УДК 519.21

Ганжа Р.О., аспірант.
Лебєдєв Є.О., д.ф.-м.н., проф.

Про нормалізовану відстань між контрольною та експериментальною вибірками

Київський національний університет імені Тараса Шевченка, 83000, м. Київ, пр-т. Глушкова, 4д, e-mail: ganzha.roman.alex@gmail.com
Київський національний університет імені Тараса Шевченка, 83000, м. Київ, пр-т. Глушкова, 4д, e-mail: leb@unicyb.kiev.ua

R. O. Ganzha, graduate student.
E. O. Lebedev, doctor of physical and mathematical sciences, professor.

About normalized distance between case and control samples

Taras Shevchenko National University of Kyiv, 83000, Kyiv, Glushkova st., 4d, e-mail: ganzha.roman.alex@gmail.com
Taras Shevchenko National University of Kyiv, 83000, Kyiv, Glushkova st., 4d, e-mail: leb@unicyb.kiev.ua

В статті розглядається метод оптимізації процесу діагностики захворювань через побудову нормалізованої відстані між контрольною та експериментальною вибірками. Запропоновано алгоритм побудови нормалізованої відстані, в основі якого лежить ідея особливого використання техніки перевірки гіпотез про значущість відмінностей між контрольною та експериментальною вибірками з використанням відстані Махаланобіса та розподілу Снедекора-Фішера, для якого отримано явні формули в термінах елементарних функцій. Ці формули лягли в основу алгоритму рекурентного типу для знаходження значень розподілу. Описано використання результатів при знаходженні нормалізованої відстані між вибірками та пошуку набору факторів, оптимального для діагностики.
Ключові слова: Розподіл Фішера, нормалізована відстань, бета-розподіл, відстань Махаланобіса.

In the process of certain disease diagnosis often there is a need to find out how close case and control samples are in a particular set of their features. While selection of the most characteristic features from a complete set is an optimization problem, the task of construction of normalized distance between case and control samples must be solved prior to solving of the optimization problem.

This paper serves to reveal one approach of building an algorithm to calculate normalized distance. The basis of the proposed algorithm for constructing normalized distance is the idea of using a special technique for testing of hypotheses about the significance of differences between case and control samples using Mahalanobis distance and F-distribution. Explicit formulas in terms of elementary functions have been obtained for a distribution of interest. Based on these formulas recurrent type algorithms for finding of distribution values have been proposed. Using of results which were obtained by applying this algorithm were showed for finding the normalized distance between two samples, as well as for finding a set of factors which can be considered as optimal in diagnosis determination.
Key-words: F-distribution, normalized distance, beta distribution, Mahalanobis distance.

Статтю представив д.т.н., проф. Заславський В.А.

В процесі медичного діагностування часто виникає необхідність побудови оцінки близькості двох вибірок по деякому спільному набору їх ознак. В той час, як вибір найбільш характеристичних ознак з їх повного набору є задачею оптимізації, задача побудови нормалізованої відстані між контрольною та експериментальною вибірками має бути розв’язана до розв’язку оптимізаційної задачі.

В основі запропонованого алгоритму побудови нормалізованої відстані лежить ідея особливого використання техніки перевірки гіпотез про значущість відмінностей між контрольною та експериментальною вибірками з використанням відстані Махаланобіса та розподілу Снедекора-Фішера.

Побудуємо нормалізовану відстань, використовуючи підхід з роботи [2]. Нехай $X^{(1)}$ та $X^{(2)}$ – дві вибірки, які складають дані вимірів факторів $F_1, F_2, ..., F_m$, в контрольній та експериментальній вибірках відповідно, $\rho(X^{(1)}, X^{(2)})$ - статистика критерію (відстань між вибірками). Через $F(x)$ позначимо функцію розподілу $\rho(X^{(1)}, X^{(2)})$. Нормалізованою відстанню $\rho_N(X^{(1)}, X^{(2)})$ між вибірками $X^{(1)}$ і $X^{(2)}$ будемо називати

$$\rho_N(X^{(1)}, X^{(2)}) = F(\rho(X^{(1)}, X^{(2)})). \quad (1)$$

Якщо $\tilde{X}^{(1)}$, $\tilde{X}^{(2)}$ - реалізації вибірок $X^{(1)}$, $X^{(2)}$, $\tilde{\rho}_N = \rho_N(\tilde{X}^{(1)}, \tilde{X}^{(2)})$, $\tilde{\rho} = \rho(\tilde{X}^{(1)}, \tilde{X}^{(2)})$, то нормалізована відстань $\tilde{\rho}_N$ є вірогідністю того, що при повторенні експерименту будуть зустрічатись $\rho(X^{(1)}, X^{(2)})$ не більші за $\tilde{\rho}$.

Оцінка відмінності між вибірками, обрахована по формулі (1), має дві важливі властивості: 1) $\tilde{\rho}_N$ стандартизована та знаходиться в діапазоні від 0 до 1; 2) чим ближче $\tilde{\rho}_N$ до 1, тим менш вірогідна нульова гіпотеза про відсутність ефекту і тим з меншим рівнем значущості (з більшою правдоподібністю) відхиляється нульова гіпотеза згідно традиційній схемі перевірки статистичних гіпотез.

Підрахунок значень розподілу Фішера

Нагадаємо, що функція розподілу Фішера має вигляд $F_{m_1,m_2}(x) = \int_0^x f_{m_1,m_2}(t)dt$, [1]

де $f_{m_1,m_2}(t)$ – щільність, що задається наступною формулою:

$$f_{m_1,m_2}(t) = \begin{cases} \dfrac{\Gamma(\frac{m_1+m_2}{2}) m_1^{\frac{m_1}{2}} m_2^{\frac{m_2}{2}}}{\Gamma(\frac{m_1}{2})\Gamma(\frac{m_2}{2})} t^{\frac{m_1}{2}-1} (m_2 + m_1 t)^{\frac{m_2+m_1}{2}}, t > 0 \\ 0, t \le 0 \end{cases}, \quad (2)$$

де $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx, z > 0$ – функція Ейлера другого роду (гамма-функція).

Якщо ξ_1, ξ_2 - незалежні випадкові величини, що мають розподіл χ^2 з m_1, m_2 ступенями свободи, то випадкова величина $\xi = \dfrac{\xi_1 / m_1}{\xi_2 / m_2}$ (3) має розподіл Фішера (F – розподіл) з (m_1, m_2) ступенями свободи. Визначимо розподіл:

$$\mathrm{B}_{a,b}(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^x t^{a-1}(1-t)^{b-1} dt = \quad (4)$$

$$= \frac{1}{\mathrm{B}_{a,b}(x)} \int_0^x t^{a-1}(1-t)^{b-1} dt, 0 \le x \le 1, a, b > 0,$$

де $B_{a,b}(\cdot)$ - бета-функція розподілу.

Розподіл (4) пов'язаний з розподілом Фішера наступним чином.

Лема 1. Якщо $F_{m_1,m_2}(x)$ – функція розподілу Фішера, а $\mathrm{B}_{a,b}(x)$ - визначається за формулою (4), то

$$F_{m_1,m_2}(x) = 1 - \mathrm{B}_{\frac{m_2}{2},\frac{m_1}{2}}(y), \quad (5)$$

де $y = \dfrac{m_2}{m_2 + m_1 x}$.

Також корисними є такі властивості:

Лема 2. (Властивість симетрії).

Якщо $F_{m_1,m_2}(x)$ – функція розподілу Фішера, то

$$F_{m_1,m_2}(x) = 1 - F_{m_2,m_1}(\frac{1}{x}), \quad (6)$$

Лема 3. Якщо $\mathrm{B}_{a,b}(x)$ визначається формулою (4), то

$$\mathrm{B}(a,b) = \frac{1}{\mathrm{B}(a,b)} \frac{x^a (1-x)^{b-1}}{a} + \mathrm{B}_{a+1,b+1}(x), \quad (7)$$

де $\mathrm{B}(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1} dx$ - бета-функція.

На основі вище наведених властивостей можна довести такий результат:

Теорема 1. Якщо $F_{m_1,m_2}(x)$ – функція розподілу Фішера, а $\mathrm{B}_{a,b}(x)$ - визначається за формулою (4), то

$$F_{m_1,m_2}(x) = 1 - \mathrm{B}_{\frac{m_2}{2},\frac{m_1}{2}}(y), \quad (8)$$

де:

1. При $m_1 = 2k, m_2 = 2p$:

$$\mathrm{B}_{\frac{m_2}{2},\frac{m_1}{2}}(y) = \mathrm{B}_{p,k}(y) = \\ = \sum_{i=0}^{k-1} y^{p+1}(1-y)^{k-1-i}\frac{(p+k-1)!}{(p+i)!\backslash(k-1-i)!}, \quad (9)$$

$$y = \frac{m_2}{m_2 + m_1 x};$$

2. При $m_1 = 2k, m_2 = 2p+1$:

$$\mathrm{B}_{\frac{m_2}{2},\frac{m_1}{2}}(y) = \mathrm{B}_{p+\frac{1}{2},k}(y) = \\ = \sum_{i=0}^{k-1} y^{p+\frac{1}{2}+i}(1-y)^{k-1-i}\frac{\prod_{\nu=1}^{p+k}\left[\nu-\frac{1}{2}\right]}{\prod_{\mu=1}^{p+k+1}\left[\mu-\frac{1}{2}\right](k-1-i)!}; \quad (10)$$

3. При $m_1 = 2k+1, m_2 = 2p+1$:

$$\mathrm{B}_{\frac{m_2}{2},\frac{m_1}{2}}(y) = \mathrm{B}_{p+\frac{1}{2},k+\frac{1}{2}}(y) = \\ = \frac{1}{\pi}\sum_{i=0}^{k-1} y^{p+\frac{1}{2}+i}(1-y)^{k-\frac{1}{2}-i}\frac{(p+k)!}{\prod_{\nu=1}^{p-1}\left[\nu-\frac{1}{2}\right]\prod_{\mu=1}^{p+i+1}\left[\mu-\frac{1}{2}\right]} + \quad (11)$$

$$+1-\frac{2}{\pi}arctg(z) - -\frac{2}{\pi}\sum_{i=1}^{p+k}\frac{\frac{1}{2i}\frac{z}{(1+x^2)}}{\prod_{j=1}^{i}(1-\frac{1}{2j})};$$

$$z = (y^{-1}-1)^{\frac{1}{2}}.$$

Доведення теореми 1 отримано на основі алгоритму побудови функції розподілу Фішера з початковою ідентифікацією параметрів за одним з чотирьох випадків:

- $m_1 = 2k, m_2 = 2p$
- $m_1 = 2k, m_2 = 2p+1$
- $m_1 = 2k+1, m_2 = 2p$
- $m_1 = 2k+1, m_2 = 2p+1.$

1. Якщо $m_1 = 2k$, $m_2 = 2p$, то

$$y = \frac{m_2}{m_2 + m_1 x},\ B_{p,k}(y) = \sum_{i=0}^{k-1}\exp\{A_i\},$$

$$A_0 = p\ln y + (k-1)\ln(1-y) + \sum_{\nu=1}^{k-1}\ln(1+p/\nu),$$

$$A_{i+1} = A_i + \ln[\frac{y}{1-y}\cdot\frac{k-1-i}{p+1+i}],\quad i = 0,1,...,k-2,$$

$$F_{m_1,m_2}(x) = 1 - B_{p,k}(y).$$

2. У випадку, якщо $m_1 = 2k$, $m_2 = 2p+1$, то

$$y = \frac{m_2}{m_2 + m_1 x},\ B_{p+1/2,k}(y) = \sum_{i=0}^{k-1}\exp\{B_i\},$$

$$B_0 = (p+1/2)\ln y + (k-1)\ln(1-y) + \\ + \sum_{\nu=1}^{k-1}\ln(1+\frac{p+1/2}{\nu}),$$

$$B_{i+1} = B_i + \ln[\frac{y}{1-y}\cdot\frac{k-1-i}{p+3/2+i}],\quad i = 0,1,...,k-2,$$

$$F_{m_1,m_2}(x) = 1 - B_{p+1/2,k}(y).$$

3. Для випадку, коли $m_1 = 2k+1, m_2 = 2p$ обчислення аналогічні попередньому пункту з аргументом $\frac{1}{x}$.

В результаті: $F_{m_1,m_2}(x) = 1 - F_{2p,2k+1}(1/x)$.

4. Якщо $m_1 = 2k+1$, $m_2 = 2p+1$, то

$$y = \frac{m_2}{m_2 + m_1 x},\ \mathrm{z} = (\mathrm{y}^{-1}-1)^{1/2},$$

$$B_{p+1/2,k+1/2}(y) = \sum_{i=0}^{k-1}\exp\{D_i\} + 1 - \frac{2}{\pi}arctg\ z - \frac{2}{\pi}\sum_{i=1}^{p+k}C_i,$$

$$D_0 = -\ln\pi + (p+1/2)\ln y + (k-1/2)\ln(1-y) + \\ + \sum_{\nu=1}^{p+1}\ln(\frac{\nu}{\nu-1/2}) + \sum_{\mu=1}^{k-1}\ln(\frac{p+1+\mu}{\mu-1/2}) - \ln(k-1/2),$$

$$D_{i+1} = D_i + \ln[\frac{y}{1-y}\cdot\frac{k-i-1/2}{p+i+3/2}],\ i = 0,1,...,k-2,$$

$$C_1 = \frac{z}{1+z^2},\ C_{i+1} = \frac{2i}{2i+1}\cdot\frac{1}{1+z^2}C_i,$$

$$i = 1,2,...,\ p+k-1,$$

$$F_{m_1,m_2}(x) = 1 - B_{p+1/2,k+1/2}(y).$$

Застосуємо алгоритм до обробки статистичних даних.

Знаходження нормованої відстані

Нехай експериментальна та контрольна вибірки представлені в виді матриць $X^{(1)} = \left\|x_{ij}^{(1)}\right\|$, $X^{(2)} = \left\|x_{ij}^{(2)}\right\|$ розміру $m_1 \times n$ та $m_2 \times n$, $x_{ij}^{(1)}$ - значення фактора F_j в i-ому вимірі експериментальної вибірки, $x_{ij}^{(2)}$ - значення фактора F_j в i-ому вимірі контрольної вибірки, m_1, m_2 - об'єм експериментальної та контрольної груп відповідно, n - число факторів. Враховуючи істотну багатовимірність вибірок, для побудови

статистики $\rho(X^{(1)}, X^{(2)})$ використовуємо відстань Махаланобіса

$D^2(X^{(1)}, X^{(2)}) = \sum_{i,j=1}^{n} \Delta x_i \Delta x_j s_{ij}^{-1}$, де

$\Delta x_i = \bar{x}_i^{(1)} - \bar{x}_i^{(2)}$, $\bar{x}_i^{(1)} = \frac{1}{m_1} \sum_{k=1}^{m_1} x_{ki}^{(1)}$,

$\bar{x}_i^{(2)} = \frac{1}{m_2} \sum_{k=1}^{m_2} x_{ki}^{(2)}$ - середнє значення фактора F_i в першій та другій вибірках відповідно,

$S^{-1} = \left\| s_{ij}^{-1} \right\|_1^n$ - матриця, обернена матриці

$S = \left\| s_{ij} \right\|_1^n$, $s_{ij} = (s_{ij}^{(1)} + s_{ij}^{(2)})/(m_1 + m_2 - 2)$,

$s_{ij}^{(p)} = \sum_{k=1}^{m_p} (x_{ki}^{(p)} - \bar{x}_i^{(p)})(x_{kj}^{(p)} - \bar{x}_j^{(p)})$, $p = 1,2$.

Відомо, що для вибірок $X^{(1)}$, $X^{(2)}$ з багатовимірних нормальних сукупностей

$$\rho(X^{(1)}, X^{(2)}) = \frac{m_1 + m_2 - n - 1}{n} \cdot \frac{m_1 m_2}{(m_1 + m_2)(m_1 + m_2 - 2)} D^2(X^{(1)}, X^{(2)}), \quad (12)$$

має функцію розподілу Фішера $F_{\nu_1,\nu_2}(x)$ з $\nu_1 = n$, $\nu_2 = m_1 + m_2 - n - 1$ ступенями свободи. Використовуючи цей факт, можна побудувати нормалізовану відстань Фішера між вибірками $X^{(1)}$, $X^{(2)}$ виду

$$\rho_N(X^{(1)}, X^{(2)}) = F_{\nu_1,\nu_2}(D^2(X^{(1)}, X^{(2)})) \quad (13)$$

Застосування

В роботі [3] описана вище рекурентна схема для $F_{\nu_1,\nu_2}(x)$ була застосована для аналізу біоелектричної активності мозку в контрольній групі тварин та групі тварин з осередковою структурно-функціональною патологією центральної нервової системи. Об'єми першої та другої груп були рівними 55. Множину факторів біоелектричної активності мозку було розбито на п'ять груп: (F_1, F_2, F_3),(F_4, F_5, F_6),(F_7, F_8, F_9), (F_{10}, F_{11}, F_{12}),(F_{13}, F_{14}, F_{15}) відповідно діапазонам [4-6 Гц] - θ_1-ритми, [6-8 Гц] - θ_2-ритми, [8-12 Гц] - α-ритми, [12-20 Гц] - β_1-ритми, [20-40 Гц] - β_2-ритми. Першим фактором в кожній групі $(F_{1+3i}, i = 0,1,2,3,4)$ була сумарна потужність (мкв^2 / гц), другим фактором $(F_{2+3i}, i = 0,1,2,3,4)$ була максимальна потужність (мкв), третім фактором $(F_{3+3i}, i = 0,1,2,3,4)$ – середня частота в діапазоні. Дані збирались при трьох режимах проведення експерименту: фон, світло, після світла. Оптимальні набори наведені в наступній таблиці відповідно:

$\rho_{N,max}$	Код набору
0,99989	(100010100000010)
0,98757	(000000000111100)
0,99931	(111110111000100)

В стопчику "Код набору" "1" вказує на присутність фактору в наборі, а "0" - на його відсутність. Таким чином код (100010100000010) відповідає набору факторів (F_1, F_5, F_7, F_{14}).

Список використаних джерел

1. Айвазян С.А. Прикладная Статистика: Основы моделирования и первичная обработка данных: справочное издание / С.А.Айвазян, И.С.Енюков, Л.Д.Мешалкин.– Москва :Финансы и статистика, 1983.-472 с.
2. Копанев В.А. Метод вероятностной оценки токсического эффект / В.А.Копанев Э.Х.Гинзбург Н.В.Семенова. – Новосибирск : Наука, 1988. - 128 с.
3. E.Lebedev, B.Shurunov, A.Cherchenko Exact formulae for basic statistical distributions and their applications, Proceedings of the 10-th International Symposium on Applied stochastic models and data analysis, 2001, Universite de Technologie de Compiegne, France, vol. 2, p. 660-666.

References

1. IVAZYAN, S., ENYUKOV, I., MESHALKIN, L. (1983) *Applied Statistics.* Moscow. Finance and Statistics. P. 472.
2. KOPANEV, V., GINZBURG, E., SEMENOVA, N. (1988) Probability method of toxic effect evaluation. Novosibirsk. Science. Siberian department. P.128.
3. LEBEDEV, E., SHURUNOV, B., CHERCHENKO, A. (2001) Exact formulae for basic statistical distributions and their applications. Proceedings of the 10-th International Symposium on Applied stochastic models and data analysis". Universite de Technologie de Compiegne, France, vol. 2, P. 660-666.

Надійшла до редколегії 27.02.2015