DỰ BÁO CHÍNH XÁC DỊCH CÚM TOÀN CẦU THÔNG QUA MÔ HÌNH THỐNG KÊ SỬ DỤNG DỮ LIỆU LỚN CỦA GOOGLE

Shihao Yang, Mauricio Santillana, và Samuel Kou, Đại học Harvard, Mỹ

(tiếp theo)

Thảo luận

Khả năng dự báo của mô hình ARGO: Từ các kết quả trình bày cho thấy khả năng dự báo của mô hình ARGO có độ chính xác cao so với tất cả các mô hình thử nghiệm khác. Kết quả dự báo sẽ còn chính xác hơn nếu nhóm nghiên cứu được tiếp cận với các biến tham số đầu vào của Google sử dụng tính toán phục vụ dự báo, vì hiện tại nhóm đang thực nghiệm dự báo với các biến đầu vào dựa trên dữ liệu chất lượng thấp của Google.

Sự kết hợp giữa thông tin tìm kiếm dịch cúm theo mùa với quyền số linh hoạt là một yếu tố quan trọng trong tính chính xác nâng cao của mô hình ARGO. Vì thông tin về mức độ hoạt động dịch cúm tuần trước thường có một tác động đáng kể vào mức độ hiện tại và những thông tin cách đây nửa năm hay 1 năm có thể cung cấp thêm thông tin, như thể hiện trong Hình 1, phản ánh sự tương quan mạnh mẽ thời gian, là hệ số tương quan dương có nghĩa rằng các thông tin về dịch cúm có mối liên quan với nhau. Bên cạnh đó, việc tính toán tích hợp các thông tin mô hình chuỗi thời gian đã đưa ra một mô hình đường cong liên tục, giúp ngăn ngừa được những điểm phát sinh đột biến không mong muốn. Việc thực hiện chỉ là thêm các điều khoản tham chiếu của mô hình chuỗi thời gian vào mô hình GFT ban đầu để trở thành một mô hình tối ưu (mô hình ARGO). Để thực hiện được điều này nhóm nghiên cứu coi toàn bộ mô hình GFT ban đầu là một biến tham số độc lập và không cho phép thay đổi thông tin của biến này trong mô hình chuỗi thời gian ở các mức độ khác nhau khi truy vấn. Như vậy, khi thông tin của mô hình chuỗi thời gian được kết hợp thêm điều khoản mới thì nhiều điều khoản đang áp dụng đối với mô hình GFT ban đầu sẽ không còn giúp ích cung cấp thêm thông tin. Tuy nhiên, trong thực tế thông tin của mô hình chuỗi thời gian chứa các thuật ngữ truy vấn đơn lẻ vẫn còn có thể giúp ích cung cấp các thông tin có giá trị về dịch cúm. Ví dụ, trong số 100 thuật ngữ truy vấn của dữ liệu có tương quan với Google được lựa chọn, thì mô hình ARGO lựa chọn 14 điều khoản kết hợp, còn các mô hình Santillana et al và mô hình GFT lựa chọn tương ứng là 38 và 45 điều khoản kết hợp. Do vậy kết quả tìm kiếm của mô hình ARGO sẽ được mở rộng phạm vi hơn. Ngoài ra, sự kết hợp độ trơn (làm cho mô hình dự báo được mịn hơn) và độ thưa (làm giảm đi những vùng có ít điểm dữ liệu điểm quan sát trong mô hình) đã giúp cho mô hình ARGO giảm đáng kể các lỗi tính toán so với các mô hình khác, như ở Bảng 1 và Bảng 2 cho thấy mô hình ARGO đã cải thiện hiệu suất khi đánh giá số liệu trong khoảng thời gian nghiên cứu và gấp đôi hiệu quả của mô hình GFT + AR(3).

Thông qua mô hình ARGO chúng ta thấy được cách bổ sung hỗ trợ lẫn nhau giữa thông

tin các nguồn dữ liệu tìm kiếm từ Google Trends và nguồn dữ liệu có tương quan với Google trong mô hình chuỗi thời gian (Hình1). Đối với mô hình chuỗi thời gian thường có xu hướng thay đổi chậm để đáp ứng với những thay đổi đột ngột khi quan sát mức độ hoạt động dịch cúm của CDC. Điều này thấy rõ thông qua "độ trễ" ở mô hình chuỗi thời gian linh hoạt AR(3). Mặc dù, mô hình AR(3) có hệ số tương quan tốt. Đối với mô hình ARGO thì ngược lại, đã xử lý rất hiệu quả với những trường hợp phát hiện sự thay đổi hoạt động đột ngột về dịch cúm, và nó cũng rất nhạy cảm với những hoạt động hành vi tăng đột biến của người dân tìm kiếm về thông tin dịch cúm.

Để hiểu rõ hơn mối quan hệ các biến tham số có ảnh hưởng đến độ chính xác của dự báo dịch cúm trong mô hình ARGO, nhóm nghiên cứu đã tính toán lượng tăng/giảm giữa các hệ số tương quan và so sánh với mô hình dự báo GFT. Lượng tăng/giảm của hệ số tương quan giữa hai mô hình theo chuỗi thời gian a_t và b_t được định nghĩa là Corr($a_t - a_{t-1}$, $b_t - b_{t-1}$). Trong Bảng 1, Mô hình ARGO (Corr(ARGO) = 0.758) có giá trị tương tự mô hình GFT và mô hình Santillana et al có nghĩa là mô hình này cũng có khả năng như mô hình GFT trong việc nắm bắt được những mức độ thay đổi trong hoạt động của dịch cúm, và nhanh hơn mô hình AR(3).

Thông tin chuỗi thời gian (mùa dịch) có xu hướng làm cho dự báo của mô hình ARGO thay đổi trong quá khứ. Điều này thấy rõ khi bắt đầu bùng phát mùa dịch cúm H1N1 năm 2009, khi đó mô hình ARGO đưa ra dự báo kết quả thấp (Hình 1 đường màu đỏ thấp nhất). Mô hình ARGO đã tự động điều chỉnh sửa lỗi hiệu quả bằng cách chuyển một phần quyền số các giá trị tìm kiếm từ miền mô hình theo chuỗi thời gian (dữ liệu báo cáo trong quá khứ) sang miền dữ liệu của mô hình truy vấn tìm kiếm của Google ở các tuần sau đó. Ngược lại, ở mùa dịch cúm 2012-2013 (10/2012 - 04/2013), các mô hình ARGO, GFT, và Santillana et al đã vượt quá giới hạn có thể theo dõi dịch cúm (mất đỉnh) (Hình 1). Điều này có thể do một sự thay đổi đột ngột nào đó chưa từng có trong quá khứ về hoạt động tìm kiếm thông tin dịch cúm. Nhưng mô hình ARGO đã xử lý nhanh bằng cách tự động điều chỉnh quyền số đối với toàn bộ các điều khoản truy vấn và tìm kiếm thông tin của Google trong toàn bộ chuỗi thời gian theo dõi, nên sự việc mất tích giới hạn theo dõi dịch cúm chỉ xảy ra trong 1 tuần. Trái lại, mô hình Santillana et al diễn ra trong 2 tuần và mô hình GFT diễn ra khoảng 4 tuần. Điều quan trọng, chúng ta thấy các cơ quan y tế ở Hoa kỳ đã sử dụng dữ liệu báo cáo dịch cúm của CDC như là thông tin tiêu chuẩn đảm bảo cho các hoạt động dự báo mức độ tình trạng dịch cúm, mà các dữ liệu có mối tương quan với Google hoặc Google Trends thì được coi như là các biến tham số độc lập. Qua đó, chúng ta có thể thấy mô hình ARGO có thể tự xử lý điều chỉnh nhanh để thích nghi phù hợp tình hình thực tế của dịch cúm với bất kỳ các tiêu chuẩn khác nhau của các biến tham số, có thể đó là biến tham số độc lập.

Hạn chế và các bước tiếp theo: Mặc dù mô hình ARGO đã thể hiện khả năng vượt trội hơn so với các mô hình khác, nhưng không có nghĩa đây là một mô hình hoàn hảo, vì cách thức tính toán của mô hình dựa trên các dữ liệu hành vi tìm kiếm thông tin dịch cúm của của người dân. Nếu có thay đổi đột ngột về cấu trúc của các công cụ tìm kiếm hoặc phương thức truy vấn tìm kiếm thông tin thì sẽ ảnh hưởng đến kết quả và độ chính xác của mô hình dự báo. Nhóm nghiên cứu hy vọng rằng mô hình ARGO sẽ nhanh chóng tự điều chỉnh được nếu có sự thay đổi như vậy xảy ra trong tương lai. Ngoài ra, với bất kỳ mô hình dự báo thì chất lượng hoạt động tốt trong quá khứ và hiện tại sẽ không đảm bảo hoạt động

tốt trong tương lai. Do vậy, bài viết này nhóm nghiên cứu đã cố định bảng các thuật ngữ truy vấn từ trước năm 2010 và so sánh với bảng kết quả các thuật ngữ từ năm 2010 trở đi với những điều khoản truy vấn tương tự trong mô hình dự báo.

Trong tương lai, các ứng dụng của mô hình ARGO có thể tiếp tục được cập nhật thường xuyên hơn và dễ sử dụng hơn để nắm bắt được mức độ hoạt động các bệnh dịch hoặc sự kiện xã hội được theo dõi thông qua hình thức truy vấn tìm kiếm thông tin trực tuyến với bất kỳ quy mô không gian và thời gian nào. Bên cạnh đó, sẽ cải thiện hơn nữa trong hoạt động dự báo dịch cúm bằng cách kết hợp nhiều yếu tố dự báo từ các nguồn dữ liệu khác nhau.

Ngay sau khi nhóm nghiên cứu về GFT gửi báo cáo ban đầu tháng 05/2015, trong đó đưa ra đề xuất về một mô hình chuỗi thời gian mới theo dõi dịch cúm dựa trên mô hình GFT. Thì Google đã thông báo cho phép các nhà khoa học nghiên cứu về GFT được tiếp cận với dữ liệu thô của họ. Đề xuất mới này đã đóng góp kịp thời và có ích trong việc cung cấp một phương pháp minh bạch cho việc theo dõi dịch bệnh trong tương lai.

Dữ liệu và phương pháp

Dữ liệu của Google

Để tránh thông tin truy vấn ngoài khoảng thời gian nghiên cứu trước năm 2009, thì những dữ liệu này đã được đưa ra ngoài mẫu nghiên cứu. Cách tiếp cận như vậy là phù hợp với nội dung nghiên cứu mô hình GFT. Ngay sau khi phát sinh đại dịch H1N1 năm 2009, nhóm nghiên cứu đã thu thập được bộ dữ liệu có mối tương quan với Google tốt nhất của CDC cho hai giai đoạn khác nhau (*tại www.google.com/trends/correlate*) thông qua hình thức truy vấn tìm kiếm thông tin trực tuyến. Giai đoạn đầu (giai đoạn trước khi xảy ra H1N1), nhóm nghiên cứu chèn dữ liệu dịch cúm từ các báo cáo của CDC trong giai đoạn 1/2004 đến 28/3/2009 (dữ liệu trước khi xảy ra đại dịch cúm) vào cùng với bộ dữ liệu mà nhóm thu được, và sử dụng các thuật ngữ tìm kiếm có tương quan tốt nhất được coi như là các biến tham số độc lập để giúp cho nhóm nghiên cứu dự đoán dịch cúm trong khoảng thời gian nghiên cứu dịch cúm 04/4/2009 đến 22/5/2010. Trong giai đoạn thứ hai (giai đoạn xảy ra dịch cúm H1N1), nhóm nghiên cứu cũng chèn dữ liệu dịch cúm từ các báo cáo của CDC từ 01/2004 đến 22/5/2010 nhưng với các điều khoản tham chiếu tìm kiếm dịch cúm áp dụng cho toàn bộ gói dữ liệu mà nhóm thu được. Các thuật ngữ tìm kiếm cuối cùng đã được sử dụng như là các biến độc lập cho tất cả các dự đoán trong quá trình nghiên cứu có hoặc không kèm thêm điều kiện. Ví dụ thuật ngữ flu.fever (trong cụm từ tìm kiếm thì cụm từ fever (cơn sốt) được coi là biến tham số độc lập, bên cạnh đó, có thêm điều kiện giả định là flu (cúm); Nhưng thuật ngữ fevers cũng có thể được tìm kiếm không kèm theo điều kiện nào). Đối với giai đoạn trước khi xảy ra dịch cúm H1N1, các giả thiết có trong dữ liệu có tương quan với Google bao gồm các điều khoản giả định 7 (điều kiện chỉ xảy ra trong quá trình nghiên cứu hoặc không thể xảy ra được trong thực tế). Tuy nhiên, những giả định này không được mô hình ARGO lựa chọn, nghĩa là mô hình ARGO sẽ lựa chọn những dữ liệu giả định này với quyền số bằng không. Qua đó nó đã thể hiện được khả năng phân loại thông tin mạnh mẽ của mô hình. Đối với khoảng thời gian sau dịch cúm H1N1, các thuật ngữ truy vấn cập nhật từ dữ liệu có tương quan với Google với các điều khoản chủ yếu liên quan đến cúm, có nghĩa các giả định nhóm nghiên cứu đưa vào đã được "lọc ra" khỏi dữ liệu của

[7] *Lazer D, Kennedy R, King G, Vespignani A (2014) Big data. The parable of Google Flu:*
Traps in big data analysis. Science 343(6176):1203–1205.

mùa dịch cúm năm sau. Trong khoảng thời gian của 28/03/2015 đến ngày gửi đi báo cáo này, nhóm nghiên cứu đã tổng hợp được tần suất tìm kiếm các thuật ngữ truy vấn từ Google Trends (*tại www.google.com/trends; cập nhật 11/7/2015*), Vì lý do ban đầu, nhóm nghiên cứu chỉ thu thập được dữ liệu có tương quan với Google đến ngày 28/3/2015.

Nhóm nghiên cứu thu thập các dữ liệu có tương quan với Google dựa trên các tiêu chuẩn về khối lượng tìm kiếm thông tin của mỗi truy vấn phải có (Sai số trung bình Mean = 0 và độ lệch chuẩn SD = 1) và chỉ xem xét trong giai đoạn từ 01/2004 đến 03/2015.

Trong quá trình chuyển đổi nguồn dữ liệu để dự báo dịch cúm, nhóm nghiên cứu nhận thấy một vấn đề làm thế nào hai nguồn dữ liệu này có thể phù hợp với nhau. Để giải quyết nhóm nghiên cứu đã thực hiện chuyển đổi bộ dữ liệu có tương quan với Google thu được thành hàm tuyến tính với quy mô [0,100] tương tự trong bộ dữ liệu mà nhóm nghiên cứu đang phân tích, vì nguồn dữ liệu này sẵn có. Sau đó chuyển sang nguồn dữ liệu Google Trends. Điều này được thể hiện rõ trong Hình 1 bởi màu nền khác nhau của nguồn dữ liệu sử dụng cho dự báo. Nhóm nghiên cứu sử dụng dữ liệu mới nhất của GFT (phiên bản 4, 05/2014). Và dữ liệu mới nhất về dịch cúm của GFT có tại www.google.org/ flutrends /(cập nhật 11/7/2015).

Dữ liệu dịch cúm của CDC

Nhóm nghiên cứu sử dụng các phiên bản dữ liệu dự báo dịch cúm có quyền số của CDC (tại gis.cdc.gov/grasp/fluview/fluportaldashboard html; cập nhật 11/7/2015). Các phiên bản dự báo hàng tuần ILI của CDC có sẵn tại trang web của CDC có tất cả thông tin mùa dịch cúm (từ tuần 40 của năm trước cho tới tuần 20 của năm tiếp theo). Ví dụ, báo cáo dự báo tình hình dịch cúm vào tuần thứ 50 của mùa dịch 2012-2013 có sẵn tại www.cdc.gov/flu/ weekly/ weeklyarchives2012-2013 /data /senAllregt50.htm; và báo cáo sửa đổi của tuần 50 này thì có vào tuần thứ 9 của mùa dịch cúm 2014-2015 (www.cdc.gov/flu/weekly/ weeklyarchives 2014-2015 / data / senAllregt09.html)

Xây dựng mô hình ARGO

Như đã đề cập ở phần giới thiệu, mô hình ARGO được xây dựng dựa trên một mô hình Markov kết hợp với dữ liệu của các báo cáo dịch cúm của CDC đã được chuyển đổi logit thành chuỗi $\{y_t\}$ (là mô hình chuỗi thời gian được tạo thành, do sự chuyển đổi hai nguồn dữ liệu dự báo, đây chính là nguyên nhân nội tại ảnh hưởng đến chất lượng của hoạt động dự báo dịch cúm). Nhóm nghiên cứu đã áp dụng một mô hình tự hồi quy với độ trễ N, nhằm giải quyết nhược điểm độ trễ của mô hình của chuỗi thời gian, trong đó tập hợp các thông tin về chuỗi $\{y_{(t-N+1):t}\}_{t\geq N}$ là một chuỗi Markov (điều này chứng tỏ rằng trong thực tế bệnh cúm chỉ kéo dài trong một khoảng thời gian thành từng đợt, không phải kéo dài mãi mãi). Trong công thức 1, chúng ta thấy các chiều hướng chuyển đổi log khối lượng dữ liệu của các truy vấn tìm kiếm của Google tại thời điểm t, X_t chỉ phụ thuộc vào các hoạt động dịch cúm tại thời điểm đó, và dữ liệu chuỗi y_t thu nhận được thông qua sự truy vấn tìm kiếm thông tin về dịch cúm của người dân từ Google (theo trực giác thì dịch cúm xảy ra khiến cho người dân phải tìm kiếm thông tin liên quan đến dịch cúm trên mạng trực tuyến). Do vậy, các thông tin về chuỗi Markov đối với khối lượng dữ liệu thu được $y_{(t-N+1):}$ là một hàm có cấu trúc mô hình ẩn như công thức (1)

$$\begin{array}{ccccc} y_{1:N} & \rightarrow y_{2:(N+1)} & \rightarrow \cdots \rightarrow & y_{(t-N+1):T} & (1) \\ \downarrow & \downarrow & & \downarrow & \\ X_N & X_{N+1} & & X_T & \end{array}$$

Các giả thuyết chính được đưa ra:

Giả thuyết 1:

$$y_t = \mu_y + \sum_{j=1}^{N} \propto_j y_{t-1} + \in_t, \in_t \underset{\sim}{iid} N(0, \sigma^2)$$

Giả thuyết 2: $X_t|y_t \sim N_K(\mu_x + y_t\beta, Q)$

Giả thuyết 3: y_t, X_t là biến độc lập $\{y_l, X_l : l \neq t\}$

Trong đó: $\beta=(\beta_1, \beta_2, \dots, \beta_k)^T$, $\mu_x=(\mu_{x1}, \mu_{x2}, \dots, \mu_{xK})^T$, và Q là ma trận hiệp phương sai.

iid: (independent and identically distributed): Lấy mẫu độc lập và có cùng một phân phối chuẩn.

Trong mô hình phân tích dữ liệu R đối với các biến tham số đơn giản, nhóm nghiên cứu thực hiện chuyển đổi log hóa toàn bộ dữ liệu dịch cúm thu được của CDC mức gốc p_t thành quy mô [0,1] bằng phần mềm R để thu được chuỗi y_t, và cũng chuyển đổi log hóa toàn bộ khối lượng dữ liệu có tương quan với Google ở mức i thành quy mô [0,100] bằng phần mềm R để thu được chuỗi X_t. Nhóm nghiên cứu sử dụng hàm log là phù hợp, vì tần suất truy vấn tìm kiếm thông tin của Google thường có tốc độ tăng theo cấp số nhân và luôn có xu hướng tiến sát tới giới hạn biên mà nhóm nghiên cứu đang cố gắng thu nhỏ để phù hợp với quy mô [0,100] bằng cách chia tối đa các đoạn dữ liệu để xử lý. Mặt khác, dữ liệu Google Trends được sử dụng là số nguyên từ 0 đến 100, nên chúng ta thêm một số lượng nhỏ δ=0,5 trước khi chuyển đổi log để tránh các giá trị log 0 là trường hợp không xác định được. Trong đó, $f(y_t|y_{1:(t-1)}, X_{1:t})$ là hàm giản đơn mô tả phân phối ước tính, với trung bình là $y_{(t-N):(t-1)}$ và X_t; và có phương sai không đổi (xem công thức 2, xác định mô hình ARGO).

Mô hình ARGO được xác định là mô hình chuỗi thời gian hay chính là hàm $y_t = \text{logit}(p_t)$;

Trong đó y_t là hàm chuyển đổi logit dữ liệu thông tin thu được về dịch cúm của CDC có quyền số, hoạt động dịch cúm mức p_t tại thời điểm t, và $X_{i,t}$ là hàm chuyển đổi log có dữ liệu tương quan với Google của mức i tại thời điểm t. Mô hình ARGO được xác định bởi:

$$y_t = \mu_y + \sum_{j=1}^{N} \propto_j y_{t-j} + \sum_{i=1}^{K} \beta_i \boldsymbol{X_{i,t}} + \in_t, \in_t \underset{\sim}{iid} N(0, \sigma^2) \qquad (2)$$

Với X_t được coi là các biến ngoại sinh trong chuỗi thời gian $\{y_t\}$.

Biến tham số dự báo của mô hình ARGO: Nhóm nghiên cứu đã lựa chọn quan sát hoạt động dịch cúm trong khoảng thời gian là 1 năm (N = 52 tuần), và lựa chọn 100 thuật ngữ (được coi là các biến tham số độc lập) có tần suất được truy vấn và tìm kiếm từ các dữ liệu có mối tương quan với google (K = 100). Vì chúng ta có số lượng biến độc lập nhiều hơn so với số biến lượng quan sát (K=100 >N=52), nên phương pháp ước lượng thường được sử dụng là phương pháp bình phương nhỏ nhất sẽ không giải quyết được. Do đó, nhóm nghiên cứu sẽ phải áp dụng các hình thức xử lý cho các biến tham số ước lượng. Nhóm nghiên cứu đã đưa ra 3 hình thức xử lý, hình thức xử lý phổ biến L1, hình thức xử lý đặc biệt L2, và hình thức xử lý kết hợp L1 và L2. Tất cả các thông số được điều chỉnh tự động ở từng tuần quan sát và được lưu lại trong một bảng dữ liệu với mỗi cột dữ liệu chứa thông tin 2 năm về dịch cúm (104 tuần).

Trong một tuần bất kỳ, mục tiêu là tìm ra các tham số μ_y, $\alpha=(\alpha_1, \dots, \alpha_{52})$, và $\beta=(\beta_1, \dots, \beta_{100})$ là nhỏ nhất.

$$\sum_t (y_t - \mu_y - \sum_{j=1}^{52} \propto_j y_{t-j} - \sum_{i=1}^{100} \beta_i X_{i,t})^2 + \lambda_\propto \|\alpha\|_1 + \eta_\propto \|\propto\|_2^2 + \lambda_\beta \|\beta\|_1 + \eta_\beta \|\beta\|_2^2 \qquad (3)$$

Với λ_α, λ_β, η_α, và η_β là 4 tham số thượng tầng ảnh hưởng tới kết quả của mô hình dự báo. Ý tưởng của nhóm nghiên cứu sử dụng phương pháp thống kê kiểm tra chéo để chọn lựa 4 tham số này. Tuy nhiên, qua bảng dữ liệu kết quả chúng ta thấy được kết quả ở mỗi tuần là khá rõ ràng, vì chúng ta chỉ có 104 điểm dữ liệu (tương ứng dữ liệu 2 năm về dịch cúm trong một cột dữ liệu). Như vậy, chúng ta cần phải xác định trước một số các tham số ở trong công thức (3) để làm gốc so sánh các kết quả dự báo. Xuất phát từ mô hình giản đơn theo công thức 1 và kết hợp với các dữ liệu quan sát trực giác từ phương pháp kiểm tra chéo, nhóm nghiên cứu đưa ra giả thuyết $\lambda_\alpha = \lambda_\beta = 0$, từ đó đề xuất hình thức xử lý L1 áp dụng toàn bộ các mô hình thử nghiệm. Với $\lambda_\alpha \# \lambda_\beta$, kết quả nhận được vẫn có sai số đáng kể. Tiếp tục, nhóm nghiên cứu xem xét giải thiết $\lambda_\alpha = \lambda_\beta$ khi đó, mô hình ARGO được xác định chính là công thức (3). Với giả thuyết $\eta_\alpha = \eta_\beta = 0$ và $\lambda_\alpha = \lambda_\beta$.

Số liệu chính xác: Các chỉ số RMSE, MAE, và MAPE của ước tính $\hat{\boldsymbol{p}}$ theo mục tiêu dự báo mức độ hoạt động **p** được xác định, tương ứng, như sau:

$$RMSE(\hat{p_t}, p_t) = \left(\frac{1}{n}\sum_{t=1}^{n}(\hat{p_t} - p_t)^2\right)^{1/2}$$

$$MAE(\hat{p_t}, p_t) = \frac{1}{n}\sum_{t=1}^{n}|\hat{p_t} - p_t|$$

$$MAPE(\hat{p_t}, p_t) = \frac{1}{n}\sum_{t=1}^{n}|\hat{p_t} - p_t|/p_t$$

Hệ số tương quan mẫu được xác định là hệ số tương quan của ước tính $\hat{p}$ theo mục tiêu dự báo mức hoạt động p. Ngoài ra, lượng tăng/giảm của các hệ số tương quan giữa $\hat{p}_t$ và p_t được xác định như sau:

Lượng tăng/giảm của các hệ số tương quan Corr. ($\hat{p_t}$, p_t) = Corr ($\hat{p_t} - \hat{P}_{t-1}$, $p_t - p_{t-1}$).

Mức độ hiệu quả tương đối của ước tính $\hat{p}^{(1)}$ *so với* $\hat{p}^{(2)}$ là $e(\hat{p}^{(1)}, \hat{p}^{(2)}) = MSE^{(2)}_{đúng}/MSE^{(1)}_{đúng}$,

với điều kiện $MSE^{(i)}_{đúng} = E\left((\hat{p}^{(i)} - p)^2\right)$ hoặc được xác định bởi công thức sau:

$$e(\hat{p}^{(1)}, \hat{p}^{(2)}) = \frac{MSE^{(2)}_{obs}}{MSE^{(1)}_{obs}},$$

Trong đó:

$$MSE^{(i)}_{obs} = \frac{1}{n}\sum_{t=1}^{n}(\hat{p}_t^{(i)} - p_t)^2 \quad (4)$$

Khoảng tin cậy 95% được xây dựng, tính toán thông qua phương pháp Bootstrap áp dụng cho các mô hình chuỗi thời gian, với giả thiết các chuỗi thời gian được nhân rộng và có cùng các lỗi sai số do sử dụng các khối ngẫu nhiên được phân bố hình học với độ dài trung bình quan sát là 52 tuần (tương ứng với 1 năm quan sát). Khi đó chúng ta tính xác định được khoảng tin cậy bằng phương pháp Bootstrap có giá trị cơ bản là $\log\{e(\hat{p}^{(1)}, \hat{p}^{(2)})\}$. Sau đó, chúng ta tiến hành lũy thừa để khôi phục lại quy mô xem xét ban đầu khi có tham số. Vì khoảng tin cậy theo phương pháp Bootstrap không có tham số sẽ làm mất đi tính tự tương quan và tương quan chéo của các lỗi trong bộ dữ liệu quan sát, và không chính xác bằng chỉ số sai số trung bình của cả dãy.

Ghi chú: Để tìm hiểu chi tiết thêm về phương pháp luận, xem thêm Phụ lục, http://arxiv.org/pdf/1505.00864v2.pdf.

Công Hoan (dịch)

Nguồn: Hội thảo khoa học quốc tế IASC-ARS2015, Hiệp hội Toán Thống kê Quốc tế, ngày 17-19/12/2015 tại Singapore với chủ đề Toán thống kê: Cơ hội và thách thức trong kỷ nguyên Dữ liệu lớn.