_1) - X_j(k_0) = 0$ với mọi j ≠ i, từ phương trình (5) nhận được

$$\frac{odds(k_1)}{odds(k_0)} = e^{\beta_i} \quad (6)$$

Công thức (6) có nghĩa là trong điều kiện các yếu tố (biến) khác không đổi, nếu biến X_i tăng/giảm 1 đơn vị thì tỷ số $\frac{odds(k_1)}{odds(k_0)}$ sẽ tăng/giảm e^{β_i} lần và khi đó xác suất để Z=1 tại quan sát k_1 là p_{k_1} được xác định bởi công thức sau:

$$p_{k_1} = \frac{p_{k_0}.e^{\beta_i}}{1 - p_{k_0}.(1 - e^{\beta_i})} \quad (7)$$

Công thức (6) và (7) được sử dụng để phân tích rủi ro dự báo.

b). Mô hình hồi quy có thứ tự

Giả sử các biến gốc X_1, X_2,…, X_n là các biến độc lập; 1, 2, …, K là các nhãn lớp của biến Z. Xác suất dự báo biến Z nhận giá trị h ứng với bộ dữ liệu quan sát được của các biến gốc X_i được xác định bởi các phương trình (5) và (6) [5-6]:

$$P_r(Z = h|X_1, X_2, \dots, X_n) = \frac{e^{\sum_{i=1}^{n} \beta_{hi}.X_i}}{1 + \sum_{q=1}^{K-1} e^{\sum_{j=1}^{n} \beta_{qj}.X_j}} \quad (8)$$

$$P_r(Z = K|X_1, X_2, \dots, X_n) = \frac{1}{1 + \sum_{q=1}^{K-1} e^{\sum_{j=1}^{n} \beta_{qj}.X_j}} \quad (9)$$

ở đây các tham số β_{hi}, β_{qj} được ước lượng bằng phương pháp cực đại ước lượng hậu nghiệm (maximum a posteriori estimation), nó là sự mở rộng của phương pháp ước lượng cực đại hợp lý nêu trên [5]; h, q =1, K-1.

c). *Chẩn đoán và làm phù hợp mô hình dự báo*

Việc chẩn đoán cũng như làm phù hợp mô hình hồi quy lôgit cũng như hồi quy có thứ tự cơ bản là được thực hiện tương tự như đối với mô hình hồi quy nhiều biến [1, 10]. Tuy nhiên có thể thực hiện một số kiểm định thống kê khác như các kiểm định Andrews và Hosmer-Lemeshow [1] đối với lớp mô hình lôgit để làm phù hợp mô hình dự báo.

Trong quá trình ước lượng các tham biến của mô hình dự báo cần luôn kiểm tra phần dư để nhận biết có dữ liệu ngoại lai trong tập dữ liệu đầu vào hay không (bằng cách kiểm tra xem có những quan sát nào ở đó có trị tuyệt đối phần dư vượt quá 1,5 lần giá trị tới hạn của nó). Nếu có, cách thông dụng và hiệu quả nhất là sử dụng biến giả (dummy) để xử lý dữ liệu bất thường đó nếu số quan sát là lớn. Trường hợp số quan sát là nhỏ thì sử dụng các kỹ thuật làm trơn số liệu (như sử dụng phương pháp trung bình trượt có trọng số, phương pháp hồi quy, …) để xử lý dữ liệu bất thường.

2.4.4. Đánh giá khả năng dự báo

Nhằm đánh giá khả năng dự báo ngoài mẫu của mô hình hồi quy lô git và mô hình hồi quy có thứ tự, cần chia tập dữ liệu đầu vào thành 2 tập. Tập thứ nhất được sử dụng để xây dựng mô hình dự báo theo công thức (3) hoặc (5) và (6) sau đó sử dụng mô hình này để tính xác suất dự báo thuộc mỗi lớp cho các quan sát trong tập dữ liệu thứ 2. Các lớp "chính xác" nhận được khi xác suất dự báo là nhỏ hơn hoặc bằng một giá trị "cắt" (cutoff) nằm giữa 0 và 1 (ngầm định là 0.5) do người dùng xác định và nhãn lớp là z_i=0 hoặc khi xác suất dự báo là lớn hơn giá trị "cắt" và nhãn lớp z_i=1.

2.4.5. Dự báo tương lai, mô phỏng và phân tích rủi ro dự báo

a. Dự báo tương lai

Để đánh giá mức độ rủi ro trong tương lai về kết quả dự báo biến đích cần:

- Thực hiện dự báo các biến gốc X_1, X_2, …, X_n trong mô hình dự báo được xây dựng theo phương trình (3) hoặc theo các phương trình (8) và (9) bằng việc sử dụng mô hình tự hồi quy AR(p) có xu thế được xác định bởi phương trình:

$$\Delta X(t) = \alpha + \rho X(-1) + \gamma_1 \Delta X(-1) + \dots + \gamma_p \Delta X(-p) + \delta t + e_t, \quad (10)$$

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

- Thực hiện ước lượng mô hình dự báo trên toàn bộ tập dữ liệu ban đầu, sau đó sử dụng mô hình này để tính xác suất dự báo biến Z nhận giá trị theo từng nhãn ứng với từng bộ giá trị của các biến gốc vừa được dự báo, từ đó ta sẽ đánh giá được khả năng rủi ro dự báo của biến đích Y trong tương lai.

- Với mỗi bộ giá trị khác nhau của các biến gốc ta có một kịch bản dự báo của biến đích. Các biến gốc được dự báo bởi mô hình được xây dựng theo phương trình (10) là thừa nhận rằng sự thay đổi của các biến gốc trong tương lai được diễn ra gần giống như hiện tại và quá khứ, và kịch bản dự báo trong trường hợp này được gọi là kịch bản cơ sở.

b. Mô phỏng và phân tích rủi ro dự báo

Thực tế cho thấy rằng trong rất nhiều trường hợp tương lai diễn ra khác xa hiện tại và quá khứ, Khi đó dự báo biến đích ở kịch bản cơ sở thường khác xa so với thực tế cả ở phương diện giá trị kết quả dự báo và/hoặc xu thế của nó.

Mô phỏng dự báo nhằm đánh giá tác động của từng biến hoặc của một nhóm biến gốc nào đó đến kết quả dự báo biến đích khi so sánh nó với kịch bản cơ sở, từ đó giúp nhận diện được những rủi ro dự báo có thể xảy ra trong tương lai để có biện pháp thích nghi hoặc phòng ngừa phù hợp. Về bản chất mô phỏng dự báo là một cách dự báo ở đó giá trị của các biến gốc "nhạy cảm" cao đối với những thay đổi về chính sách điều hành của chính phủ, cũng như sự biến động của tình hình kinh tế, chính trị, xã hội trong và ngoài nước được xây dựng theo các giả định, các biến gốc còn lại được xem là diễn ra gần giống như hiện tại và quá khứ.

Trong ngữ cảnh tập dữ liệu đầu vào chiều cao, việc tiến hành mô phỏng dự báo không nên được thực hiện trên các thành phần chính (là các biến thay thế cho tập biến gốc) mà vẫn phải thực hiện trên tập các biến gốc ban đầu.

III. ỨNG DỤNG PHƯƠNG PHÁP

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 này được xác định theo cách tiếp cận như trong [13]. Cụ thể, các biến số phản ánh các điều kiện phát triển chung của nền kinh tế gồm 6 biến: 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. Các biến số liên quan đến biến động giá cổ phiếu FBT có 41 biến gồm: Giá của 29 mã cổ phiếu cổ phiếu BULUECHIP (bài báo sử dụng mã cổ phiếu làm 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 nhóm ngành, Công nghiệp: CNINDEX, Khoáng sản: KSINDEX, Ngân hàng: NHINDEX, Năng lượng: NLINDEX; Chỉ số gía tiêu dùng: CPI, chỉ số giá vàng: GOLDINDEX, chỉ số giá đô la ở Việt Nam: USINDEX; Các biến số phản ánh kinh tế- chính trị thế giới và của những nước lớn, các biến cố chính trị như sự xuất hiện và sự ra đời của các sự kiện chính trị quan trọng gồm 04 biến, các chỉ số chứng khoán NASDAD 100 và NASDAD 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; và chưa có biến số nào phản ánh 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, …

Chi tiết về các biến cũng như nguồn sản sịnh dữ liệu của chúng có thể tham khảo trong [13]. Số liệu của hầu hết các biến số này được thu thập theo ngày kể từ ngày 2/01/2012 đến hết ngày 31/05/2017 và giá trị trung bình theo các ngày trong tháng của mỗi biến số được xem là dữ liệu tháng của biến số đó. Như vậy tập dữ liệu đầu vào để xây dựng mô hình mô phỏng và phân tích rủi ro dự báo của biến FPT là dữ liệu của 65 quan cho 51 biến kinh tế - tài chính đã nêu ở trên.

3.2. Giảm chiều dữ liệu

3.2.1. Giảm chiều lần 1

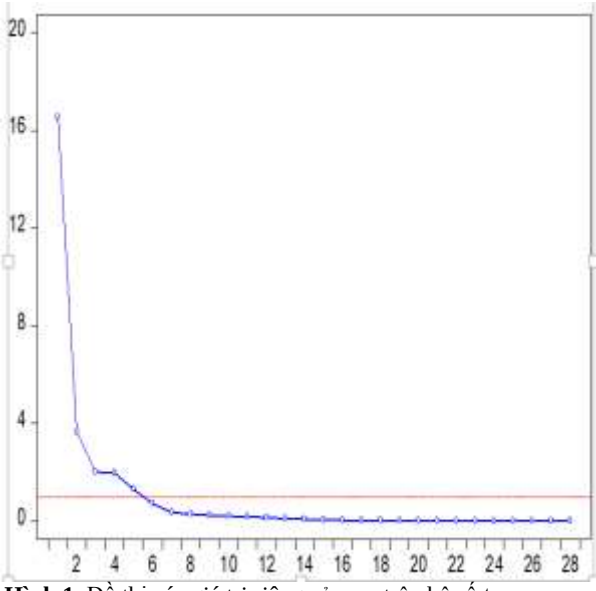
Phương pháp giảm chiều dữ liệu và tập dữ liệu đầu vào ở bài báo này là hoàn toàn tương tự như trong [13]. Theo đó tập con biến được lựa chọn bằng sử dụng thuật toán **ChonTapcon** chỉ gồm những biến có hệ số tương quan cao với biến đích FPT và không phải là biến dư thừa [13]. Tập con biến đó được chỉ ra trong Bảng 1 ở dưới:

Bảng 2. Tập con biến có ý nghĩa với biến FPT và không phải biến dư thừa của 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			

3.2.2. Giảm chiều lần 2:

Thực hiện phân tích PCA của 28 biến gốc phản ánh với dữ liệu của 60 quan sát từ tháng 1/2012 đến tháng 12/2016 theo quy trình được nêu trong [13], ta nhận được **Bảng 2** và **Hình 1**. Theo đó ta chỉ cần chọn 5 thành phần chính ứng với 5 giá trị riêng lớn nhất để làm đại diện cho tập các biến gốc đầu vào và 5 thành phần chính đó giải thích được đến 90.99% sự thay đổi trong tập dữ liệu của 51 biến gốc. 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.



Hình 1. Đồ thị các giá tri riêng của ma trận hệ số tương quan của 28 biến gốc

Số thứ tự	Giá trị riêng	Chênh lệch	Tỷ lệ	Giá trị riêng tích lũy	Tỷ lệ tích lũy
1	16.5212	12.8593	0.5900	16.5212	0.5900
2	3.6619	1.6483	0.1308	20.1831	0.7208
3	2.0136	0.0406	0.0719	22.1966	0.7927
4	1.9730	0.6655	0.0705	24.1696	0.8632
5	1.3075	0.5942	0.0467	25.4771	0.9099
6	0.7133	0.3610	0.0255	26.1903	0.9354
7	0.3523	0.0631	0.0126	26.5426	0.9479
8	0.2892	0.0475	0.0103	26.8318	0.9583
9	0.2417	0.0334	0.0086	27.0735	0.9669
10	0.2083	0.0258	0.0074	27.2818	0.9744
...	...	...	...	...	...

Bảng 2. Các giá trị riêng và tỷ lệ tích lũy

3.3. Xây dựng mô hình dự báo, mô phỏng và phân tích rủi ro dự báo

3.3.1. Xây dựng mô hình mô phỏng và phân tích rủi ro dự báo xu thế

a. Gán nhãn cho tập dữ liệu gốc

Việc gán nhãn cho tập dữ liệu đầu vào của 5 thành phần chính dựa vào giá trị của biến đích FPT và cho 60 quan sát đầu tiên ứng với các tháng từ tháng 1/2012 đến tháng 12/2016 được thực hiện theo công thức (1). Bây giờ ta có thể xây dựng mô hình dự báo biến Z theo 5 thành phần chính được chọn PC_1, PC_2, PC_3, PC_4, PC_5 như là các biến gốc mà không sợ hồi quy sai do các thành phần chính là trực giao nên chúng độc lập với nhau.

b. Xây dựng mô hình dự báo

Thực hiện phương pháp hồi quy cực đại hợp lý biến Z theo 5 thành phần chính theo các công thức (3) trên tập dữ liệu gồm 60 quan sát đầu tiên, đồng thời sử dụng kiểm định WALD [10] để kiểm tra xem những biến thành phần chính nào sẽ nằm trong mô hình dự báo một cách có ý nghĩa thống kê, ta sẽ nhận được:

$$P_r(Z=1/PC_1, PC_2, PC_3, PC_4, PC_5) = \frac{1}{1+e^{-(0.592*PC4 + 0.813*PC5 + 0.706)}} \quad (11)$$

c. Đánh giá khả năng dự báo

Sử dụng mô hình được xác định theo công thức (11) tính xác suất dự báo để Z =1 cho 5 quan sát từ tháng 1/2017 đến tháng 5/2017 với dữ liệu của 5 thành phần chính PC_1, PC_2, PC_3, PC_4, PC_5. Ta nhận được Bảng 3:

Bảng 3. Xác suất dự báo để Z=1 ở 5 tháng đầu năm 2017

Quan sát	FPT	Tần suất xuất hiện của Z =1 thực tế (%)	$P_r(Z=1/PC_1, PC_2, PC_3, PC_4, PC_5)$
Tháng 12/2016	36.4	37/60 =0.617	0.56208
Tháng 01/2017	38.28	38/61= 0.623	0.70139
Tháng 02/2017	38.59	39/62= 0.629	0.83404
Tháng 03/2017	39.55	40/63=0.635	0.88181
Tháng 04/2017	39.58	41/64=0.641	0.87105
Tháng 05/2017	41.29	42/65=0.646	0.88897

Bảng này cho thấy xác suất để giá cổ phiếu FPT nằm trong xu hướng giá tháng sau cao hơn tháng trước từ tháng 01/2017 đến tháng 05/2017 luôn khá cao, ở tháng 1/2017, xác suất trên 70%, các tháng còn thấp nhất là trên 83%, còn lại đều gần 90%. Điều đó cho thấy nếu lấy giá trị "cắt" là 0.7 (hay 70%) thì cả 5 quan sát ứng với 5 tháng đầu năm 2017 đều được phân vào lớp nhãn Z=1 (tức lớp có giá cổ phiếu FPT tháng sau cao hơn tháng trước). Số liệu thực tế về giá cổ phiếu FPT cũng cho thấy ở 5 tháng này, giá cổ phiếu FPT cũng luôn nằm trong xu hướng tăng (hay Z=1). Cần lưu ý là; như đã nói ở trên, ta không thể lấy tần suất các quan sát có nhãn Z=1 (mặc dù cũng khá cáo, trên 60%) để so sánh với xác suất để Z=1 tại các quan sát đó vì xác suất còn phụ thuộc vào giá trị cụ thể của các biến gốc PC_i trong mỗi quan sát.

Chọn giá trị "cắt" là 0.5, sử dụng mô hình dự báo được xây dựng trên tập dữ liệu của 60 quan sát đầu tiên để tính xác suất dự báo phục vụ cho việc phân lớp của tất cả 65 quan sát từ tháng 1/2012 đến tháng 5/2017, ta nhận được kết quả trong Bảng 4.

Bảng 4: Độ chính xác phân lớp theo giá trị "cắt" = 0.5

	Z=0	Z=1	Tổng
P(Z=1)<=0.5	13	7	20
P(Z=1)>0.5	10	35	45
Tổng	23	42	65
Chính xác	13	35	48
% Chính xác	56.52	83.33	73.85
% Không chinh xác	43.48	16.67	26.15

Từ bảng này có thể thấy rằng, với giá trị "cắt" không thiên vị thì phân lớp với nhãn là 1 cho độ chính xác cao hơn với nhãn là 0. Khi giá trị cắt càng nhỏ hơn 0.5 thì độ chinh xác phân lớp nhãn là 1 càng cao và ngược lại độ chính xác phân lớp nhãn 0 càng thấp. Trái lại độ chính xác phân lớp nhãn là 1 sẽ giảm dần, nhãn là 0 sẽ tăng dần khi giá trị cắt là gần đến 1.

d. Dự báo, mô phỏng và phân tích rủi ro dự báo

d1. Dự báo mức độ rủi ro về kết quả dự báo theo xu thế

Từ phương trình (11), để dự báo được mức độ rủi ro (hay xác suất) về kết quả dự báo giá cổ phiếu FPT theo xu thế, ta chỉ cần dự báo các thành phần chính PC_4, PC_5 được ước lượng theo phương trình (10) như sau:

$$\mathbf{d(PC_4)} = -0.194*PC_4(-1) + 0.359*d(PC_4(-1)) + 0.293*d(PC_4(-4)) + 0.223*d(PC_4(-11)) \quad (12)$$

Std: (0.065) (0.131) (0.128) (0.101)

R^2: 0.28; DW: 1.93; SMPL: 54, sau khi đã điều chỉnh bởi trễ.

$$\mathbf{d(PC5)} = -0.33*PC5(-1) + 0.45*D(PC5(-1)) + 0.27*D(PC5(-3)) + 0.27*D(PC5(-5)) + 0.31*D(PC5(-10))$$

Std: (0.079) (0.125) (0.134) (0.117) (0.098)

$$- 0.44*D2015M07 + 0.49*D2016M08 \quad (13)$$

(0.181) (0.180)

R^2: 0.47; DW: 1.95; SMPL: 54, sau khi đã điều chỉnh bởi trễ.

Với 51 biến gốc hoàn toàn tương tự như trong bài báo này, bài báo [13] đã cung cấp kết quả dự báo giá cổ phiếu FPT cũng như sai số trung bình của nó ở 3 tháng 6, 7, 8 của năm 2017 và thực hiện dự báo khả năng xảy ra rủi ro về kết quả dự báo theo các phương trình (12) và (13) ta nhận được Bảng 4 ở dưới:

Bảng 5. Dự báo khả năng xảy ra rủi ro về kết quả dự báo theo xu thế

	Tháng 5/2017	**Tháng 6/2017(f)**	**Tháng 7/2017(f)**	**Tháng 8/2017(f)**	**Giải thích/Nguồn**
PC_4	1.635	1.460	1.177	1.101	Biến gốc (f)
PC_5	0.499	0.614	1.090	0.721	Biến gốc (f)
FPT	**41.29**	**41.451**	**41.147**	**41.068**	[13]
Sai số trung bình		+/- 1.582	+/- 1.565	+/- 1.561	
Sai số trung bình (%)		+/- 3.82	+/- 3.80	+/- 3.80	[13]
$\mathbf{P_r}$ (Z=1/ $PC_1,\ldots,PC_5$)	**0.889**	**0.888**	**0.908**	**0.875**	(f): Dự báo

Từ Bảng này ta có thể nhận được: giá cổ phiếu của công ty FPT trong các tháng 6, 7, 8 bây giờ sẽ là: [41.290, 41.451+1.582]. [41.451, 41.147+1.565] và [41.147, 41.068 + 1.561] với xác suất xảy ra tương ứng là 0.888, 0.908 và

0.875 chứ không phải là các khoảng [41.451-1.582, 41.451+1.582]. [41.147-1.565, 41.147+1.565] và [41.068 - 1.561, 41.068 + 1.561] như theo Bài báo [13]. Khả năng giá cổ phiếu FPT không nằm trong 3 khoảng đã nêu ở trên ứng với các tháng 6, 7, 8 năm 2017 chỉ vào khoảng 10%. Như vậy việc tính xác suất của rủi ro dự báo theo xu thế đã hỗ trợ làm thu hẹp khoảng dự báo giá cổ phiếu.

d2. Mô phỏng và phân tích rủi ro dự báo

Kết quả dự báo giá cổ phiếu FPT trong Bảng 4 sẽ xảy ra với xác suất khoảng 0.9 (hay 90%) nếu 3 tháng 6,7,8 năm 2017 không có những biến động đặc biệt về kinh tế, chính trị, xã hội trong nước và ngoài nước, trong đó nhất là những thay đổi về chính sách điều hành kinh tế của Chính phủ. Ngược lại rủi ro dự báo có thể xảy ra. Nhằm phát hiện và lường trước tác động của những rủi ro này, cần thực hiện mô phỏng dự báo. Có vô vàn kịch bản mô phỏng dự báo giá cổ phiếu FPT. Trong bài báo này chỉ thực hiện một mô phỏng như là một minh họa điển hình.

Phân tích 28 biến gốc có tác động đến sự thay đổi của FPT, ta thấy có những biến như liên quan trực tiếp đến chính sách điều hành của Chính phủ như dư nợ tín dụng: DUNO, giá thế giới về dầu thô: OIL và chỉ số chứng khoán S$P500: SP500 là những biến mà sự thay đổi của chúng trong rất nhiều trường hợp không theo quy luật của hiện tại và quá khứ.

Trong kịch bản mô phỏng này, ta giả sử rằng ở 3 tháng 6, 7, 8 năm 2017 dư nợ tín dụng sẽ giảm 1% mỗi tháng, giá dầu thô tăng 2%/ tháng và chỉ số chứng khoán SP500 tăng 2%/tháng, các biến khác diễn ra một cách bình thường. Thế thì điều gì sẽ xảy ra về xu thế tăng/giảm giá của cổ phiếu FPT ?

Thực hiện dự báo 25 biến gốc khác trừ 3 biến đã nêu bằng phương pháp hồi quy nhiều biến bằng mô hình dự báo được ước lượng theo phương trình (10) ở 3 tháng 6, 7, 8 năm 2017 và tính các thành phần chính ở 3 tháng đã nêu, sau đó tính xác suất để biến đích FPT có xu hướng tăng theo phương trình (11), ta nhận được kết quả trong Bảng 6.

Bảng 6: Xác suất để giá FPT có xu hướng tăng của KBMP và KB cơ sở

	Tháng 6/2017	**Tháng 7/2017**	**Tháng 8/ 2017**	**Ghi chú/Nguồn**
PC1	7.352	7.754	7.941	Kịch bản mô phỏng
PC2	-0.625	-0.695	-0.484	KBMP
PC3	0.26	0.274	-0.262	KBMP
PC4	1.319	1.392	1.191	KBMP
PC5	0.483	0.873	0.663	KBMP
$\mathbf{P_r}$ (Z=1/ $PC_1,\ldots,PC_5$)	0.888	0.908	0.875	Kịch bản cơ sở
$\mathbf{P_r}$ (Z=1/ $PC_1,\ldots,PC_5$)	0.88983	0.91113	0.88132	KBMP
KBMP so với KBCS	**0.00183**	**0.00313**	**0.00632**	

Bảng 6 cho thấy khi Chính phủ hạn chế dư nợ tín dụng, và khi giá thế giới về dầu thô và chỉ số giá chứng khoán S&P500 có xu hướng tăng thì xác suất để giá cổ phiếu FPT cũng có xu hướng tăng giá. Mức độ tăng của giá cổ phiếu FPT phụ thuộc vào mức độ giảm và tăng cụ thể tương ứng của dư nợ tín dụng, giá dầu thô và chỉ số chứng khoán S&P500 như được chỉ ra.

3.3.2. Xây dựng mô hình mô phỏng và phân tích rủi ro khác biệt về kết quả dự báo

a. Gán nhãn cho tập dữ liệu gốc

Việc phân chia 60 quan sát đầu tiên của tập dữ liệu đầu vào thành 6 lớp với 6 nhãn từ 1 đến 6 được dựa vào giá trị của cổ phiếu FPT và theo công thức (2) như sau:

$$z_i = \begin{cases} 1,\ nếu\ 100 * \dfrac{(FPT_i - FPT_{i-1})}{FPT_{i-1}} < -3\% \\ 2, nếu -3\% \le 100 * \dfrac{(FPT_i - FPT_{i-1})}{FPT_{i-1}} < -1\% \\ 3, nếu\ -1\% \le 100 * \dfrac{(FPT_i - FPT_{i-1})}{FPT_{i-1}} < 1\% \\ 4, nếu\ 1\% \le 100 * \dfrac{(FPT_i - FPT_{i-1})}{FPT_{i-1}} < 3\% \\ 5,\ nếu\ 3\% \le 100 * \dfrac{(FPT_i - FPT_{i-1})}{FPT_{i-1}} < 5\% \\ 6,\ nếu\ 5\% \le 100 * \dfrac{(FPT_i - FPT_{i-1})}{FPT_{i-1}} \end{cases}$$

ở đây Z = $\{z_i\}$ là biến nhãn lớp, i là quan sát thứ i, i = 1, …, 60.

b. Xây dựng mô hình dự báo

Mô hình dự báo xác suất xảy ra của mỗi lớp được ước lượng theo mô hình hồi quy có thứ tự (8), (9) với các biến PC_i (i=1, 2, …,5) là các biến gốc thay thế cho 51 biến gốc ban đầu ở 60 quan sát đầu tiên như sau:

$$P_r(Z=1/PC_1,..,PC_5) = \frac{1}{1+e^{-(1.90+0.29*PC3+\ 0.67*PC4+0.47*PC5)}}$$

$$P_r(Z=2/PC_1,..,PC_5) = \frac{1}{1+e^{-(0.99+0.29*PC3+\ 0.67*PC4+0.47*PC5)}} - \frac{1}{1+e^{-(1.90+0.29*PC3+\ 0.67*PC4+0.47*PC5)}}$$

$$P_r(Z=3/PC_1\ldots,PC_5) = \frac{1}{1+e^{-(0.26+0.29*PC3+\ 0.67*PC4+0.47*PC5)}} - \frac{1}{1+e^{-(0.99+0.29*PC3+\ 0.67*PC4+0.47*PC5)}}$$

$$P_r(Z=4/PC_1,..,PC_5) = \frac{1}{1+e^{-(0.40+0.29*PC3+\ 0.67*PC4+0.47*PC5)}} - \frac{1}{1+e^{-(0.26+0.29*PC3+\ 0.67*PC4+0.47*PC5)}}$$

$$P_r(Z=5/PC_1\ldots,PC_5) = \frac{1}{1+e^{-(1.10+0.29*PC3+\ 0.67*PC4+0.47*PC5)}} - \frac{1}{1+e^{-(0.40+0.29*PC3+\ 0.67*PC4+0.47*PC5)}}$$

$$P_r(Z=6/PC_1,\ldots,PC_5) = 1 - \frac{1}{1+e^{-(1.10+0.29*P\ C3+\ 0.67*PC4+0.47*PC5)}} \quad (14)$$

c. Đánh giá khả năng dự báo của mô hình

Việc đánh giá khả năng dự báo của mô hình hồi quy có thứ tự là tương tự như mô hình hồi quy loogit, nên mục này sẽ được bỏ qua.

d. Dự báo, mô phỏng và phân tích rủi ro

Tương tự như đối với mô hình hồi quy logit, trong mục này sẽ giới thiệu 2 kịch bản dự báo. Kịch bản cơ sở (KBCS) và Kịch bản mô phỏng (KBMP) giả sử cũng được giả định tương tự như trường hợp mô phỏng và phân tích rủi ro dự báo xu thế ở trên. Khi đó xác suất dự báo để phân lớp cho 3 tháng 6, 7, 8 năm 2017 theo bộ dữ liệu đầu vào của hai kịch bản này được cho trong Bảng 7.

Bảng 7: Hai kịch bản xác suất dự báo thuộc về mỗi lớp theo dữ liệu đầu vào 3 tháng 6,7,8/2017

Dự báo giá cổ phiếu FPT	**Tháng 6/2017**	**Tháng 7/2017**	**Tháng 8/2017**	**Ghi chú**
Thấp nhất	41.290	41.451	41.147	Do kết hợp kết quả dự báo trong [13] và xác suất dự báo theo xu thế
Cao nhất	43.033	42.712	42.629	Kết quả dự báo trong [13]
$P_r(Z=1/PC_1,..,PC_5)$	0.0507	0.0418	0.0598	**KBCS**
$P_r(Z=2/PC_1,..,PC_5)$	0.0642	0.0541	0.0741	KBCS
$P_r(Z=3/PC_1,..,PC_5)$	0.1157	0.1007	0.1291	KBCS
$P_r(Z=4/PC_1,..,PC_5)$	0.1391	0.1272	0.1482	KBCS
$P_r(Z=5/PC_1,..,PC_5)$	0.1697	0.1650	0.1712	KBCS
$P_r(Z=6/PC_1,..,PC_5)$	0.4606	0.5112	0.4175	KBCS
~~**KBMP so với KBCS**~~				
~~$P_r(Z=1/PC_1,..,PC_5)$~~	~~-0.0006~~	~~-0.0011~~	~~-0.0024~~	KBMP-KBCS
$P_r(Z=2/PC_1,..,PC_5)$	-0.0006	-0.0013	-0.0025	KBMP-KBCS
$P_r(Z=3/PC_1,..,PC_5)$	-0.0009	-0.0019	-0.0033	KBMP-KBCS
$P_r(Z=4/PC_1,..,PC_5)$	-0.0007	-0.0017	-0.0021	KBMP-KBCS
$P_r(Z=5/PC_1,..,PC_5)$	-0.0002	-0.0009	-0.0001	KBMP-KBCS
$P_r(Z=6/PC_1,..,PC_5)$	0.0029	0.0068	0.0104	KBMP-KBCS

FPT thuộc vào lớp giá tháng sau tăng cao hơn 5% so với tháng trước là tăng, trong khi xác suất để giá cổ phiếu FPT tăng hoặc giảm theo các mức độ khác đều giảm. Kết quả dự báo giá cổ phiếu FPT trong [13], kết hợp với xác suất dự báo giá cổ phiếu FPT tăng/giảm theo xu thế ở mục 4.3.1 đã làm thu hẹp khoảng dự báo giá cổ phiếu FPT, và việc tính được xác suất dự báo thuộc về mỗi một trong 6 lớp của tập dữ liệu đầu vào ở 3 tháng 6,7,8 năm 2017 theo kịch bản cơ sở và kịch bản mô phỏng như được trình bầy trong Bảng này sẽ góp phần làm thu hẹp khoảng dự báo hơn nữa, đồng thời cũng cho biết chi tiết hơn về mức độ rủi ro của kết quả dự báo.

IV. KẾT LUẬN

Bài báo này đã đề xuất khung lý thuyết để dự báo, mô phỏng và phân tích rủi ro dự báo trong ngữ cảnh tập dữ liệu của các biến gốc đầu vào để phục vụ dự báo có số chiều cao, hơn nữa trong tập biến gốc đầu vào có thể có những biến không có ý nghĩa tác động đến sự thay đổi của biến đích và/hoặc không cần thiết phải được quan tâm khi dự báo biến đích (biến dư thừa). Việc giảm chiều dữ liệu trong bài báo này là hoàn toàn tương tự như trong [13] đó là sử dụng đồng thời kỹ thuật lựa chọn thuộc tính bởi thuật toán được đề xuất trong [13] và kỹ thuật học thuộc tính PCA nhằm đảm bảo không chỉ giảm chiều, loại bỏ biến gốc kém ý nghĩa và dư thừa, mà còn bảo toàn nhiều nhất như có thể quan hệ giữa biến gốc và biến đích. Họ mô hình hồi quy lôgit được đề xuất trong bài báo này có thể được xem là giải pháp phù hợp, hiệu quả nhất cho đến thời điểm này trong việc xây dựng mô hình dự báo, mô phỏng và phân tích rủi ro dự báo biến đích trong lĩnh vực kinh tế - xã hội.

Bài báo đã thực hành ứng dụng phương pháp được đề xuất trong việc xây dựng mô hình dự báo xác suất kết quả dự báo của giá cổ phiếu FPT thuộc về mỗi lớp theo tập dữ liệu đầu vào của các biến gốc. Việc thực hành mô phỏng xác suất dự báo theo các kịch bản khác nhau của tập dữ liệu gốc đầu vào sẽ giúp nhà đầu tư thu hẹp khoảng dự báo của giá cổ phiếu FPT và đồng thời cũng nhận biết được mức độ rủi ro có thể xảy ra. Những kết quả phân tích ấy thực sự có ích đối với việc ra quyết định của nhà đầu tư cổ phiếu. Kết quả nghiên cứu của bài báo này và bài báo [13] có thể được xem là một nghiên cứu khá hoàn chỉnh về dự báo giá và phân tích rủi ro dự báo về giá, một trong những bài toán được xem là khó nhất trong lĩnh vực dự báo kinh tế - xã hội.

TÀI LIỆU THAM KHẢO

[1]. Allison, P. D. (2014). Measures of Fit for Logistic Regression, Statistical Horizons LLC and the University of Pennsylvania, working paper: 1485.

[2]. Aven, T. et all (2015). Risk Analysis Foundations, Society of Risk Analysis.

[3]. Barros, R. C., Basgalupp, M. P., Carvalho, A. C. P. L. F., Freitas, Alex A. (2011). A Survey of Evolutionary Algorithms for Decision-Tree Induction. IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews, vol. 42, n. 3, p. 291-312, May 2012.

[4]. Berg, H. P. (2010). Risk management: procedures, methods and experiences, RT&A # 2(17), (Vol.1) 2010, June.

[5]. Hosmer, D.W., Lemeshow, S. (2000). Applied Logistic Regression, New York, USA: John Wiley and Sons.

[6]. McFarland, H. R. and Richards D. St. P. (2002). Exact Misclassification Probabilities for Plug-In Normal Quadratic Discriminant Functions, II. The Heterogeneous Case, Journal of Multivariate Analysis, 2002, vol. 82, issue 2, pages 299-330.

[7]. Piesse, J. and Lin, L. (2004). Financial risk assessment in takeovers: The effect on bidder firm shareholders' wealth, Research Paper 028, The Management Centre, King's College, University of London.

[8]. Pohar, M., Blas, M., Turk, S. (2004). Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study, Metodološki zvezki, Vol. 1, No. 1, 2004, 143-161.

[9]. Rausand, M. (2011). Risk Assessment: Theory, Methods, and Applications, Society of Risk Analysis.

[10]. Simar W. L. (2007). Applied Multivariate Statistical Analysis, Springer Berlin Heidelberg, pp. 289–303.

[11]. Stoica, C. R. and Constantin, B. (2012). The assessment of risks that threaten a project, Economy transdisciplinarity cognition, Vol.15, Issue 2/2012.

[12]. Tixier, J., Dusserre, G., Salvi, O., Gaston D. (2014). Review of 62 risk analysis methodologies of industrial plants, HAL Id: ineris-00961858, https://hal-ineris.ccsd.cnrs.fr/ineris-00961858.

[13]. Thành D. V. (2017), Dự báo giá cổ phiếu trong ngữ cảnh dữ liệu số chiều cao, Bài gửi hội nghị FAIR X, Đà Năng 17-18/8/2017.

[14]. Yu, H. F. ; Huang, F. L., Lin, C. J. (2011). Dual coordinate descent methods for logistic regression and maximum entropy models, Machine Learning, **85**: 41–75. doi:10.1007/s10994-010-5221-8.

[15]. Zhang, Y. (2009). Remote-sensing Image Classification Based on an Improved Probabilistic Neural Network, Sensors, **9** (9): 7516–7539, doi:10.3390/s90907516.

SIMULATION AND ANALYSIS OF FORECAST RISKS ON HIGH DIMENSIONAL DATA SET

Thanh Do Van[1] and Hieu Do Duc[2]

1: Departement of Information Technology, Nguyen Tat Thanh University, dvthanh@ntt.edu.vn

2: Departement of Information and Communication Technology, University of Science and Technology of Hanoi, Vietnamese Academy of Science and Technology, vincentdo2310@gmail.com

***ABSTRACT**: Using quantitative models for forecast means that the future is assumed to close to the present and the past. But the reality is not so therefore the forecasts are always difficult and there are many cases the forecast results are far different, even are opposed to the fact even though the forecast model is carefully diagnosised and tested, and assessed as goodness - of - fit. This phenomenon is called forecast risks. The forecast risk is often measured by the probability of its occurrence.*

The purpose of this paper is to propose a theoretical framework for simulation and analysis of forecast risks in the context that the factors affecting the problem (or the target variable) needed to be forecasted are very large. The paper uses data dimension reduction techniques to transfer high-dimensional data sets (number of original variables and/or number of observations is huge) into low-dimensional data sets so that the relationships between the target variable and the original variables change as little as possible and basically the low-dimensional data sets reflect rather fully information in the high-dimensional data sets. At the same time the paper uses also the logistic regression model or the ordered regression model to build models for caculating forecast probabilities under trends or under different levels of forecast results. The paper applies the proposed methodology in the economic-financial field.

***Keyword**: dimentionality reduction, high-dimensional data, simulation of forecast risk, risk analysis, logistic and odered regression.*