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**

Tóm tắt:

Theo dõi dịch cúm chính xác trong thời gian thực (thời điểm theo dõi và ghi chép lại dịch cúm là đồng thời hoặc ngay sát sau khi bùng phát dịch cúm). Điều này sẽ giúp các cơ quan y tế đưa ra quyết định kịp thời và có thể cứu sống được nhiều người. Nhóm nghiên cứu đã đề xuất một mô hình theo dõi dịch cúm ARGO (mô hình tự hồi quy sử dụng dữ liệu lớn được tìm kiếm trực tuyến của Google, bao gồm nguồn dữ liệu đầu vào từ Google Trends sẵn có công khai và nguồn dữ liệu có tương quan với Google). Với một nền tảng thuật toán thống kê tốt, giúp cho mô hình ARGO xử lý nhanh hơn so với các mô hình theo dõi trước đó của Google. Mô hình này có khả năng kết hợp dữ liệu các vùng dịch cúm theo mùa vụ và nắm bắt được những xu hướng thay đổi trong hành vi tìm kiếm trực tuyến của người dân về dịch cúm theo thời gian. Ngoài ra, mô hình này cũng rất linh hoạt, tự động điều chỉnh sửa lỗi,v.v... làm cho nó trở thành một công cụ có khả năng mạnh mẽ có thể được sử dụng để theo dõi các sự kiện xã hội trong thời gian thực ở nhiều mức độ quan sát không gian và thời gian khác nhau.

Hiện nay, dữ liệu lớn được tạo ra liên tục là từ các hoạt động thu thập thông tin của hàng triệu người dùng trên mạng internet toàn cầu. Nhiều nghiên cứu đã cho thấy tiềm năng to lớn của dữ liệu lớn để phát hiện quản lý các ổ dịch (cúm, Ebola, sốt xuất huyết) dự báo giá cổ phiếu và giá nhà đất, v.v… Năm 2009, một hệ thống phát hiện dịch bệnh kỹ thuật số GFT (gọi là Google Flu Trends) sử dụng dữ liệu là khối lượng lớn các thuật ngữ (từ ngữ) được truy vấn của Google để dự báo hoạt động, hướng lây lan và phát triển của dịch bệnh giống dịch cúm hiện nay (gọi tắt ILI). Sự hoạt động của mô hình GFT đã được xác định bởi nhiều tiêu chí. Ví dụ, làm thế nào dữ liệu lớn sẽ làm thay đổi cách thức phân tích dự báo thống kê truyền thống.

Tuy nhiên, sự khác biệt giữa các kết quả dự báo dịch cúm của GFT và Trung tâm kiểm soát dịch bệnh (CDC - Centers for Disease Control) ở Hoa Kỳ trong những năm sau đó đã dẫn đến nghi ngờ về độ chính xác dự báo của mô hình GFT. Mặc dù nhiều báo cáo đã chỉ ra sai sót trong phương pháp luận của thuật toán GFT ban đầu và đã có những chỉnh sửa[9]. Do đó, theo tác giả mô hình ARGO có nhiều điểm vượt trội so với các mô hình khác, nhưng không hẳn đây là một mô hình hoàn hảo. Trong bài viết này, tác giả chỉ tập trung vào những điểm nổi bật của một mô hình có thuật toán xử lý thông minh, nhanh hơn so với tất cả các mô hình hiện tại cho phép nó có thể theo dõi chính xác tình hình dịch bệnh bằng cách sử dụng dữ liệu lớn được truy vấn tìm kiếm trực tuyến của Google.

Bài trình bày tại 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.

[9] *xem http:// www. googlesearch.Blogspot.com/2014/10/google-flu-trends-gets-brand-new-engine.html*

Dịch cúm gây ra hơn 500.000 ca tử vong mỗi năm trên toàn cầu, trong đó có khoảng 3.000 - 50.000 ca tử vong mỗi năm ở Hoa Kỳ. Theo tác giả, để mô hình này đạt được hiệu quả cho việc chuẩn bị và ứng phó với các dịch cúm phụ thuộc nhiều vào sự sẵn có của các dữ liệu báo cáo chính xác về hoạt động của dịch bệnh trong thời gian thực. Hiện nay, dự báo theo mô hình chuỗi thời gian, khoảng thời gian, và tầm quan trọng về dịch cúm vẫn còn chưa được nhận thức đầy đủ. Cũng như các phương pháp hiện có chưa đủ khả năng để theo dõi hoạt động dịch cúm toàn cầu. Ví dụ như, hệ thống theo dõi hoạt động dịch cúm của CDC ở Hoa Kỳ, luôn có sẵn những báo cáo về số lượng bệnh nhân đến khám ở các bệnh viện với các triệu chứng dịch bệnh giống cúm (www.cdc.gov/flu/). Mặc dù, đây chỉ là những báo cáo thông tin không đầy đủ về tình hình dịch cúm tại Hoa Kỳ, nhưng nó lại rất quan trọng giúp cho các nhà quản lý phân bổ nguồn lực để chuẩn bị và ứng phó với các đợt tăng số lượng các bệnh nhân đến khám tại các cơ sở bệnh viện sẽ có khả năng xảy ra tại các khu vực tại Hoa Kỳ.

Ý nghĩa: Dữ liệu lớn được tạo ra từ Internet có tiềm năng lớn trong theo dõi và dự đoán các hoạt động xã hội. Trong bài viết này, nhóm nghiên cứu tập trung vào theo dõi dịch bệnh cúm. Nhóm nghiên cứuđề xuất một mô hìnhsử dụng dữ liệu tìm kiếm được công bố công khai trên Google để ước tính mức độ hoạt động các loại bệnh cúm. Mô hình của nàynhanh hơn so với tất cả mô hình có sẵn theo dõi thời gian thực của Google dựa trên tìm kiếm các mô hình dịch cúm ở cấp quốc gia của Hoa Kỳ, bao gồm cả GTF. Mô hình này linh hoạt,tự điều chỉnh, mạnh mẽ, và khả năng mở rộng, là một công cụ mạnh mẽ có thể được sử dụng để đánh giá và dự đoán chính xác về mức độ thời gian và không gian cho các sự kiện xã hội khác nhau.

Báo cáo dịch cúm của CDC thường bị chậm từ 1-3 tuần, vì phải sàng lọc, xử lý và tổng hợp thông tin. Khoảng thời gian này là không tối ưu cho mục đích ra quyết định. Để rút ngắn thời gian, nhiều phương pháp đã được đề xuất kết hợp để dự báo dịch cúm, như kết hợp thêm dữ liệu khí hậu, dân số và dịch tễ học với các mô hình toán học. Trong những năm gần đây, phương pháp khai thác thông tin trên Internet cũng đã được đề xuất, như tìm kiếm mạng trực tuyến Google, Yahoo, Baidu, Twitter, kho cơ sở dữ liệu mở Wikipedia, cơ sở dữ liệu sàng lọc truy vấn, và dữ liệu điện toán đám mây có thể tạo dữ liệu báo cáo tự động dựa trên ứng dụng thiết bị di động như mạng dự báo dịch cúm châu Âu, Hệ thống giám sát sức khỏe trực tuyến để phát hiện dịch cúm (Flutracking) ở Úc, và Dự án về theo dõi Dịch cúm của Hoa Kỳ. Trong số đó, mô hình GFT của Google đã nhận được sự chú ý nhiều nhất, vì lấy ý tưởng từ hệ thống phát hiện dịch bệnh kỹ thuật số. Điều thú vị, Google luôn bảo mật thông tin dữ liệu thô dùng cho dự báo. Do đó, chúng ta không thể mô phỏng giống như các kết quả dự báo của GFT.

Trong bài viết này, nhóm nghiên cứu chỉ tập trung vào ba hạn chế liên quan đến thuật toán của mô hình GFT ban đầu đã được xác định.

Thứ nhất, cách tiếp cận vấn đề cứng nhắc. Nghĩa là, thuật toán của mô hình GFT ban đầu đã không cập nhật các kết quả mới nhất trong các báo cáo dịch cúm của CDC về quá trình phát triển của mùa dịch cúm. Do đó, đưa ra mô hình dự báo không rõ ràng và dẫn đến các ước tính không chính xác về tình hình dịch bệnh.

Thứ hai, trong mô hình GFT ban đầu các thuật ngữ được coi như là các biến tham số độc lập. Do vậy, ý tưởng gộp nhóm các biến

tham số (thuật ngữ độc lập) được truy vấn nhiều vào thành một nhóm biến duy nhất, mà không cho phép thay đổi kết quả hiển thị của các thuật ngữ khi người dân truy vấn tìm kiếm thông tin liên quan đến dịch cúm. Do đó đã làm thay đổi kết quả ước tính về dịch cúm.

Thứ ba, thuật toán của mô hình GFT ban đầu bỏ qua các yếu tố có liên quan đến các biến tham số của mô hình chuỗi thời gian khi quan sát dịch cúm. Như là, khi xem xét quá trình hoạt động dịch bệnh cúm theo mùa vụ trong quá khứ, để từ đó nhìn nhận ra những thông tin quan trọng có thể giúp các dự đoán chính xác ILI trong thời gian thực tại.

Đóng góp của nhóm nghiên cứu

Mô hình ARGO đã đưa ra cách thức dự báo mức độ hoạt động dịch bệnh mạnh mẽ và chính xác bằng cách giải quyết ba hạn chế của mô hình GFT ban đầu đã được đề cập ở trên. Ngoài ra, đây cũng là lần đầu tiên (theo tác giả), chúng ta được biết đến một phương pháp luận thống kê sử dụng ngôn ngữ Văn học để giải thích mô hình thống kê theo dõi và phát hiện dịch bệnh thông qua kết hợp mô hình nhân quả với một mô hình thống kê Markov (Mô hình dùng để xử lý các thuật ngữ tự nhiên). Mô hình này được coi là một trong các mô hình thống kê đặc biệt. Mô hình ARGO đã đạt được một số mục tiêu: (i) tự động kết hợp với các thông tin từ báo cáo ILI của CDC mới nhất, (ii) tự động lựa chọn các truy vấn tìm kiếm của Google có liên quan đến dự báo về dịch bệnh, (iii) cải thiện được công tác dự báo trong khoảng thời gian dài hạn theo chu kỳ (mùa dịch), từ mùa dịch cúm 2015 vừa qua các thông tin theo dõi dịch cúm đã được cập nhật như là dữ liệu đầu vào của hệ thống dự báo dịch cúm, (iv) Đối với nghiên cứu, chúng ta có thể theo dõi mô hình thông qua bảng dữ liệu với mỗi cột là thông tin dịch cúm được ghi nhận lại trong 2 năm liên tiếp, cho phép chúng ta nắm bắt những thay đổi gần đây nhất và những hành vi tìm kiếm thông tin của người dùng về dịch bệnh. Đóng góp của nhóm nghiên cứu là xây dựng được mô hình dự báo thống kê dịch bệnh hiệu quả từ hai nguồn dữ liệu (nguồn dữ liệu truy vấn trực tuyến từ Google và nguồn dữ liệu từ các báo cáo dịch cúm của CDC trong quá khứ). Việc xây dựng mô hình dự báo chỉ đơn giản là kết hợp giữa các điều khoản tham chiếu của mô hình chuỗi thời gian với mô hình GFT ban đầu dựa trên các thông tin dữ liệu sẵn có, vì mô hình GFT ban đầu không được tổng hợp tối ưu để cung cấp thông tin theo chuỗi thời gian. Hơn nữa, nhóm nghiên cứu cũng cung cấp một công cụ so sánh hiệu quả ý nghĩa thống kê các cải tiến của phương pháp ARGO so với các phương pháp khác. Ví dụ, phương pháp sử dụng ARGO của nhóm nghiên cứu là phương pháp đo lường chính xác gấp 2 lần so với các phương pháp thử nghiệm khác. Cuối cùng, mặc dù mô hình ARGO chỉ sử dụng dữ liệu được công bố công khai, chất lượng thấp của Google Trends và dữ liệu có tương quan với Google, nhưng đã cải thiện đáng kể mức độ dự báo so với các phiên bản mới nhất của Google theo dõi dịch cúm.

Nhóm nghiên cứu đặt tên cho mô hình là ARGO (Auto Regression with Google search data - mô hình tự hồi quy với dữ liệu truy vấn của Google). Trong thống kê, ARGO là một mô hình tự hồi quy, với các biến tham số trong mô hình là các biến ngoại sinh, và sử dụng các hình thức xử lý phổ biến L_1 (và các hình thức xử lý tiềm năng L_2) được lựa chọn theo quy tắc tự động để có được các dữ liệu về những thông tin dịch cúm có liên quan chặt chẽ với nhau.

Kết quả

Chúng ta có thể sử dụng mô hình ARGO để dự báo về mức độ hoạt động dịch cúm ILI của CDC trong khoảng thời gian từ 29/3/2009 đến 11/7/2015. Khi chúng ta truy cập vào báo cáo dịch cúm của CDC trong quá khứ, chúng ta

chỉ thấy có dữ liệu cập nhật đến trước tuần chúng ta muốn dự báo về dịch cúm. Với mô hình ARGO, chúng ta có được những thông tin so sánh và dự báo chính xác nhất cho tuần hiện tại: vì CDC luôn cập nhật báo cáo các mức độ hoạt động quan trọng ILI. Đối với việc công bố thường chậm 1 hoặc 2 tuần là do phải xử lý các sai số liên quan đến số liệu mô tả dịch cúm như: căn bậc hai của sai số trung bình (RMSE), sai số trung bình tuyệt đối (MAE), sai số phần trăm trung bình tuyệt đối (MAPE), hệ số tương quan, và lượng tăng/giảm của hệ số tương quan. Để so sánh, chúng ta cần phải tính toán những số liệu chính xác về: (i) GFT (cập nhật số liệu 11/7/2015), (ii) Mô hình dự báo Santillana et al.[10], (iii) Mô hình dự báo kết hợp giữa mô hình GFT với mô hình tự hồi quy AR(3) gọi là mô hình GFT+AR(3), (iv) Mô hình tự hồi quy với thời gian linh hoạt AR(3); (v) Mô hình dự báo đơn giản (thô) chỉ sử dụng các giá trị ILI của CDC từ các tuần trước đó làm dự báo cho thời điểm hiện tại. Để so sánh một cách công bằng, các mô hình chuẩn (ii-iv) được theo dõi tự động bằng một bảng với mỗi cột là thông tin dữ liệu 2 năm liên tiếp.

Bảng 1 mô tả tóm tắt các số liệu chính xác của các mô hình dự báo trong khoảng thời gian nghiên cứu. Ở cột "**Khoảng thời gian theo dõi**" cho thấy kết quả ước tính của mô hình ARGO tốt hơn tất cả các mô hình thử nghiệm khác. Các cột khác trong Bảng 1 cũng cho thấy chi tiết hiệu suất của từng mô hình dự báo cho mùa bùng phát dịch cúm H1N1 năm 2009, và các mùa dịch cúm thông thường từ năm 2010 trở đi.

Bảng 1: So sánh các dự báo các dịch bệnh cúm bằng các mô hình khác nhau

Mô hình	Khoảng thời gian theo dõi (từ 29/3/2009 đến11/7/2015)	Bệnh cúm theo mùa H1N1 (Do siêu virus gây ra)	Quan sát mùa cúm thông thường (tuần thứ 40 của năm trước đến tuần thứ 20 năm hiện tại)				
			2010-2011	2011-2012	2012-2013	2013-2014	2014-2015
Căn bậc hai của sai số trung bình							
ARGO	**0.608**	0.640	**0.596**	**0.807**	**0.687**	**0.306**	**0.438**
GFT (10/2014)	2.216	0.773	1.110	3.023	4.454	0.986	0.700
Santillana et al	0.915	0.833	0.881	2.027	1.090	0.446	0.663
GFT+AR(3)	0.912	**0.580**	0.602	1.382	1.279	0.993	0.906
AR(3)	0.957	0.813	0.794	1.051	1.191	0.969	0.928
Đơn giản/thô	1(0.348)	1(0.600)	1(0.339)	1(0.163)	1(0.499)	1(0.350)	1(0.465)
Sai số trung bình tuyệt đối							
ARGO	**0.649**	0.584	**0.574**	**0.748**	**0650**	**0.391**	**0.530**
GFT (10/2014)	1.834	0.777	1.260	3.277	5.028	0.891	0.770
Santillana et al	1.052	0.719	1.010	2.211	1.029	0.610	0.820
GFT+AR(3)	0.888	**0.570**	0.613	1.308	1.016	1.034	0.839
AR(3)	0.925	0.777	0.787	0.951	0.988	0.917	0.934
Đơn giản/thô	1(0.201)	1(0.425)	1(0.259)	1(0.135)	1(0.325)	1(0.212)	1(0.295)

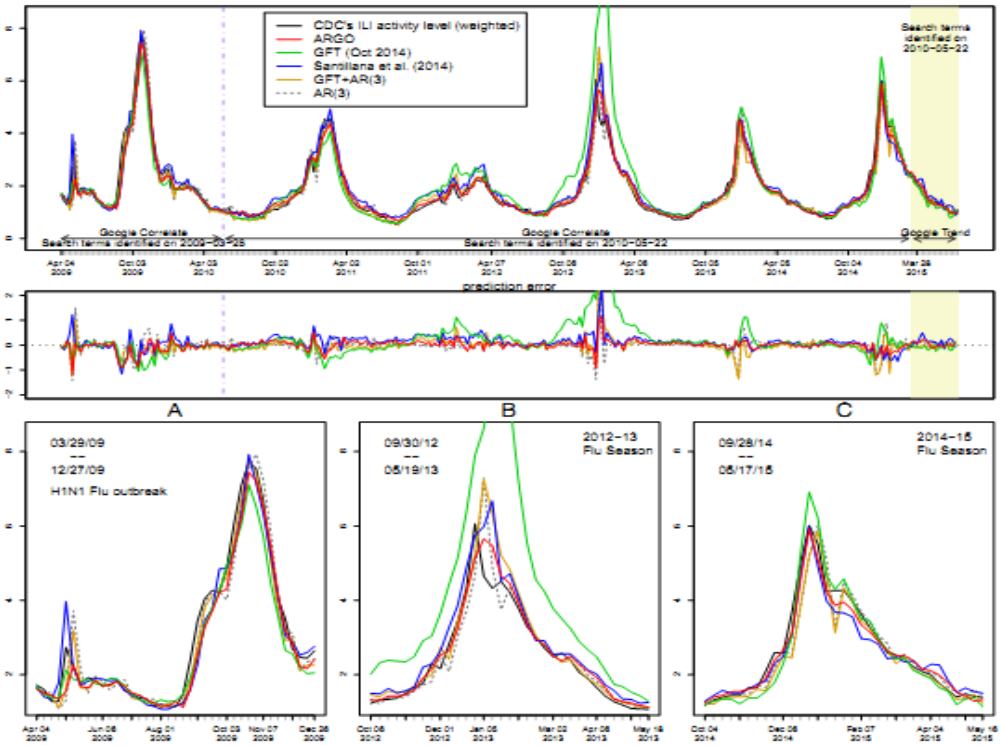
[10] *Xem Santillana M, Zhang DW, Althouse BM, Ayers JW (2014) What can digital disease detection learn from (an external revision to) Google Flu Trends? Am J Prev Med 47(3):341–347*

Mô hình	Khoảng thời gian theo dõi (từ 29/3/2009 đến11/7/2015)	Bệnh cúm theo mùa H1N1 (Do siêu virus gây ra)	Quan sát mùa cúm thông thường (tuần thứ 40 của năm trước đến tuần thứ 20 năm hiện tại)				
			2010-2011	2011-2012	2012-2013	2013-2014	2014-2015
Sai số phần trăm trung bình tuyệt đối							
ARGO	**0.787**	**0.620**	**0.663**	**0.770**	**0.719**	**0.453**	**0.620**
GFT (10/2014)	1.937	0.721	1.394	3.442	5.419	0.892	0.895
Santillana et al	1.381	0.765	1.380	2.306	1.251	0.754	0.958
GFT+AR(3)	1.037	0.683	0.698	1.407	0.986	1.062	0.828
AR(3)	1.003	0.894	0.814	0.947	0.939	0.891	0.916
Đơn giản/thô	1(0.090)	1(0.139)	1(0.105)	1(0.081)	1(0.110)	1(0.084)	1(0.097)
Hệ số tương quan							
ARGO	**0.986**	0.985	**0.989**	**0.928**	**0.968**	**0.993**	**0.993**
GFT (10/2014)	0.875	**0.989**	0.968	0.833	0.926	0.969	0.986
Santillana et al	0.971	0.967	0.983	0.927	0.956	0.985	0.984
GFT+AR(3)	0.967	0.986	0.985	0.879	0.929	0.945	0.957
AR(3)	0.964	0.968	0.971	0.877	0.903	0.927	0.945
Đơn giản/thô	0.961	0.951	0.954	0.887	0.924	0.923	0.937
Lượng tăng/giảm của hệ số tương quan							
ARGO	**0.758**	0.806	**0.810**	0.286	**0.527**	**0.938**	0.912
GFT (10/2014)	0.706	**0.863**	0.702	0.484	0.502	0.847	**0.918**
Santillana et al	0.690	0.776	0.693	**0.510**	0.367	0.915	0.889
GFT+AR(3)	0.512	0.708	0.708	0.165	0.141	0.534	0.587
AR(3)	0.385	0.585	0.569	0.077	0.011	0.404	0.493
Đơn giản/thô	0.436	0.602	0.570	0.095	0.134	0.406	0.514

GFT+AR(3) là mô hình dự báo GFT có phương trình: $p_t = \mu + a_1 p_{t-1} + a_2 p_{t-2} + a_3 p_{t-3} + \beta GFT(t)$; các biến tham số trong mô hình là các biến ngoại sinh. Những số liệu in đậm trong Bảng 1 thể hiện hiệu suất tốt nhất trong từng giai đoạn nghiên cứu về dịch cúm. Các giá trị RMSE, MAE, và MAPE là các lỗi sai liên quan đến các phương pháp nghiên cứu; Sai số tuyệt đối của phương pháp đơn giản (thô) được ghi trong ngoặc đơn; và tất cả so sánh đều dựa trên quy mô mức độ hoạt động ILI ban đầu.

Hình 1 cho thấy trong mùa dịch cúm thông thường sau năm 2009, mô hình ARGO đều đưa ra các kết quả tốt hơn so với tất cả các mô hình dự báo thay thế khác dựa trên các chỉ số RMSE, MAE, MAPE, và hệ số tương quan. Mô hình ARGO đã tránh được những vấn đề của mô hình GFT1. Mùa bùng phát dịch cúm H1N1 năm 2009, ARGO đã tính toán MAPE nhỏ nhất với RMSE = 0,640 , MAE = 0,584) có hiệu suất tốt đứng thứ 2 chỉ sau mô hình GFT + AR (3) (RMSE = 0,580 & MAE = 0,570). Xét hệ số tương quan, ARGO là (r = 98,5%) có hiệu suất tương đương (khả năng có dữ liệu cần tìm kiếm trong mẫu dữ liệu thu thập) với mô hình GFT (r= 98,9%) (14) và mô hình GFT+AR (3) (r = 98,6%) và vượt trội hơn so với tất cả các mô hình khác thử nghiệm.

Hình 1: Đồ thị các ước tính của các mô hình so sánh với báo cáo mức độ hoạt động ILI của CDC



(Hình trên cùng): là các đường mô hình dự báo mức độ hoạt động ILI của các phương pháp khác nhau theo dõi dịch cúm: ARGO: màu đỏ đậm; CDC: màu đen đậm trái ngược với ARGO dự đoán; GFT: màu xanh lá cây; Santillana et al: màu xanh dương; GFT+AR(3): màu vàng đậm; AR(3): màu xám; Hai màu nền: màu trắng là nguồn dữ liệu từ các nguồn dữ liệu có tương quan với Google, và màu vàng là nguồn dữ liệu có nguồn từ Google Trends. Các đường kẻ dọc đứng nét đứt mầu tím nhằm phân tách dữ liệu có tương quan với Google theo điều khoản tìm kiếm được xác định từ trước ngày 28/3/2009 và từ sau ngày 28/3/2009 đến ngày 22/5/2010.

(Hình ở giữa) Sai số dự báo, được xác định là giá trị ước tính trừ đi giá trị mức độ hoạt động ILI của CDC;

(Hình dưới cùng) Kết quả ước tính trong các giai đoạn nghiên cứu khác nhau đã được phóng đại (A) Các giai đoạn bùng phát dịch cúm H1N1. (B) Mùa cúm thông thường mùa dịch 2012-2013. (C) Mùa cúm thông thường mùa dịch 2014-2015. (Một mùa cúm thông thường được định nghĩa là tuần 40 của một năm hiện tại đến tuần 20 của năm sau đó).

Để đánh giá ý nghĩa thống kê về mức độ cải thiện khả năng dự báo chính xác của ARGO, nhóm nghiên cứu đã tính toán và xây dựng một khoảng tin cậy 95% cho phép xác định các điểm đạt hiệu quả tương đối của mô hình ARGO và so sánh với các mô hình khác. Hiệu quả tương đối của mô hình 1 so với mô hình 2 được xác định là tỷ lệ trung bình của bình phương các sai số (MSE) của mô hình 2 so với mô hình 1, có thể được ước tính theo giá trị quan sát của nó (xem công thức 4 phần phương pháp luận); khoảng tin cậy của mô hình được xây dựng bằng phương pháp chọn mẫu có lặp cố định về chuỗi thời gian số dư sai số. Bảng 2 cho thấy mô hình ARGO được ước tính có hiệu quả ít nhất bằng hai lần các mô hình thay thế khác, và cải thiện độ chính xác là rất quan trọng về mặt thống kê.

Bảng 2: So sánh mức độ dự báo hiệu quả tương đối của ARGO so với các mô hình khác với khoảng tin cậy (CI) 95%

Mô hình	Số điểm dự báo chính xác	Khoảng tin cậy (CI) 95%
GFT (10/2014)	12.85	(5.18, 91.82)
Santillana et al	2.02	(1.36, 2.83)
GFT+AR(3)	2.17	(1.23, 4.53)
AR(3)	2.40	(1.56, 3.69)

Bảng 2 được tính toán dựa trên dữ liệu từ 29/03/2009 đến 17/05/2015

Mặc dù nhóm nghiên cứu cũng biết rằng báo cáo có trọng số của CDC sẽ được sửa đổi số liệu ở tuần sau khi công bố, điều này mâu thuẫn với tính thống nhất về dữ liệu trong hệ thống theo dõi, nhưng sự sửa đổi đó cũng giúp cho việc dự báo chính xác hơn về tình hình dịch cúm. Như vậy, các thông tin dịch cúm của CDC trong quá khứ luôn sẵn có trong một tuần bất kỳ, và không nhất thiết phải chính xác như khi nó đã được công khai lần đầu. Nhóm nghiên cứu đã kiểm tra về độ chính xác khi sử dụng dữ liệu đã bị sửa đổi này bằng cách thu thập dữ liệu thông tin hoạt động về dịch cúm của CDC từ các báo cáo ban đầu chưa sửa đổi và các báo đã sửa đổi trong quá khứ, cùng với các dữ liệu từ các trang web theo dõi dịch cúm của CDC ghi nhận tại những ngày dữ liệu đã được sửa đổi, cập nhật trong khoảng thời gian nghiên cứu. Sau đó xây dựng một mô hình dự báo chuỗi thời gian của các mô tả ước tính về toàn bộ về dịch cúm, rồi tiến hành so sánh các mức độ dự báo về dịch cúm giữa các mô hình với nhau và thấy rằng mô hình ARGO vẫn vượt trội hơn tất cả. Hơn nữa, các giá trị về số liệu của mô hình ARGO dự báo cho cả năm là chính xác và không thay đổi. Điều đó cho thấy khả năng không thể sửa đổi các kết quả theo ý muốn của chúng ta trong các báo cáo hoạt động về dịch cúm của CDC.

Tuy nhiên, nhóm nghiên cứu đã phải đối mặt với một thách thức trong việc sản xuất số liệu ước tính dự báo dịch cúm cho các khoảng thời gian mới nhất của mùa dịch cúm 2014-2015. Vì tại thời điểm viết bài này, các dữ liệu mới nhất chỉ cập nhật đến ngày 28/03/2015, và sau thời điểm này các dữ liệu dùng để tính toán được lấy từ nguồn Google Trends. Các thông tin này có chất lượng thấp, thậm chí chất lượng dữ liệu thấp hơn cả các nguồn dữ liệu có tương quan với Google thể hiện rõ thông qua sự cập nhật dữ liệu thay đổi theo tuần. Sự thay đổi về nguồn dữ liệu để dự báo là điều chúng ta không mong muốn, vì sẽ ảnh hưởng đến chất lượng mô hình dự báo. Do vậy, để đánh giá chất lượng, sự ổn định dự báo của mô hình ARGO bởi sự thay đổi nguồn dữ liệu này, nhóm nghiên cứu tiến hành thu thập dữ liệu là các thuật ngữ liên quan đến dịch cúm có tần suất tìm kiếm với cùng một kiểu cách thức truy vấn của Google Trends trong 25 ngày khác nhau ở tháng 04/2015 và sử dụng mô hình ARGO để sản xuất một bộ dữ liệu gồm 25 số liệu ước tính tương ứng với 25 ngày trong tháng. Kết quả cho thấy ba thông số RMSE, MAE, MAPE của mô hình ARGO là ổn định hơn so với mô hình Santillana et al, và nhìn chung thì tốt hơn so với các mô hình khác. Mặc dù mô hình ARGO chỉ sử dụng dữ liệu chất lượng thấp của Google Trends.

(Còn nữa)

Công Hoan (lược dịch)