TƯƠNG LAI CỦA THỐNG KÊ HỌC
(Tiếp theo)

4.4. Công nghệ thông tin

Sự gia tăng nhanh chóng của kho dữ liệu tin học quy mô lớn đã tác động một cách sâu sắc đến nhiều hoạt động của con người. Đây chính là thời điểm các nhà Thống kê học hoạt động sôi nổi nhất trong các lĩnh vực liên quan đến công nghệ thông tin.

Sự phát triển của mạng toàn cầu và dung lượng tăng lên theo cấp số nhân của các hệ thống máy tính đã mở ra những khả năng không thể tưởng tượng trước đây về trao đổi thông tin, khả năng thu thập và phân tích các bộ số liệu cực lớn về nhiều loại khác nhau từ các nguồn khác nhau, cũng như truyền tải các kết quả thống kê. Sự phát triển của các phần mềm nguồn mở đã mở rộng khả năng ảnh hưởng của năng lực và ý tưởng của các nhà nghiên cứu. Do đó, những thách thức mới trong việc nghiên cứu và xây dựng mô hình thống kê từ dữ liệu là rất lớn. Phần này sẽ tập trung vào các lĩnh vực có tác động lớn được chọn ra.

4.4.1 Thông tin liên lạc

Rất nhiều các số liệu về Thông tin liên lạc được tạo ra trong mỗi phút. Mỗi cuộc gọi trên đường dây hoặc không dây tạo ra một bộ dữ liệu cho biết các thông tin về người thực hiện cuộc gọi, người nhận cuộc gọi, thời gian và địa điểm thực hiện cuộc gọi, thời gian cuộc gọi kéo dài cũng như chi phí cuộc gọi. Mỗi yêu cầu tải một tập tin về từ một địa chỉ Internet được ghi lại trong 1 log file (tập tin nhật ký) hay tải lên một phiên trò chuyện trực tuyến trong một diễn đàn công khai đều được ghi lại.

Các hồ sơ trong thông tin liên lạc như vậy được rất nhiều đối tượng quan tâm, trong đó có các kỹ sư mạng - những người chuyên thiết kế mạng và phát triển các dịch vụ mới, các nhà Xã hội học - những người quan tâm đến cách thức con người kết nối và hình thành các nhóm xã hội ra sao, các nhà cung cấp dịch vụ - những người cần phải tìm ra các gian lận một cách nhanh nhất, cũng như các cơ quan hành pháp và an ninh muốn tìm ra các hoạt động hoạt tội phạm và khủng bố.

Một loạt các vấn đề thách thức về thống kê cần phải được giải quyết trước khi rất nhiều dữ liệu được chuyển thành hàng loạt thông tin.

4.4.2 Máy học và Khai thác số liệu

Các nghiên cứu về Máy học và Khai thác số liệu chủ yếu được thực hiện tại các bộ phận về khoa học vi tính, còn các nghiên cứu về ước lượng phi tham số thì được thực hiện chủ yếu tại các bộ phận thống kê. Có thể thấy ranh giới giữa hai loại nghiên cứu này đã trở nên mờ nhạt đi rất nhiều. Trong thực tế, các nhà Thống kê học thường xuyên sử dụng "Máy học" và "Khai thác số liệu". Phạm vi chính của các nghiên cứu hiệu quả trong các bộ phận thống kê bao gồm các phương pháp mới để phân loại, tập hợp và xây dựng mô hình dự báo. Các nhà thống kê đã mất khá nhiều thời gian để phát triển các công cụ phân loại, nhưng chính sự bùng nổ về khả năng vận hành máy tính cùng với những thành quả đạt được trong các nghiên cứu gần đây đã mang lại những tiến bộ mới quan trọng.

Một trong những tiến bộ đạt được về phân loại mang lại lợi ích trong thực tiễn chính là các Máy Véc tơ hỗ trợ. Phương pháp này rất phổ biến trong các cộng đồng Máy học thuộc Khoa học máy tính và hưởng lợi từ đầu vào của các nhà Thống kê - những người đã góp phần vào việc tìm hiểu về các thuộc tính của phương pháp. Tuy nhiên bên cạnh đó còn có những cơ hội quan trọng khác để hiểu sâu hơn về các thuộc tính lý thuyết của công cụ này cũng như về phương pháp thích hợp và hiệu quả nhất để sử dụng công cụ nhằm phục hồi thông tin từ số liệu trong nhiều trường hợp.

4.4.3 Ngành mạng máy tính

Ngành nghiên cứu về lưu lượng Internet nhìn chung có thể được chia thành các lĩnh vực: đo lường và mô hình hóa lưu lượng truy cập, Topo mạng (cấu trúc hình học không gian của mạng) và phân lớp mạng. Tất cả các lĩnh vực này đều có những thách thức thống kê quy mô lớn.

Chính nhu cầu cùng nhau nâng cao hiệu quả và chất lượng dịch vụ đã thúc đẩy hơn nữa việc nghiên cứu thêm về lĩnh vực đo lường và mô hình hóa. Các phương pháp tiếp cận hiện nay về chất lượng dịch vụ dựa trên quá nhiều nguồn. Điều này vừa lãng phí vừa không có hiệu quả vì sự bùng nổ về lưu lượng truy cập một phần là do giao thức không phù hợp và các quy trình lộn xộn. Nhiều ý kiến đã được đề xuất nhằm giải quyết những vấn đề này do đó cần phải so sánh giữa các ý kiến và thực hiện mô phỏng thường xuyên. Điều này đòi hỏi việc mô hình hóa phải được thực hiện theo một cách nào đó có thể giải quyết được vấn đề nghiêm trọng về Mức độ phù hợp (Good of fit). Các phương pháp và kỹ thuật thống kê cổ điển thường trở nên không thực tế do xuất hiện tại nhiều điểm của các phân phối đuôi nặng (heavy-tailed distributions) (thường không sử dụng đến ngay cả các công cụ tiêu chuẩn như biến thiên và tương quan) cũng như sự phụ thuộc và bất ổn tầm xa (vượt ra ngoài cả những giả định cơ bản nhất về chuỗi thời gian cổ điển). *Lĩnh vực Topo mạng* có nhiều vấn đề thống kê khác nhau. Mục đích chính của nó là để hiểu được cấu trúc kết nối của Internet. Các khái niệm lý thuyết đồ thị cùng với sự biến thiên theo thời gian và kết quả lấy mẫu là cần thiết để tạo ra tiến bộ quan trọng trong lĩnh vực này.

Phân lớp mạng là lĩnh vực suy luận cấu trúc của Internet chỉ dựa trên hoạt động của tín hiệu được gửi qua Internet. Hiểu một cách đúng đắn thì việc phân tích và mô hình hóa các lượng bất định phức tạp liên quan đến quá trình này là vô cùng quan trọng để tạo ra tiến bộ trong lĩnh vực *Phân lớp mạng.*

4.4.4 Các dòng số liệu

Các phân tích thống kê về các bộ số liệu lớn về cơ bản thường được thực hiện hàng loạt. Để thu thập và chuẩn bị các bộ số liệu này có thể mất hàng năm trời. Và để có các phân tích thống kê tương ứng với các bộ số liệu đó có thể phải cần thêm một khoảng thời gian tương tự. Tuy nhiên, Khai thác số liệu theo thời gian thực phù hợp hơn rất nhiều với các nhà Thống kê. Các tình huống như vậy có thể phát sinh, ví dụ như trong cảm biến từ xa - nơi băng thông có giới hạn nằm giữa một vệ tinh trên quỹ đạo và trạm mặt đất của nó ngăn cản việc truyền tất cả các số liệu chưa qua xử lý. Một ví dụ khác là các trang web thương mại như Hệ thống đặt chỗ của một hãng hàng không - nơi dữ liệu tổ hợp phím chi tiết khiến cho các lệnh đặt chỗ thành công hoặc thất bại không được lưu lại.

Thách thức phải đối mặt chính là tạo ra các công cụ thống kê chạy trên hầu hết các thời gian tuyến tính, nghĩa là tạo ra các công cụ có thể chạy song song với

dòng số liệu theo thời gian thực. Đối với các số liệu thống kê đơn giản ví dụ như các mô men mẫu thì không có gì khó khăn. Tuy nhiên, các công cụ này phải có khả năng thích ứng trong thời gian thực. Hơn nữa, công tác *Khai thác số liệu* hầu như sử dụng tất cả các công cụ thống kê hiện đại (ví dụ như Thuật toán Phân cụm, các sơ đồ hình cây hay hồi quy Logistic). Chuyển các công cụ thống kê có sẵn thành các công cụ mong muốn cần phải có óc tưởng tượng, sự tài tình lẫn kết hợp với các chuyên gia thuật toán trong các lĩnh vực Khoa học Toán học khác.

4.5. Các ngành khoa học Vật lý

Trong lịch sử, Thiên văn học là một trong những nguồn cảm hứng cũng như ứng dụng đầu tiên và quan trọng nhất cho các ý tưởng Thống kê. Trong thế kỷ 18, các nhà Thiên văn đã sử dụng trung bình của các phép đo các số lượng bằng nhau dưới những điều kiện như nhau. Điều này đã đưa đến phương pháp Bình phương nhỏ nhất vào đầu thế kỷ 19.

Trong những năm gần đây, ngành Thiên văn học đã mở rộng mạnh mẽ cả về quy mô và độ phức tạp của các bộ số liệu để ước lượng được các thông số trong vũ trụ của vụ nổ Big Bang thông qua phân cụm không đẳng hướng các thiên hà, phổ dao động của bức xạ phông vi sóng, vv... Một loạt các vấn đề thống kê cơ bản khác phát sinh từ Đài quan sát ảo - một liên hợp gồm nhiều Terabyte đa bước sóng các cơ sở dữ liệu khảo sát Thiên văn học.

Mặc dù hai ngành Thống kê và Thiên văn học có cùng khởi nguyên và liên quan đến nhau trong phân tích số liệu, tuy nhiên chỉ những năm gần đây giữa hai ngành này mới có những hợp tác đáng kể (ví dụ minh họa cho điều này đã được nhắc tới ở phần trên khi bàn về Thống kê nòng cốt).

Sự khác biệt đã tồn tại trong một thời gian dài giữa Thống kê học và Thiên văn học là một ví dụ minh họa chung cho các ngành khoa học Vật lý. Thống kê học hoạt động bằng sự tích lũy hiệu quả các dấu hiệu từ các nguồn thông tin nhiễu riêng biệt. Nhìn chung, ta có thể mô tả sự phổ biến của hệ phương pháp thống kê trong lịch sử như là "các lĩnh vực nhiễu đầu tiên": Thống kê hộ tịch, Kinh tế học, Nông nghiệp, Giáo dục, Tâm lý học, Y học, Di truyền học và Sinh học. Các ngành "Khoa học cứng" được gọi tên bắt nguồn từ Tỷ lệ tín hiệu trên nhiễu (signal-to-noise-ratio) gần như hoàn hảo có thể đạt tới được trong thử nghiệm cổ điển, nên có thể hiểu rằng: đây là các ngành ít phù hợp nhất để áp dụng các phương pháp luận Thống kê.

Tuy nhiên, xu hướng gần đây chính là làm mềm các ngành khoa học cứng, do đó nhu cầu về các nguyên tắc và phương pháp thống kê ngày càng tăng lên. Công nghệ hiện tại cho phép các dự án thu thập dữ liệu trở nên lớn hơn và nhiều tham vọng hơn, chẳng hạn như Dự án Đài quan sát neutrino Sudbury (để quan sát về hạt sơ cấp Neutrino) và Chương trình Thăm dò vi sóng bất đẳng hướng. Các dự án và chương trình này phải trích xuất được các thông tin quan trọng từ rất nhiều các số liệu nhiễu. (Tỷ lệ tín hiệu trên nhiễu tại Đài quan sát Sudbury chưa đến một phần triệu). Rõ ràng là các phương pháp thống kê đóng vai trò hết sức quan trọng và to lớn trong các dự án này.

Để thấy rõ vai trò đầy hứa hẹn trong tương lai của ngành Thống kê đối với các ngành khoa học Vật lý, chúng tôi đưa ra 3 ví dụ thống kê chuyên sâu ngắn gọn trong các ngành Vật lý hạt nhân, Quang phổ hóa học và Thiên văn học.

4.5.1. Khoảng tin cậy trong dò tìm hạt nhân

Tình trạng sau đây xảy ra khi nghiên cứu về các hạt lẩn tránh: một máy dò chạy trong một thời gian dài, có x điểm quan tâm; một máy dò tương tự dò tìm các hạt lẩn tránh được che chắn khỏi "tín hiệu gây nhiễu" có y điểm. Vậy giới hạn cận trên nào cho tỉ lệ thực sự của các hạt? Các vấn đề thống kê trở nên nghiêm trọng nếu y vượt quá x, khi đó ước lượng tỷ lệ khách quan sẽ bị âm. Vấn đề tính đến tiếp theo là liệu giới hạn cận trên có đủ lớn để tăng khả năng phát hiện.

Ngay cả trong những tình huống thực tế có thể liên quan đến nhiều hiệu chỉnh nền phức tạp hơn, thì vấn đề này vẫn thu hút rộng rãi sự quan tâm trong cộng đồng Vật lý. Một trong những nguồn tham khảo được trích dẫn khá nhiều là Feldman và Cousins (năm 1998). Louis Lyons, giáo sư Vật lý tại trường Đại học Oxford đã tổ chức một cuộc Hội thảo vào tháng 9 - 2003 tại Trung tâm Máy gia tốc tuyến tính Stanford để bàn về các vấn đề thống kê trong Vật lý hạt nhân, Vật lý thiên văn học và Vũ trụ học.

(www-conf.slac.stanford.edu/phystat2003/)

4.5.2 Các thí nghiệm so sánh trong Quang phổ hóa học

Giáo sư Richard Zare thuộc khoa Hóa học trường Đại học Stanford đã phát triển một loại máy quang phổ khối lượng tiên tiến có thể đồng thời tính thời gian các chuyển động bay của nhiều hạt có khối lượng. Điều này cho phép so sánh các nhóm hạt khác nhau thu được trong những điều kiện khác nhau, ví dụ như các phân tử phức tạp phát triển trong các môi trường hóa học khác nhau.

Một phổ điển hình bao gồm số lượng hạt chuyển động trong các đơn vị thời gian, khoảng 15.000 chuyển động trong 1 dải điển hình. So sánh 2 quang phổ như vậy có thể tìm ra các chuyển động khác nhau đáng kể giữa hai điều kiện. Đây chính là một hoạt động trong thử nghiệm giả thuyết đồng thời. Với 15.000 chuyển động, tính đồng thời là rất lớn. Hệ phương pháp thống kê ban đầu được phát triển để việc phân tích các vùng siêu nhỏ có thể liên quan tới các so sánh quang phổ, nhưng mối liên hệ giữa các chuyển động theo thời gian không giống như mối liên hệ giữa các gen, điều đó cho thấy một hệ phương pháp mới là vô cùng cần thiết.

4.5.3 Phân tích biến cố và Thiên văn học

Một ví dụ tiêu biểu về sự phát triển song hành của hai ngành Thống kê và Thiên văn học đó là hai ngành đã đưa ra các lý thuyết liên quan chặt chẽ với nhau nhằm giải quyết vấn đề thiếu hụt số liệu. Lĩnh vực hoạt động này trong các tài liệu thống kê được gọi là Phân tích biến cố (Survival Analysis). Có nhiều nguyên nhân khác nhau dẫn tới thiếu hụt số liệu. Các nhà Thiên văn ở Trái đất không thể quan sát rõ ràng các điểm trong không gian hoặc không thể quan sát được các điểm ở khoảng cách quá xa, điều này dẫn tới tình trạng “cắt cụt”. Bên cạnh đó, hiện tượng “loại bỏ” cũng xảy ra trong các trong các thử nghiệm Y học khi người bệnh không để lại sự việc quan trọng nào, ví dụ như tái phát hoặc tử vong trước khi thử nghiệm kết thúc. Phương pháp Lynden-Bell và Ước lượng Kaplan-Meier chính những giải pháp mà ngành Thiên văn học và ngành Thống kê đã đưa ra để giải quyết vấn đề thiếu hụt số liệu, và về cơ bản thì 2 phương pháp này là giống nhau.

Đến những năm 1980, Thống kê học và Thiên văn học mới bắt đầu quan tâm và tìm hiểu về nhau. Một loạt các hội thảo chung có tầm ảnh hưởng của hai ngành đã được tổ chức Penn State Babu và Feigelson. Các hội thảo này đã đưa đến sự hợp tác và

phát triển trong việc phân tích các số liệu Thiên văn học. Ví dụ như các bùng phát tia gamma có nguồn gốc từ ngoài dải Ngân hà và điều này đã được chứng minh thông qua phân tích biến cố trước khi các nghiên cứu Vật lý chuyên sâu có thể chứng minh nó.

4.6. Tăng cường các hoạt động hợp tác

Một đặc điểm khác biệt của ngành Thống kê đó là các giá trị trí tuệ được tạo ra trong cả công tác phát triển phương pháp luận thống kê và các hoạt động liên ngành (ví dụ như trong các ứng dụng thống kê vào các ngành Sinh học, Y học, Khoa học Xã hội, Thiên văn học, Kỹ thuật, Đường lối Chính phủ và An ninh Quốc gia). Trong đó, các hoạt động liên ngành là nhân tố thúc đẩy quan trọng để phát triển các phương pháp mới. Về cơ bản thì tất cả các nhà Thống kê đều tham gia vào cả công tác nghiên cứu phương pháp luận thống kê và công tác ứng dụng mặc dù trạng thái cân bằng giữa hai hoạt động này trong các thời điểm khác nhau là không giống nhau.

Thông qua các tác động lẫn nhau như vậy mà các nhà Thống kê phát triển các công cụ hết sức quan trọng cho những phát hiện trong các ngành khoa học và kỹ thuật khác. Bên cạnh đó, các nhà Thống kê cũng tìm ra những tương đồng giữa các vấn đề có vẻ như không liên quan trong các ngành khác nhau, từ đó góp phần hoặc tạo ra các tác động lẫn nhau giữa các lĩnh vực khoa học khác nhau.

Tuy nhiên, theo ghi nhận của Báo cáo Odom thì những gì chúng ta đạt được còn khá hạn hẹp. Báo cáo có nêu: *Cả trong các ứng dụng và các dự án liên ngành vẫn còn tồn tại những vấn đề nghiêm trọng về việc lạm dụng các mô hình thống kê cũng như về chất lượng đào tạo các phương pháp thống kê cho các nhà khoa học, kĩ sư, các nhà nghiên cứu xã hội và cả những người sử dụng khác. Do các quan sát tạo ra nhiều số liệu hơn nên các nhà Thống kê trong các nhóm nghiên cứu cần phải thường xuyên giải quyết vấn đề trên.*

Ngoài ra, các nhà Thống kê tham gia vào các hoạt động này luôn phải đối mặt với những thách thức lớn, trong đó có yêu cầu bám sát tất cả các lĩnh vực liên quan và cung cấp các phần mềm liên quan để thực hiện các phân tích thống kê. Không những thế, mặc dù các hoạt động liên ngành luôn được khuyến khích nhưng việc đánh giá các hoạt động này lại hết sức khó khăn, thậm chí có lúc gây tranh cãi.

Tóm lại, chúng tôi lưu ý rằng:

• Thực sự có rất nhiều số liệu được tạo ra trong tất cả các ngành. Điều này đang làm cho nhu cầu về các nhà Thống kê tăng lên. Cùng với việc đưa ra các phương pháp mới để hướng dẫn thiết kế thí nghiệm và phân tích số liệu, các nhà Thống kê có thể kết hợp với các nhà khoa học trong các lĩnh vực khác.

• Như đã được đề cập rõ hơn trong báo cáo đầy đủ, một thách thức về phần mềm vẫn đang tồn tại và ăn sâu vào nhiều ngành khác nhau. Đó chính là nhu cầu rất lớn về các phương pháp thống kê được đưa vào các sản phẩm phần mềm nguồn mở cũng như thiếu sự hỗ trợ để đáp ứng nhu cầu này.

• Cần phải có nguồn vốn dài hạn cho các dự án liên ngành để các nhà Thống kê có đủ điều kiện nâng cao các kiến thức khoa học cần thiết cho việc hợp tác.

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

Nguồn: A Report on the Future of Statistics http://www.biostat.jhsph.edu/~kbroche/Stat%20-%20PDF/lindsay%20et%20al%20future%20of%20stat.pdf