KHÁM PHÁ MỘT THẾ GIỚI MỚI THÔNG QUA KHOA HỌC THỐNG KÊ

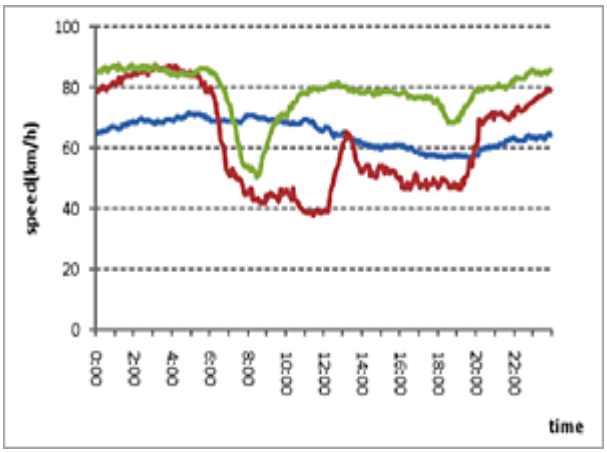
Giới thiệu

Trong những năm gần đây thế giới đã có những tiến bộ nhanh chóng trong công nghệ đo lường và hiệu suất máy tính. Kết quả là các loại dữ liệu mới đã được đo lường và lưu trữ, công nghệ phân tích thống kê phát triển làm cho việc sử dụng máy tính được hiệu quả hơn. Trong thời gian tới, những quyết định dựa trên các dữ liệu đó sẽ được ứng dụng trong nhiều lĩnh vực, từ khoa học công nghệ tiên tiến đến cuộc sống hàng ngày. Vì vậy, có thể nói vai trò quan trọng của khoa học thống kê đang được tăng lên.

Lĩnh vực của tôi nghiên cứu là khoa học thống kê, và mục tiêu nghiên cứu của tôi là phát triển các phương pháp thống kê mới cho các loại dữ liệu khác nhau. Tuy nhiên, trong bài viết này, tôi sẽ nhấn mạnh khía cạnh ứng dụng thực tế và giới thiệu các ví dụ về phân tích dữ liệu và nhiều phương pháp thử nghiệm được rút ra từ các đề tài nghiên cứu của tôi đã thực hiện.

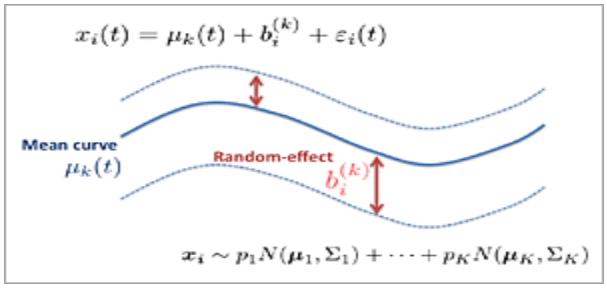
Chức năng Phân tích dữ liệu

Phần giới thiệu tôi đã nói, những tiến bộ gần đây trong công nghệ đo lường đã làm cho dữ liệu đạt đến mức có thể có được nhiều định dạng dữ liệu đầu ra khác nhau. Dữ liệu chức năng là một loại dữ liệu. Dữ liệu này được quan sát , nhìn thấy trong các định dạng của một hàm tuyến tính, hay cụ thể hơn, một đường cong (hoặc bề mặt cong). Phân tích dữ liệu là một phương pháp thống kê để xử lý dữ liệu chức năng như vậy. Ví dụ:



Mô hình tốc độ trung bình của người xe lái xe trên đường cao tốc nhất định.

Trên trục ngang thể hiện thời gian trong ngày (0:00 đến 24:00) và trục dọc cho thấy tốc độ. Phương tiện đi lại trên đường cao tốc màu xanh dương duy trì gần như cùng một tốc độ trong suốt thời gian 24 giờ. Tuy nhiên, có thể thấy rằng các đường cao tốc màu đỏ và màu xanh lá cây có thể có cùng một giá trị vào cùng một thời điểm giờ cao điểm, hoặc một thời điểm khi tốc độ xe giảm xuống. Những thông tin này được đo 5 phút mỗi lần và do đó thu được thông qua các điểm rời rạc. Tuy nhiên về bản chất chúng có thể được dùng như đo lường của một đường cong.

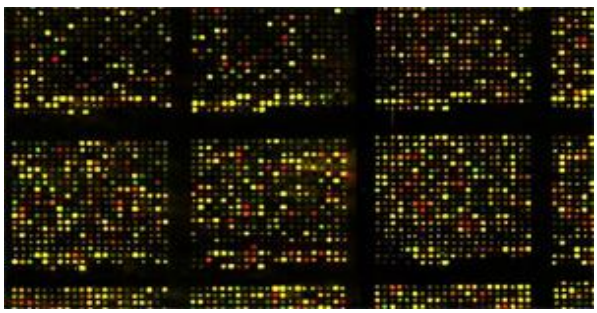


Mô hình tốc độ trung bình của người xe lái xe mô tả qua hàm số.

Khi một số lượng lớn các dữ liệu chức năng này được thu gom lại và phân loại được thực hiện cho mỗi đường cong tương tự của các mô hình thay đổi, hiện tại chúng tôi đang cố gắng phát triển một phương pháp mới, trong đó có sự kết hợp của các mô hình phức tạp và mô hình hỗn hợp gồm một mô hình GMM và một mô hình phi tham số Bayes, để tìm kiếm số lượng thích hợp của nhóm.

Bằng cách sử dụng các mô hình thống kê khác nhau, nhằm phát hiện những cấu trúc nằm sâu bên trong dữ liệu mà không thể được nhìn thấy bằng cách nhìn vào những dữ liệu phức tạp.

Nhiều phương pháp kiểm tra



Nhiều thử nghiệm là một phương pháp để xác minh đồng thời các giả thuyết.

Ví dụ, trong lĩnh vực phân tích bộ gen, nhiều thử nghiệm và một loạt các phương pháp thống kê khác nhau được sử dụng để giải quyết vấn đề như dự đoán mạng lưới gen trong đó sử dụng dữ liệu biểu hiện gen (sơ đồ) để tìm kiếm các mối quan hệ kiểm soát giữa các gen, cũng như xác định các gen liên quan đến bệnh thông qua việc sử dụng các cấu trúc đa hình từ những nucleotit đơn lẻ (SNP).

Nói chung, đặc điểm các lĩnh vực có chứa số lượng rất lớn dữ liệu thì các giả thiết đó xác nhận cùng một thời điểm và những giả thuyết này có mối quan hệ lẫn nhau tuy không rõ ràng, cấu trúc thưa thớt phân bố không đồng đều (chỉ có một số lượng rất nhỏ các quan sát có giá trị ảnh hưởng) trong một số trường hợp. Đặc điểm như vậy làm cho phương pháp thống kê truyền thống tỏ ra không hiệu quả.

Trong trường hợp này, có thể áp dụng một số phương pháp hoán vị và phương pháp bootstrap, cả hai đều là phương pháp phân tích thống kê sử dụng bằng máy tính. Tuy nhiên, về mặt lý thuyết thì đều không được chấp nhận vì cho rằng các phương pháp đó thường chỉ được sử dụng trong các lĩnh vực phân tích bộ gen. Thật vậy, một số đề tài đã chỉ ra những vấn đề về mặt lý thuyết và đã được xuất bản trong những năm gần đây. Hiện nay, chúng tôi đang làm việc để xác minh lại giá trị lý thuyết của những phương pháp thử nghiệm ở trên và để phát triển các phương pháp thử nghiệm mới. Sự phát triển của phương pháp thống kê như nhiều phương pháp kiểm tra sẽ trở thành một công cụ cho việc mở ra thế giới mới trong bộ gen được phân tích, một lĩnh vực mới mà nhiều thứ chúng ta còn chưa biết.

Thống kê trong thể thao

Phân tích thống kê cũng được sử dụng trong thể thao. Phân tích thống kê là điều cần thiết để tiến hành đánh giá khách quan trong một loạt các lĩnh vực như việc lựa chọn chiến lược phù hợp, đánh giá các vận động viên và quản lý đội hình. Tôi muốn giới thiệu như vậy vì việc sử dụng số liệu thống kê bằng cách tập trung vào môn bóng chày là một trong

những môn thể thao tiên tiến nhất trên quan điểm của phân tích thống kê về những môn thể thao.

Từ những năm 1980, phương pháp phân tích thống kê với bóng chày được gọi là SABRmetrics (một thuật ngữ được hình thành bằng cách kết hợp từ viết tắt của Hội Nghiên cứu bóng chày Mỹ và các số liệu(SABR), với một thuật ngữ đại diện cho đo lường metrics) đã được sử dụng rộng rãi tại Hoa Kỳ. Một số người có thể đã nghe nói về SABRmetrics từ Billy Beane (là một tổng giám đốc MLB người nổi tiếng với những công việc chuyển nhượng quyền thương mại ở Hoa Kỳ).

Thay vì dựa vào các giá trị truyền thống, SABRmetrics dựa vào khái niệm của việc sử dụng dữ liệu để tiến hành đánh giá khách quan về chiến lược và vận động viên. Ví dụ, số liệu thống kê như số lần đánh bóng trung bình (RBI) và số lần chạy về nhà(điểm đứng đón bóng trên sân bóng chày) được sử dụng để đánh giá tổng thể batters (người đánh bóng) tại Mỹ. Tuy nhiên, nó không phải là cách đánh giá hợp lý cho vấn đề này tại Nhật bản. Ví dụ, một cầu thủ đánh bóng chạy được một đường bóng duy nhất chính xác sẽ ghi được 01 điểm kết quả, trong khi đó thống kê về trường hợp này thì không được đánh giá mà chỉ đánh giá số lần đánh bóng và số lần chạy trên đường bóng của vận động viên.

Thay vì sử dụng số liệu thông thường truyền thống thì, SABRmetrics đánh giá vận động viên thông qua một số tiêu chí như OPS (tỷ lệ phần trăm trên cơ sở + tỷ lệ phần trăm của số truy cập ngoài cơ sở). Chỉ số này có mối tương quan cao giữa ghi bàn do lượt chạy và ghi bàn trong khi chạy có liên quan đến kết quả trận thắng và thua. Vì vậy, có thể nói rằng việc sử dụng các tiêu chí như OPS làm cho nó có thể nhìn thấy rõ sự đóng góp của một vận động viên.

Tương tự như vậy, SABRmetrics đề xuất một số lượng lớn các tiêu chí khác nhau mà mục tiêu đánh giá các vị trí như bình (là cầu thủ ném bóng chày từ gò của đỉnh sân về phía bắt bóng bắt đầu trò chơi) và người bắt bóng.v.v..

Một xu hướng hiện tại ở Hoa Kỳ là vấn đề chuyển đổi dữ liệu quỹ đạo của quả bóng ném và quỹ đạo phong trào của vận động viên. Dữ liệu này được phân tích và áp dụng thông qua phân tích thống kê đặc biệt. Tuy nhiên, việc này không khả quan khi áp dụng ở Nhật Bản vì có sự khác biệt về công nghệ so với Hoa Kỳ.

Một ví dụ về khác về vấn đề này là phương pháp bunt hy sinh. Nó đã được chứng minh về mặt thống kê phương pháp bunts hy sinh làm suy giảm khả năng ghi điểm trong một hiệp đấu. Vì lý do này, bunts hy sinh hầu như không được sử dụng trong giải bóng chày lớn tại Hoa kỳ. Tuy nhiên, nó lại được sử dụng như một chiến thuật cơ bản tại Nhật Bản. Một số người có thể nghĩ rằng do điều kiện khác nhau trong bóng chày Nhật Bản và Hoa Kỳ, nên ở Nhật Bản thường gọi là bóng nhỏ mà không phải là bóng chày. Tuy nhiên, phương pháp này cũng được xác nhận rằng việc sử dụng nó làm giảm xác suất ghi điểm trong bóng chày Nhật Bản.

Chúng tôi đang bắt đầu nghiên cứu một số phương pháp đánh giá đúng đắn về bình. Trong bóng chày chuyên nghiệp Nhật Bản, bình thường được đánh giá thông qua sử dụng số liệu thống kê như ERA liên quan đến kết quả trận thắng và thua. Tuy nhiên, thống kê này là không thích hợp để đánh giá bình. Ví dụ, khi một bình bắt đầu ném tốt, thì có thể

không kiếm được một chiến thắng nếu người đánh bóng không hỗ trợ bình bằng cách cho điểm chạy. Hơn nữa, bình sẽ bị mất quyền của mình như là bình thưởng nếu chạy được mà điểm lại tính cho bình cứu trợ. Ngược lại, có những trường hợp một bình mang lại một số lượng lớn điểm do nhiều lần chạy có thể nhận được một chiến thắng thông qua sự hỗ trợ của người đánh bóng đội mình. Vì vậy, chúng ta xem xét việc sử dụng một hỗ trợ là chiến thắng trung bình (phương pháp Lỗ). Phương pháp này xác định số lượng chiến thắng (hoặc lỗ) là một bình bắt đầu sẽ có nếu người đánh bóng đội mình và bình cứu trợ thực hiện trên một mức độ tiêu chuẩn. Nói cách khác, chỉ số này là phù hợp có thể đánh giá bình bằng cách loại bỏ tất cả các yếu tố khác hơn so với bình mình.

Hiện tại tôi đang nói tới môn bóng chày tuy nhiên, trong tương lai, tôi cũng muốn áp dụng phân tích thống kê cho môn bóng đá.Tôi chắc chắn rằng hầu hết mọi người vẫn còn có những kỷ niệm đẹp về đội bóng Nhật Bản trong năm World Cup 2010. Trong World Cup gần đây nhất, tổng khoảng cách chạy của mỗi vận động viên trong một trận đấu đã được tính toán. Ở Nhật Bản, các cầu thủ như Endo và Honda chạy khoảng 11 km trong một trận theo thống kê đã được thông báo trên truyền thông. Khoảng cách chạy được tính từ dữ liệu quỹ đạo có được thông qua cách làm mô phỏng theo sự chuyển động của cầu thủ và quả bóng bằng máy ảnh. Tuy nhiên, dữ liệu quỹ đạo này có nhiều ứng dụng khác mà không chỉ đơn giản là tính toán khoảng cách chạy. Chúng tôi muốn sử dụng dữ liệu giá trị này để giúp thúc đẩy các môn thể thao bóng đá. Hơn nữa, bằng cách phát triển mô hình thống kê mới, chúng tôi cũng muốn tiến hành nghiên cứu với mục tiêu phát triển khoa học thống kê.

Kết luận

Trong bài viết này, tôi giới thiệu các ý tưởng nghiên cứu của tôi đã thực hiện cũng như các ứng dụng thực tế đi kèm. Khoa học thống kê được sử dụng trong một loạt các lĩnh vực là một phương pháp để phát hiện ra kiến thức mới. Để đáp ứng nhu cầu đó, các nhà nghiên cứu khoa học thống kê đã làm việc hàng ngày để tìm ra các lý thuyết và phương pháp mới.

Ra quyết định thông qua đánh giá thống kê là không có giới hạn trong những công nghệ khoa học tiên tiến hoặc lĩnh vực chuyên môn mô tả. Trên thực tế, hồ sơ công việc, được ưu tiên trong cuộc sống hàng ngày. Ví dụ, một loạt các hồ sơ lưu trữ đồ thị được sử dụng bởi phương tiện truyền thông như truyền hình và báo chí.Tuy nhiên, số liệu thống kê hoặc các đồ thị thường không chính xác. Một xã hội phát triển và có những công dân thông minh, có khả năng giải thích một cách chính xác thông tin, điều quan trọng là có được tài liệu thống kê. Tầm quan trọng này được thể hiện trong các hướng dẫn chương trình giáo dục mới của các trường tiểu học, trường trung học cơ sở và trung học. Các chương trình đào tạo mới cho toán học và đặt trọng tâm khoa học về thống kê, xác suất. Chúng tôi đang làm việc nỗ lực để truyền bá và phát triển thống kê thông qua một loạt các sự kiện và các chiến dịch, cũng như bằng cách cung cấp một số lượng lớn các tài liệu giáo dục. Tôi hy vọng rằng bài viết này sẽ tạo sự quan tâm trong khoa học thống kê của các độc giả.

Fumitake Sakaori, Tiến sĩ (Khoa học)
Phó Giáo sư Khoa học Thống kê, Khoa Khoa học và Kỹ thuật, Đại học Chuo, Nhật Bản.

Công Hoan dịch:
Nguồn tin http://www.yomiuri.co.jp /adv/chuo/dy/research/20101028.htm