TƯƠNG LAI CỦA THỐNG KÊ HỌC

(Tiếp theo)

3.3. Những thách thức trong công tác nghiên cứu

Thống kê nòng cốt đang phải đối mặt với những thách thức nào trong công tác nghiên cứu? Việc xác định được những thách thức đó trong ngành Thống kê là hoàn toàn khác với các ngành khoa học khác. Ví dụ, nếu như trong Toán học có cả một danh sách các vấn đề nổi cộm với những thách thức mang tính chất lâu dài thì trong ngành Thống kê, các vấn đề luôn luôn biến đổi và liên quan đến sự phát triển của các cấu trúc dữ liệu mới cũng như các công cụ tính toán mới. Ngoài ra, khác với các ngành khoa học thí nghiệm, Thống kê học không có những vấn đề nghiêm trọng phải đầu tư nhiều tiền của vào các phòng thí nghiệm để chạy đua hoặc hợp tác trên các lĩnh vực chủ chốt. So với các lĩnh vực khoa học khác, dự đoán được những tiến bộ quan trọng nhất của ngành Thống kê có lẽ sẽ khó khăn hơn. Chính vì thế mà chúng ta cần phải có một triết lý cơ bản đủ linh hoạt để có thể thích ứng với các thay đổi. Đồng thời, điều quan trọng là nghiên cứu này về tương lai của ngành Thống kê sẽ không trở thành những phương pháp vô ích.

Có thể xác định được một số chủ đề khái quát đang chi phối hoạt động nghiên cứu trong lĩnh vực Thống kê nòng cốt hiện đại. Đó là những thách thức trong việc phát triển các khuôn khổ khái niệm và các lý thuyết xấp xỉ tiệm cận thích hợp để xử lý rất nhiều các quan sát (có thể có) với nhiều thông số, nhiều tỉ lệ và nhiều biến số phụ thuộc phức tạp. Các mục nhỏ dưới đây sẽ cho thấy chi tiết hơn về những vấn đề này.

3.3.1. Quy mô của số liệu

Sự bùng nổ về số liệu thu thập được đã không còn là chuyện gì mới mẻ. Tuy vậy, các nhà phân tích số liệu toàn phương và thống kê tuyến tính vẫn đề cập: các dữ liệu đã tăng theo cấp số nhân. Nguyên tắc phân loại kích thước số liệu năm 1994 của Huber chia kích thước số liệu thành các dạng: rất nhỏ 10^2, nhỏ 10^4, trung bình 10^6, lớn 10^8, cực lớn 10^{10} gần đây có vẻ như không còn phù hợp (theo Wegman, 1995). Ví dụ như cơ sở dữ liệu đơn cho một thí nghiệm vật lý hạt nhân đơn sử dụng máy dò BaBar tại Trung tâm Máy gia tốc Tuyến tính Stanford có tới 5×10^{15} byte.

Các vấn đề sẽ được tiếp tục nghiên cứu trên tất cả các quy mô còn hiện tại chúng ta vẫn chưa giải quyết hết các vấn đề của các bộ số liệu dưới 100. Tuy nhiên, một thách thức mới của Thống kê là các vấn đề lẫn lộn với nhau, chẳng hạn như khả năng khái quát hóa, khả năng mở rộng, tính thiết thực cũng như chiều sâu hiểu biết khoa học về số liệu sẽ thay đổi theo quy mô và bối cảnh. Hơn nữa, rõ ràng là công tác nghiên cứu và đào tạo sau đại học của chúng ta vẫn chưa nhận thức được đầy đủ về vấn đề tính toán cũng như các vấn đề khác liên quan đến các quy mô lớn hơn.

3.3.2. Thu nhỏ và nén dữ liệu

Chúng ta cần phải có các "nguyên tắc thu nhỏ" mới. R.A. Fisher đã cho chúng ta rất nhiều ý tưởng quan trọng để thu nhỏ dữ liệu, chẳng hạn như các lập luận đầy đủ, phụ thuộc và có điều kiện, các chuyển

đổi, các phương pháp quan trọng, tối ưu tiệm cận và tính bất biến. Tuy nhiên, rõ ràng là cần phải tìm kiếm những ý tưởng mới để định hướng cho các mảng như lựa chọn, dự báo và phân loại mô hình.

Một trong những ý tưởng đó là sử dụng "nén dữ liệu" như là một mô hình định hướng cho phân tích dữ liệu. Ý tưởng cơ bản ở đây chính là: am hiểu về cấu trúc dữ liệu có liên quan đến khả năng lưu trữ thu gọn dữ liệu mà không làm mất khả năng "giải nén" dữ liệu cũng như có thể phục hồi gần như nguyên trạng các thông tin ban đầu. Ví dụ, trong phạm vi các dữ liệu tín hiệu và hình ảnh, các hàm biến thiên là không tối ưu trong việc hiển thị và nén các đường cong. Điều này cho thấy cần phải có các hệ thống hiển thị mới để thực hiện nén dữ liệu tốt hơn.

3.3.3. Phân tích dữ liệu ngoài thống kê

Nhiều phương pháp và chiến lược tính toán, chẳng hạn như Máy học (Machine Learning) hay các mạng Nơ-ron (Neural Network), đã phát triển ngoài lĩnh vực Thống kê. Phần lớn các phương pháp này còn chưa được biết đến rộng rãi như là các phương tiện vô cùng hữu hiệu hỗ trợ cho ngành Thống kê. Vì vậy, những nghiên cứu trong tương lai cần phải liên quan chặt chẽ tới việc tích hợp nhiều phương pháp phân tích các bộ số liệu lớn và phức tạp mà cộng đồng Máy học và các cộng đồng khác đang phát triển nhằm phục vụ cho Thống kê nòng cốt.

Thông thường thì nghiên cứu này có thể được tiến hành dựa trên xây dựng các mô hình và cấu trúc cho phép mô tả về các nguy cơ cũng như đánh giá dựa trên số liệu. Sau đó, nghiên cứu sẽ bao gồm các công cụ nguyên tắc phát triển nhằm thích ứng với việc thực hiện xây dựng mô hình. Một khả năng khác được Breiman đưa ra (năm 2001) là "Nếu mục tiêu của chúng ta là nhằm tạo ra một lĩnh vực sử dụng số liệu để giải quyết các vấn đề thì chúng ta cần phải tránh phụ thuộc vào những mô hình số liệu duy nhất và chấp nhận nhiều công cụ khác".

3.3.4. Phân tích đa biến cho p lớn, n nhỏ

Trong nhiều ứng dụng thống kê quan trọng, các biến số (p) xuất hiện nhiều hơn so với các đơn vị (n), ví dụ như trong phân tích số liệu đường cong, quang phổ, hình ảnh và vi mạng ADN. Một Hội thảo gần đây có tên gọi "Số liệu Đa chiều: p >> n trong Thống kê Toán học và trong các Ứng dụng Y Sinh" đã được tổ chức tại Leiden, Hà Lan. Hội thảo đã nhấn mạnh tầm quan trọng của các nghiên cứu hiện nay về chủ đề này trong nhiều lĩnh vực Thống kê.

Ví dụ sau đây sẽ giải thích cụ thể hơn về việc những đổi mới trong các lĩnh vực khác hữu ích ra sao đối với vấn đề này, qua đó cho thấy Thống kê nòng cốt luôn tìm kiếm các ý tưởng. Lý thuyết ma trận ngẫu nhiên mô tả các mô hình và phương pháp Vật lý Toán học được xây dựng trong hơn 40 năm qua, bắt đầu với việc nghiên cứu các mức năng lượng hạt nhân phức tạp. Trong những năm gần đây, các ý tưởng này đã thu hút nhiều sự quan tâm về xác suất và các tổ hợp.

Đây là thời điểm chín muồi để áp dụng và phát triển những phương pháp này vào các vấn đề đa chiều của Thống kê cũng như trong phân tích dữ liệu. Các nhà khoa học trong nhiều lĩnh vực khác nhau phải làm việc với các ma trận số liệu lớn [các quan sát (n) và nhiều biến số (p)], tuy nhiên hiện nay có rất ít lý thuyết thống kê có thể hỗ trợ và nắm bắt được các phương pháp heuristic nhằm thu gọn các số liệu này trong các phân tích thành phần chủ yếu, các phân tích tương quan tiêu chuẩn, v.v...

Kết quả ban đầu cho thấy: trong một số trường hợp, lý thuyết "n lớn – p lớn" có thể phù hợp và hiệu quả hơn so với các tiệm cận cổ điển "n lớn – p cố

định". Ví dụ, phân phối Tracy-Widom cho "Đồng trực giao Gauss" đưa ra một phân phối đơn, trong đó trọng tâm và tỷ lệ thích hợp đưa ra các mô tả khá chính xác về các phân phối của các thành phần chính và các tương quan tiêu chuẩn trong các trường hợp giả thuyết Không.

3.3.5. Phương pháp Bayes và ước lượng chệch

Thập kỷ 90 mang lại các phương pháp tính toán và khả năng áp dụng hoàn chỉnh các phương pháp Bayes vào nhiều dạng mô hình khác nhau. Thách thức trong thời gian tới chính là phát triển hoàn chỉnh và khai thác mối liên hệ giữa các phương pháp Bayes và những phương pháp về số liệu không tham số và bán tham số hiện đại, bao gồm cả nghiên cứu về sự kết hợp có thể giữa trường phái thống kê Bayes và trường phái tần số.

Rõ ràng là đối với các mô hình có dữ liệu khổng lồ với nhiều biến số, các ý tưởng về ước lượng chệch hoặc "gần" không chệch (như đối với ước tính khả năng tối đa) không còn hữu ích. Nguyên nhân là vì ý tưởng về tổng hợp dữ liệu tiềm ẩn trong hệ phương pháp thống kê không thể sử dụng được với độ phức tạp và tính biến thiên của các phương pháp chệch. Điều này cho thấy cần phải có một "lý thuyết ước lượng chệch" bao quát hơn và các lý thuyết mới nhằm phù hợp với lượng dữ liệu khổng lồ với nhiều biến số.

Với việc sử dụng ngày càng nhiều các phương pháp đó trong tất cả các dạng bài tập xây dựng mô hình, nhu cầu phân tích chuyên sâu hơn về các phương pháp suy luận Monte Carlo ngày càng trở nên rõ ràng hơn.

3.3.6. Khả năng dung hòa giữa kiểm chứng và thử nghiệm tính toán

Một thách thức khác đối với công tác lý luận trong những thập kỷ tới chính là tăng cường khả năng dung hòa giữa tốc độ kiểm chứng (quá chậm) và thử nghiệm tính toán tự do (quá độc đoán và không thuyết phục). Có rất nhiều vấn đề, trong đó các kiểm chứng Toán học chặt chẽ có thể bị bỏ qua trong cả kiểm chứng và thử nghiệm tính toán bởi vì thực hiện các kiểm chứng Toán học là rất khó và cũng không phải quan trọng hàng đầu.

Ví dụ, mặc dù đã được nghiên cứu nhiều năm nhưng vẫn có các nhóm mô hình thống kê quan trọng, chẳng hạn như các mô hình hỗn hợp, có khả năng bị bỏ qua vì việc phân tích khó khăn và liên quan đến hàng loạt các cấu trúc mô hình cần phải nghiên cứu tỉ mỉ.

3.4. Xây dựng và duy trì các hoạt động Thống kê nòng cốt

Việc khai thác các cơ hội đa dạng trong các ngành khoa học hiện nay đã làm tăng nhu cầu về các kiến thức và chuyên môn rộng hơn trong các ứng dụng. Điều này tạo ra thách thức cho ngành Thống kê, cụ thể là tạo ra áp lực cho Thống kê nòng cốt. Theo thời gian, áp lực đó có thể làm giảm hiệu quả trong việc củng cố kiến thức thống kê cũng như khả năng cung cấp kiến thức Thống kê cho các ngành khoa học. Về cơ bản, các hoạt động liên ngành đang ngày một mở rộng và đa dạng hơn. Với tốc độ như vậy, các hoạt động này thậm chí có thể đe dọa tới sự gắn kết trong ngành Thống kê.

Số liệu thu thập và nhu cầu phân tích số liệu tăng lên theo cấp số nhân thì tại sao lại liên quan đến công tác nghiên cứu của Thống kê nòng cốt? Đó là bởi vì thống nhất được các ý tưởng có thể điều chỉnh được sự tăng trưởng này và Thống kê nòng cốt chính là nơi có thể nảy sinh và truyền tải những ý kiến này trong giới khoa học. Vì vậy, trên quan điểm tổ chức và trao đổi có hiệu quả các tiến bộ trong phân tích dữ liệu thì

thúc đẩy hoạt động nghiên cứu của Thống kê nòng cốt thực sự là mục tiêu cơ bản quan trọng cho khoa học.

Thống kê nòng cốt phát triển mạnh (thông qua việc liên kết mạnh mẽ với các ứng dụng) là hy vọng phát triển lớn nhất trong bối cảnh các lĩnh vực đang bùng nổ các phương pháp phân tích dữ liệu như vậy. Tóm lại, đây chính là cơ sở hạ tầng quan trọng đối với các ngành khoa học nói chung.

Trong Chương 4 của báo cáo đầy đủ, chúng ta đã xác định và thảo luận chi tiết những cơ hội và nhu cầu sau đây cho Thống kê nòng cốt:

• *Thích ứng với phân tích số liệu bên ngoài*

Nhu cầu về số liệu không ngừng tăng lên đã mang đến cho các nhà Thống kê những thách thức riêng trong việc cung cấp kiến thức để phát triển các phương pháp phân tích số liệu trong các lĩnh vực khác.

• *Sự phân rã trong hoạt động nghiên cứu của Thống kê nòng cốt*

Các hoạt động bên ngoài không ngừng tăng lên do nhiều nguyên nhân tích cực. Chúng tôi nghĩ rằng sự tăng trưởng này sẽ dẫn tới hậu quả khôn lường – đó là sự sao lãng trong nghiên cứu cơ bản và nguy cơ kèm theo là sự phân rã của ngành Thống kê.

• *Vấn đề nhân lực*

Hơn bao giờ hết, nguồn nhân lực làm việc trong lĩnh vực nghiên cứu của Thống kê nòng cốt tại Hoa Kỳ hiện nay đang bị thu hẹp. Vấn đề nhân lực ngày càng đi xuống một phần là do sự thiếu hụt chung về nhân sự của ngành Thống kê, một phần khác là do các hoạt động Thống kê bên ngoài đang kéo các nhà Thống kê ra khỏi lĩnh vực nghiên cứu nòng cốt.

• *Nhu cầu nghiệp vụ tăng lên*

Các hoạt động nghiên cứu cốt yếu của Thống kê mang tính đa ngành. Điều này thể hiện trong các công cụ của ngành: (Ít nhất) Thống kê vay mượn từ lý thuyết thông tin, Khoa học máy tính, Vật lý cũng như Xác suất và các ngành Toán học truyền thống. Khi các nhà thống kê ngày càng tập trung vào số liệu (để giải quyết các vấn đề về kích thước và phạm vi của các số liệu hiện nay) thì nhu cầu về các kỹ năng cần thiết trong Thống kê nòng cốt cũng tăng lên. Điều này đã đưa ra một thách thức để có thể đảm bảo rằng Thống kê nòng cốt tiếp tục là nơi hội tụ các ý tưởng thống kê.

• *Vấn đề kinh phí nghiên cứu*

Rõ ràng là kinh phí cho hoạt động nghiên cứu của Thống kê nòng cốt chưa theo kịp với sự phát triển của lĩnh vực này. Nguồn kinh phí được tập trung cho các hoạt động Thống kê bên ngoài hoặc hoạt động tư vấn thay vì vướng vào vấn đề kinh phí khó khăn để nuôi dưỡng nhân tài, tạo thời gian và không gian nghiên cứu cho nhà thống kê có thâm niên cũng như khuyến khích những người còn ít thâm niên gắn bó với lĩnh vực nghiên cứu này.

• *Chiến lược mới để tìm kiếm nguồn kinh phí*

Những ý tưởng mới để tìm kiếm nguồn kinh phí cho công tác nghiên cứu có thể giúp cho ngành Thống kê đương đầu với những thách thức của mình. Báo cáo đầy đủ đã đưa ra một ví dụ mở rộng về chiến lược tìm kiếm kinh phí cho phép các nhà Thống kê có thể thực hiện nhiều nghiên cứu Thống kê cơ bản cùng với các hoạt động liên ngành mà không xa rời những nghiên cứu nòng cốt.

(Còn tiếp)

Quỳnh Trang (dịch) - Đoàn Dũng (hiệu đính)

Nguồn: A Report on the Future of Statistics
http://ww.biostat.jhsph.edu/...