NGHIÊN CỨU VÀ ỨNG DỤNG MỘT PHƯƠNG PHÁP THỐNG KÊ NHIỀU CHIỀU TRONG Y HỌC

NGUYỄN QUANG HÒA
NGUYỄN TUẤN HOA
NGUYỄN VĂN HÙNG
Viện KH tính toán và Điều khiển

I – MỞ ĐẦU

Một trong những phương hướng nghiên cứu của ngành thống kê – phân tích các bảng số liệu phân lớp chéo (hay bảng tiếp liên) đã được phát triển từ rất sớm. Những công trình của Pearson và Yule xuất bản năm 1900 đã đặt nền móng cho các phương pháp phân tích, về các hệ số đánh giá thống kê cũng như các phương pháp kiểm định sự phù hợp của mô hình với bảng số liệu được thu thập. Tuy nhiên hơn 30 năm sau đó những nghiên cứu vẫn chưa đi quá việc phân tích các bảng tiếp liên hai chiều. Điều này được giải thích do sự chậm phát triển của công cụ tính số cũng như độ phức tạp lý thuyết mà nhà thống kê gặp phải khi phân tích những bảng tiếp liên nhiều chiều hơn. Kể từ năm 1935 khi Barlett cho xuất bản công trình về kiểm định đối với các bảng tiếp liên ba chiều dạng $2 \times 2 \times 2$ thì các công trình nghiên cứu phân tích các bảng tiếp liên nhiều chiều đã chiếm một vị trí hứa hẹn trong các tạp chí thống kê thế giới. Hai mươi năm trở lại đây, sự phát triển như vũ bão của kỹ thuật máy tính điện tử đã cung cấp cho các nhà thống kê một công cụ hữu hiệu để phân tích những mô hình toán học mô tả tương tác giữa các biến có các thể hiện bằng số ở bảng tiếp liên nhiều chiều. Người ta đưa ra ngày càng nhiều mô hình toán, nhiều thuật toán để đánh giá tốt hơn tương tác thực sự của các biến. Số lượng các công trình nghiên cứu về vấn đề này tăng vọt trong những năm gần đây.

Tuy nhiên ở Việt Nam, vấn đề này hầu như còn rất mới mẻ. Trong quá trình làm việc với các cơ sở thực tế – chủ yếu thuộc bộ Y tế – chúng tôi nhận thấy việc áp dụng toán thống kê thường chỉ là sử dụng những hệ số tương quan, kiểm định hai biến, tính trung bình và các khoảng tin cậy. Trường hợp có nhiều biến tác động người ta phân thành những cặp xét độc lập với các tương tác khác. Điều này dẫn đến các kết luận sai lầm nếu như các biến không thỏa mãn một điều kiện độc lập như định lý 3. 1 [1] đã nêu. Mục đích bài này của chúng tôi là phân tích các bảng tiếp liên ba chiều và kết quả ứng dụng trong các bài toán thực tế mà chúng tôi đã giải quyết.

II – BA BÀI TOÁN PHÂN TÍCH BẢNG TIẾP LIÊN BA CHIỀU

Bài toán 1. (của Vụ Dược chính – Bộ Y tế).

Để nghiên cứu tác dụng của thuốc nam và tân dược trong việc điều trị một chứng bệnh X, người ta quan sát một mẫu N người mắc bệnh đó từ thời điểm t_0. Sau một khoảng thời gian t người ta đếm số người khỏi và chưa khỏi bệnh trong bốn nhóm:

– Nhóm những người không dùng thuốc
– Nhóm những người dùng tân dược theo liều lượng và không dùng thuốc nam.
– Nhóm những người dùng thuốc nam theo liều lượng và không dùng tân dược.
– Nhóm những người dùng cả tân dược và thuốc nam theo liều lượng.

Bài toán này dẫn đến việc phân tích một bảng tiếp liên ba chiều cỡ $2 \times 2 \times 2$, qua đó rút ra những kết luận thích đáng để điều trị, chú trọng đến việc xét tương tác giữa thuốc nam và tân dược.

Bài toán 2. (của bệnh viện Bạch Mai Hà Nội).

Nghiên cứu mối liên quan giữa bệnh vành tim (coronary heart disease) với tỉ lệ Cholesterol trong huyết thanh (mg/100cc) và huyết áp (mm Hg), người ta quan sát một mẫu N cá thể bị bệnh vành tim và không bị bệnh vành tim và đếm số những người đó trong các nhóm sau:

- Lượng Cholesterol trong huyết thanh ở mức C_i ($i = 1, 2, 3, 4$) trong đó:

  C_1 : dưới 200mg/100cc
  C_2 : 200 → 219 mg/100cc
  C_3 : 220 → 259 mg/100cc
  C_4 : Trên 260 mg/100cc

- Huyết áp ở mức H_i ($i = 1, 2, 3, 4$) trong đó:

  H_1 : dưới 127mm Hg
  H_2 : 127 → 146 mm Hg
  H_3 : 147 → 166 mm Hg
  H_4 : Trên 166 mm Hg

Bài toán trên dẫn đến việc phân tích bảng tiếp liên 3 chiều cỡ $2 \times 4 \times 4$.

Bài toán 3. (của trường Đại học Y Hà Nội).

Nghiên cứu mối liên quan giữa bệnh viêm phế quản mãn với sự ô nhiễm môi trường (chủ yếu là nồng độ SO_2 và nồng độ bụi trong không khí) người ta đã điều tra ở một số địa điểm khác nhau thuộc nội ngoại thành Hà Nội. Một mẫu N người ($N = 6389$) được đếm theo số người có hoặc không có biểu hiện i của bệnh lý ($i = 1, 2, \ldots 12$) ở các vùng sau:

- Nồng độ SO_2 và nồng độ bụi ở mức cho phép (SO_2^+, $bụi^+$)

- Nồng độ SO_2 ở mức cho phép, nồng độ bụi quá mức cho phép (SO_2^+, $Bụi^-$)

- Nồng độ bụi ở mức cho phép, nồng độ SO_2 quá mức cho phép (SO_2^-, $bụi^+$)

- Nồng độ SO_2 và nồng độ bụi đều quá mức cho phép (SO_2^-, $Bụi^-$)

Bài toán này dẫn đến việc phân tích một bảng tiếp liên ba chiều cỡ $2 \times 2 \times 2$.

Trong phần IV, chúng tôi sẽ đưa ra các kết quả và phân tích của bài toán 3. Số liệu điều tra do GS_1. PTS Đào Ngọc Phong (đại học Y Hà Nội) cung cấp.

III – CƠ SỞ TOÁN HỌC CỦA PHƯƠNG PHÁP PHÂN TÍCH

Trong phần này chúng tôi giới hạn trong việc phân tích các bảng tiếp liên 3 chiều. Về mặt lý thuyết có thể áp dụng cho trường hợp 4, 5 hoặc nhiều chiều hơn. Tuy vậy trong thực tế, khi xét những trường hợp đó, chúng ta phải xây dựng lại các công thức ước lượng cũng như cân nhắc khi sử dụng để tránh các sai số lớn.

1. Ba biến có các thể hiện số trong bảng tiếp liên ba chiều được ký hiệu v_1, v_2, v_3 có tập giá trị hữu hạn trong đó 1 hoặc 2 biến là biến trả lời (response variable) và các biến còn lại là biến giải thích (explanatory variable), các biến trả lời coi như cố định và các biến giải thích là điều khiển được. Tuy nhiên sự phân biệt giữa hai loại biến này không phải lúc nào cũng rõ ràng như vậy. Trong ba ví dụ ở trên các biến trả lời lần lượt là:

1) Người bệnh (khỏi, không khỏi),

2) Người bệnh (có bệnh, không có bệnh),

3) Người được quan sát (có biểu hiện bệnh, không có biểu hiện bệnh).

Những biến còn lại là những biến giải thích.

Các trường hợp có những biến có giá trị liên tục chúng tôi không đề cập ở đây vì chúng có liên quan đến những phương pháp thống kê khác; ví dụ trường hợp biến giải thích rời rạc, biến trả lời liên tục là liên quan đến phương pháp phân tích phương sai.

2. Các giá trị đếm được xắp xếp vào các bảng tiếp liên ba chiều cỡ $I \times J \times K$ với các quan sát x_{ijk} ($i \in \overline{1, I}$, $j \in \overline{1, J}$, $k \in \overline{1, K}$). Ví dụ các quan sát ở bài toán 3. Có thể được sắp xếp vào bảng sau:

Cho: $v_1 = \begin{cases} 1: \text{nếu có dấu hiệu bệnh lý } i \\ 2: \text{nếu không có dấu hiệu bệnh lý } i \end{cases}$

$v_2 = \begin{cases} 1: \text{nếu nồng độ } SO_2 \text{ cho phép} \\ 2: \text{nếu nồng độ } SO_2 \text{ quá mức cho phép} \end{cases}$

$v_3 = \begin{cases} 1: \text{nếu nồng độ bụi cho phép} \\ 2: \text{nếu nồng độ bụi quá mức cho phép} \end{cases}$

Bảng 1

	$v_3 = 1$		$v_2 = 2$	
	$v_3 = 1$	$v_3 = 2$	$v_3 = 1$	$v_3 = 2$
$v_1 = 1$	x_{111}	x_{112}	x_{121}	x_{122}
$v_1 = 2$	x_{211}	x_{212}	x_{221}	x_{222}

3. Ta xét các mô hình log tuyến tính và dùng kiểm định χ^2 để xác nhận xem mô hình đó có phù hợp với bảng số liệu quan sát hay không. Vì ta chỉ xét các mô hình phân cấp nên trường hợp bảng tiếp liên ba chiều sẽ có 9 mô hình sau:

Mô hình 1 (mô hình độc lập):

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)}$$

Mô hình 2:

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)}$$

Mô hình 3:

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{13(ik)}$$

Mô hình 4:

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{23(jk)}$$

Mô hình 5:

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)}$$

Mô hình 6:

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{23(jk)}$$

Mô hình 7:

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{13(ik)} + u_{23(jk)}$$

Mô hình 8:

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)}$$

Mô hình 9 (mô hình đầy đủ):

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ij)} + u_{23(jk)} + u_{123(ijk)}$$

Trong đó:

m_{ijk} là kỳ vọng của ô (ijk) ($i \in \overline{1, I}$, $j \in \overline{1, J}$, $k \in \overline{1, K}$)

u là trung bình lớn: $u = \dfrac{1}{IJK} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \log m_{ijk}$.

$u + u_{1(i)}$ là trung bình của các log kỳ vọng ở mức i của biểu thứ nhất

$$u + u_{1(i)} = \frac{1}{J.K} \sum_{j=1}^{J} \sum_{k=1}^{K} \log m_{ijk}$$

v.v...

Nếu hai biến v_1 và v_2 độc lập thì $u_{12(ij)} = 0$ (xem [1]) và do vậy $u_{123(ijk)} = 0$. Vậy sự tồn tại của các u trong mô hình nêu lên sự tương tác được tồn tại giữa các biến tương ứng. Ký hiệu: $\hat{m}_{ijk}$ là kỳ vọng ước lượng của ô thứ (ijk)

$$X_{i++} := \sum_{j=1}^{J} \sum_{k=1}^{K} X_{ijk}; \quad X_{+j+} := \sum_{i=1}^{I} \sum_{k=1}^{K} X_{ijk}, \quad X_{++k} := \sum_{i=1}^{I} \sum_{j=1}^{J} X_{ijk}$$

$$X_{ij+} := \sum_{k=1}^{K} X_{ijk}, \quad X_{i+k} := \sum_{j=1}^{J} X_{ijk}, \quad X_{+jk} := \sum_{i=1}^{I} X_{ijk}$$

$$N := X_{+++} := \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} X_{ijk}$$

ta có thể ước lượng kỳ vọng của các ô theo công thức sau (lần lượt cho các mô hình 1, 2, 3, 4, 5, 6, 7):

$$\hat{m}_{ijk} = \frac{X_{i++} . X_{+j+} . X_{++k}}{N^2}, \qquad \hat{m}_{ijk} = \frac{X_{ij+} . X_{++k}}{N}$$

$$\hat{m}_{ijk} = \frac{X_{i+k} . X_{+j+}}{N}, \qquad \hat{m}_{ijk} = \frac{X_{+j+} . X_{i++}}{N}$$

$$\hat{m}_{ijk} = \frac{X_{i+k} . X_{ij+}}{X_{i++}}, \qquad \hat{m}_{ijk} = \frac{X_{ij+} . X_{+jk}}{X_{+j+}}$$

$$\hat{m}_{ijk} = \frac{X_{i+k} . X_{+jk}}{X_{++k}}$$

Trong đó: $i \in \overline{1, I}$, $j \in \overline{1, J}$, $k \in \overline{1, K}$

với mô hình 8 ta dùng thuật toán lặp để ước lượng kỳ vọng các ô.

Cho $\hat{m}^{(0)}_{ijk} = 1$ $(i \in \overline{1, I}, \ j \in \overline{1, J}, \ k \in \overline{1, K})$

bước 1:
$$\hat{m}^{(3v+1)}_{ijk} = \frac{X_{ij+}}{\hat{m}^{(3v)}_{ij+}} . \hat{m}^{(3v)}_{ijk}$$

bước 2:
$$\hat{m}^{(3v+2)}_{ijk} = \frac{X_{i+k}}{\hat{m}^{(3v+1)}_{i+k}} . \hat{m}^{(3v+1)}_{ijk}$$

bước 3:
$$\hat{m}^{(3v+3)}_{ijk} = \frac{X_{+jk}}{\hat{m}^{(3v+2)}_{+jk}} . \hat{m}^{(3v+2)}_{ijk}$$

Sau đó quay lại bước 1 với $v = v + 1$

Những công thức đã nêu ở trên có thể chứng minh từ giả thiết sự độc lập hay phụ thuộc giữa các biến. Ví dụ đối với mô hình 1:

$$m_{ijk} = N . P(v_1 = i, \ v_2 = j, \ v_3 = k)$$

Do tính độc lập của các biến:

$$P(v_1 = i, \ v_2 = j, \ v_3 = k) = P(v_1 = i) . P(v_2 = j) . P(v_3 = k)$$

nên ta có thể ước lượng:

$$\hat{m}_{ijk} = N \cdot \frac{X_{i++}}{N} \cdot \frac{X_{+j+}}{N} \cdot \frac{X_{++k}}{N} = \frac{X_{i++} \cdot X_{+j+} \cdot X_{++k}}{N^2}$$

Sau khi đã có các kỳ vọng ước lượng, ta tính χ^2 theo công thức:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(X_{ijk} - \hat{m}_{ijk})^2}{\hat{m}_{ijk}}$$

và so sánh với mức 0.05 để xem có loại bỏ mô hình hay không.

Bậc tự do của mô hình 1 là 4, của các mô hình 2, 3, 4 là 3, các mô hình 5, 6, 7 là 2 và của mô hình 8 là 1, nếu tất cả các mô hình từ 1 đến 8 đều bị loại bỏ thì ta chấp nhận mô hình 9 (mô hình đầy đủ).

IV – CÁC KẾT QUẢ TÍNH TOÁN CỦA BÀI TOÁN II.3.

1. Chúng tôi xét một hai dấu hiệu của người bị viêm phế quản mãn lấy số liệu cho bảng tiếp liên ba chiều. Với một dấu hiệu i, chúng tôi được 1 bảng như bảng 1 trong đó x_{ijk} là các giá trị cụ thể.

Các dấu hiệu đó (theo bảng mã của phiếu điều tra – trường đại học Y Hà Nội như sau

Dấu hiệu 1: Ho thường xuyên lúc ngủ dậy sáng, mùa đông.

Dấu hiệu 2: Mùa đông ho 5 – 6 cơn 1 ngày, 5 ngày một tuần.

Dấu hiệu 3: Ho tất cả mọi ngày liên tục trong 3 tháng một năm.

Dấu hiệu 4: Ho 3 tháng liên tục trong 2 năm liên tục.

Dấu hiệu 14: Không đi bộ được do tim, phổi.

Dấu hiệu 15: Mùa đông khó chịu, khó thở khi đi vội, trèo dốc.

Dấu hiệu 19: Thỉnh thoảng bị thở rít lên.

Dấu hiệu 20: Thở rít trong hầu hết các ngày và đêm.

Dấu hiệu 21: Có cơn khó thở kịch phát.

Dấu hiệu 23: Có cơn hen kịch phát.

Dấu hiệu 30: Thường xuyên bị tắc mũi và chảy nước mũi.

Dấu hiệu 32: Như 30 liên tục trong 2 năm.

Tổng số phiếu điều tra đã làm là 6389. Các tiêu chuẩn về nồng độ bụi, SO_2 căn cứ theo tiêu chuẩn quốc tế của OMS.

Các vùng được điều tra là (thuộc Hà Nội):

$v_2 = 1$, $v_3 = 1$: Tứ Hiệp, Liên Minh (Thanh Trì)

$v_2 = 1$, $v_3 = 2$: Ngọc Hà

$v_2 = 2$, $v_3 = 1$: Thị trấn Văn Điển

$v_2 = 2$, $v_3 = 2$: Vĩnh Quỳnh, phân lân Văn Điển, Tam Thiệp, Phó Đức Chính, Quan Thánh, Hàng Bún, Phạm Hồng Thái, Nhà máy điện Yên Phụ.

Kết quả tính toán:

Mô hình 9 phù hợp cho các dấu hiệu: 1, 2, 3, 4.

Mô hình 8 phù hợp cho các dấu hiệu: 13, 21, 26.

Mô hình 6 phù hợp cho các dấu hiệu: 15, 19, 20, 32.

Nhận xét:

– Trong tất cả các mô hình được chấp nhận đều có u_{23}, điều này phù hợp vì nồng độ SO_2 và bụi của các vùng không thay đổi với các dấu hiệu bệnh lý khác nhau.

– Không có với 1 dấu hiệu bệnh lý nào các mô hình đơn giản được chấp nhận mà các mô hình phức tạp hơn lại bị loại bỏ. Điều này chứng tỏ công thức ước lượng là tốt.

– Tốc độ hội tụ của thuật toán dùng cho mô hình 6 là tương đối cao.

2. Chương trình tính:

Chương trình viết bằng ngôn ngữ MBASIC được thực hiện trên máy vi tính của viện. Các ký hiệu:

Số liệu: X(I, J, K), I = 1, 2, J = 1, 2, K = 1, 2.

Các biến:

$B12(I, J) = X_{ij}$. $B13(I, K) = X_{i+k}$, $B23(JK) = X_{+jk}$,

$B1(I) = X_{i++}$, $B2(J) = X_{+j+}$, $B3(K) = X_{++k}$

Listing chương trình và Copy của các bảng kết quả hiện đang được các tác giả lưu giữ.

TÀI LIỆU THAM KHẢO

1. Stephen E. Fienberg, the Analysis of Cross-classified Categorical Data, the MIT Press, England (1977)

2. Ku, H.H, and Kullback, S. Loglinear models in Contigency table analysis, Amer – Statist. 28, 115–122 (1974).

ABSTRACT

A Study of a Multi – Dimensional Statistical method and its Application in biometry

In this paper we study a method of analysing three dimensional contigency tables and apply this method in biometry. Concretely, we detect the connection between the dust measure, the SO_2 measure and the lung diseases, the work has been done for the contract with a researching group of the Medical University of Hanoi. The results are suitable to those have been got from other medical researchs and have been well appreciated.

M — PHỦ TỐI TIỂU VÀ CÁC HỆ...

(Tiếp theo trang 17)

TÀI LIỆU THAM KHẢO

1. Lucchesi C. L., Osborn S. L., Canditdate keys for relation. J. of Computer and System Sciences, 17/1978, 270–297.

2. Demetrovics. J., On the equivalence of canditdate key's with Sperner systems. Acta Cybernetica Szeged 4/1979, 247–252.

3. Ho Thuan, Some romarks on the algorithm of Lucchesi and Osborn, MTA SZTAKI, Kozlemenyek 35/1986.

4. Ho Thuan and Le Van Bao, Some results about keys for relational schemas, Acta Cybernetica Szeged, 7/1985, 99–113.

ABSTRACT

M – minimal cover and Sperner systems with application to the key finding problem for relation Scheme

Pham The Que

In this paper, we investigate the properties of M–minimal covers when a finite set H and the Sperner system on H are given. Specially, we establish the necessary and sufficient condition for which two sperner systems are the set of all representative sets of each other. This means that from the given set of all keys for a relation scheme, we can construct its set of all representative sets and conversely, from the set all representative sets for the set of keys we can determine the set of keys for the relation scheme.

The set of keys for the relation scheme is just the set of all representative sets for the set of all representative sets for the set of keys.