Журнал обчислювальної та прикладної математики **2012, №4(110)** Journal of Computational & Applied Mathematics

УДК 519.234.3+519.234.7 519.71

НЕПАРАМЕТРИЧНИЙ КРИТЕРІЙ КЛАСИФІКАЦІЇ З ДВОСТОРОННІМИ ДОВІРЧИМИ ІНТЕРВАЛАМИ

К. М. Голубєва

Резюме. Ю. І. Петунін зробив значний внесок до статистичного розпізнавання образів. В статті наводиться основні результати, які стосуються p-статистики, отримані в рамках його математичної школи. Була показана перевага критерію, що ґрунтується на p-статистиках, запропонована нова міра близькості між вибіркою та групою вибірок з однієї генеральної сукупності та побудуваний критерій класифікації для розпізнавання раку молочної залози на основі нової статистики.

Вступ

Перші роботи, присвячені непараметричним методам оцінювання, з'явилися ще на початку XX ст. В своїх працях Спірмен та Кендалл запропонували методи, які ґрунтуються на коефіцієнтах рангової кореляції [1, 2]. То були лише перші спроби, становлення ж класичної непараметричної статистики відбулося у 30-х рр. XX ст. Саме тоді з'явилися роботи Колмогорова А. М. та Смірнова М. В., в яких розглядалися різниці між емпіричними та теоретичними функціями розподілу [3]. Критерії типу Колмогорова-Смірнова ще й досі часто використовуються в різних дослідженнях. Після другої світової війни непараметрична статистика почала розвиватися швидкими темпами. Вагомий внесок своїми працями зробили Уілкоксон Ф. [4], Манн Х. Б. та Уітні Д. Р. [5]. Запропоновані ними критерії перевірки гіпотез почали узагальнюватися на багатовимірні випадки. В цьому напрямку працювали Крускал У. та Уолліс А. [6], Фрідман М. [7]. Насправді на сьогоднішній день існує велика кількість непараметричних критеріїв, але робота в цій галузі не припиняється. В першу чергу це пов'язано із спробами вирішити проблему малих вибірок, тобто побудувати непараметричний критерій, який буде задовільно працювати із вибірками малих об'ємів. По-друге, ще досі є потреба в критерії з високою потужністю, який можна застосовувати не лише до теоретичних даних, а й до вибірок, отриманих в результаті проведення деякого експерименту, оскільки застосування існуючих критеріїв не завжди дає результат задовільної точності. Особливо важливим є значення похибок при аналізі біологічних та медичних даних, тому що в цьому випадку від отриманих значень залежить здоров'я, а інколи і життя людини. Ґрунтовний огляд існуючих непараметричних критеріїв показав, що майже всі вони розглядають гіпотезу про рівність математичного сподівання або дисперсії, звідки випливає,

що є потреба в більш універсальному критерії, який не просто буде чутливим до зсуву математичного сподівання чи дисперсії, а буде розглядати узагальнену гіпотезу про рівність функцій розподілу. Отже, вищезазначені дослідження є актуальними і дуже важливими. Петунін Юрій Іванович зробив значний внесок у цьому напрямку. Своїми науковими дослідженнями він збагатив і розвинув теоретичні основи статистики [8], зокрема його внесок у розділ непараметричної статистики має велику цінність для практичних задач в біології та медицині, які він розв'язував [9].

1. Непараметричний критерій еквівалентності генеральних сукупностей, що ґрунтується на усередненні p-статистик

Для розв'язання задач перевірки гіпотез використовується p-статистика. Наведемо основні відомості про неї [10]. Припустимо, що абсолютно неперервні функції розподілу $F_1(u)$ та $F_2(u)$ генеральних сукупностей G_1 та G_2 відповідно співпадають. Нехай $x=(x_1,...,x_n)\in G_1$ та $y=(y_1,...,y_n)\in G_2$, $x_{(1)} \le ... \le x_{(n)}$, $y_{(1)} \le ... \le y_{(n)}$ — порядкові статистики. Позначимо через $A_{ij}^{(k,n)}$, $k = 1,2,...,n$, випадкову подію, що має вигляд $A_{ij}^{(k,n)} = \left\{y_k \in \left(x_{(i)}, x_{(j)}\right)\right\}$. В такому випадку

$$p_{ij}^{(n)} = P(A_{ij}^{(k,n)}) = P\left(y_k \in \left(x_{(i)}, x_{(j)}\right)\right) = \frac{j-i}{n+1}.$$

Покладемо

$$p_{ij}^{(n,1)} = \frac{h_{ij}^{(n)} n + 4,5 - 3\sqrt{h_{ij}^{(n)}(1-h_{ij}^{(n)})n + 2,25}}{n+9},$$

$$p_{ij}^{(n,2)} = \frac{h_{ij}^{(n)} n + 4,5 + 3\sqrt{h_{ij}^{(n)}(1-h_{ij}^{(n)})n + 2,25}}{n+9},$$

де $h_{ij}^{(n)}$ — частота появи події $A_{ij}^{(k,n)}$ в n випробуваннях.

Позначимо через $N = n\,(n-1)\,/2$ кількість всіх довірчих інтервалів $I_{ij}^{(n)} = \left(p_{ij}^{(n,1)}, p_{ij}^{(n,2)}\right)$, L — кількість подій $B_{ij}^{(n)} = \left\{p_{ij}^{(n)} \in I_{ij}^{(n)}\right\}$.

Покладемо

$$h^{(N)} = \rho\,(x,y) = \frac{L}{N}.$$

Довірчий інтервал $I^{(N)} = \left(p^{(N,1)}, p^{(N,2)}\right)$ для імовірності $p(B_{ij}^{(n)})$ обчислюється за формулами

$$p^{(N,1)} = \frac{h^{(N)} N + 4,5 - 3\sqrt{h^{(N)}(1-h^{(N)})N + 2,25}}{N+9},$$

$$p^{(N,2)} = \frac{h^{(N)} N + 4,5 + 3\sqrt{h^{(N)}(1-h^{(N)})N + 2,25}}{N+9}.$$

Статистика $h^{(N)}$ є мірою близькості $\rho\,(x,y)$ між вибірками x та y, вона має назву p-статистика. Довірчі інтервали $I_{ij}^{(n)} = \left(p_{ij}^{(n,1)}, p_{ij}^{(n,2)}\right)$ та

$I^{(N)} = \left(p^{(N,1)}, p^{(N,2)}\right)$ будемо називати інтервалами, побудованими за правилом 3σ. Описана міра близькості використовується для розв'язання задачі про дві вибірки.

Був проведений обчислювальний машинний експеримент для порівняння p-статистики з іншими непараметричними критеріями, які використовуються для перевірки нульової гіпотези про те, що вибірки належать одній генеральній сукупності [11]. Розглядалися п'ять випадків: 1) порівнювались вибірки з однієї генеральної сукупності; 2) вибірки з генеральних сукупностей з різними дисперсіями; 3) вибірки з генеральних сукупностей з різним математичними сподіваннями; 4) вибірки з різних генеральних сукупностей із інтервалами розташування вибіркових даних, що перетинаються; 5) вибірки з різних генеральних сукупностей із співпадаючими інтервалами розташування вибіркових даних.

В поставленому експерименті оцінка похибки першого роду дорівнює нулю для всіх критеріїв, що розглядалися. Оцінка середньої похибки другого роду для розглянутих п'яти випадків для критерію, що ґрунтується на p-статистиках, дорівнює нулю, тоді як для критерію Колмогорова-Смірнова складає 20%, для критерію знаків – 44%, для Уілкоксона — 38%, а для Манна-Уітні — 40%. Максимальна довжина інтервалу, в якому коливається значення p-статистик, дорівнює 0,235. Для критерію Колмогорова-Смірнова максимальна довжина інтервалу коливання його p-значення — 0,845, для Манна-Уітні — 0,84, для критерію знаків — 0,864, для Уілкоксона — 0,91. Таким чином p-статистика в проведеному обчислювальному експерименті виявила себе більш стійкою та потужною в порівнянні з іншими непараметричними критеріями, які розглядалися. Тому саме вона використовувалася в подальших дослідженнях.

2. Усереднена p-статистика та асимптотичний рівень значущості її довірчого інтервалу

Для роботи з групою вибірок було зроблене узагальнення p-статистики [12]. Припустимо, що абсолютно неперервні функції розподілу $F_1(u)$ та $F_2(u)$ генеральних сукупностей G_1 та G_2 відповідно співпадають. Нехай $x = (x_1, ..., x_n) \in G_1$ та $y^{(m)} = \left(y_1^{(m)}, ..., y_n^{(m)}\right) \in G_2$, $m = 1, 2, ..., M$; $x_{(1)} \leq ... \leq x_{(n)}$, $y_{(1)}^{(1)} \leq ... \leq y_{(n)}^{(1)}$, $\ldots$, $y_{(1)}^{(m)} \leq ... \leq y_{(n)}^{(m)}$ — порядкові статистики. Тоді для перевірки гіпотези еквівалентності генеральних сукупностей буде використовуватись середня міра близькості між вибіркою x та вибірками $y^{(m)}$, $m = \overline{1, M}$:

$$h = \frac{1}{M} \sum_{m=1}^{M} h^{(n,m)},$$

де $h^{(n,m)} = \rho\left(x, y^{(m)}\right)$.

Міра близькості h називається усередненою p-статистикою. За довірчий інтервал для усередненої p-статистики h будемо використовувати інтервал

$I^{(M)} = (p^{(M,1)}, p^{(M,2)}) = (h - 2s_h, h + 2s_h)$. Для обґрунтування коректності використання усередненої p-статистики був оцінений рівень значущості довірчих інтервалів $I^{(M)}$. Це було зроблено у вигляді теореми, доведення якої наводиться в роботі [13].

Теорема 1. *Асимптотичний рівень значущості інтервалу*

$$I^{(M)} = \left(p^{(M,1)}, p^{(M,2)}\right)$$

не перевищує 0,05.

Для перевірки працездатності запропонованого критерію був проведений обчислювальний експеримент, в якому розглядалися п'ять випадків вищезазначені у попередньому обчислювальному експерименті. Критерій, що ґрунтується на усередненій p-статистиці, порівнювався з критеріями Крускала-Уолліса та Фрідмана. Усереднена p-статистика виявила себе стійкою та потужною: вона жодного разу не припустилася помилки в проведеному обчислювальному експерименті, а максимальні межі її коливання складають 0,09. Для порівняння: середня оцінка похибки першого роду для критеріїв Крускала-Уолліса та Фрідмана дорівнює 10%, середня оцінка похибки другого роду для розглянутих п'яти випадків складає по 40%, а максимальні межі коливання p-значень 0,949 та 0,981 відповідно.

3. Критерій класифікації для розпізнавання раку молочної залози

Працездатність запропонованих непараметричних методів досліджувалася на експериментальних даних. Розв'язання практичних біологічних задач відбувалося в рамках співпраці із Інститутом експериментальної патології, онкології та радіобіології ім. Р. Є. Кавецького НАН України.

Матеріалом для дослідження були препарати букального епітелію слизової оболонки порожнини рота пацієнтів хворих на рак молочної залози (G_1 — 38 пацієнтів), фіброаденоматоз (G_2 — 44 пацієнта) та практично здорових (G_3 — 33 людини). В кожного пацієнта розглядалося близько 50 інтерфазних ядер. Яскравість ядра була зареєстрована та цифрові зображення цитологічних препаратів букального епітелію отримували за допомогою світлового мікроскопу Olympus BX41, при чому досліджувалася лише зелена компонента зображення. Такі зображення мають назву сканограми; методика одержання сканограм докладно описана в роботах [14].

Для кожної сканограми був проведений розрахунок морфо- та денситометричних показників ядер: площа ядра; мінімальна яскравість зображення ядра; максимальна яскравість зображення ядра; середня яскравість зображення ядра; середньоквадратичне відхилення розподілу яскравості зображення ядра; медіана розподілу яскравості зображення ядра; коефіцієнт асиметрії розподілу яскравості зображення ядра; енергія розподілу яскравості зображення ядра; ентропія за Шенноном розподілу яскравості зображення ядра; ексцес розподілу яскравості зображення ядра; верхній

квартіль розподілу яскравості зображення ядра; нижній квартіль розподілу яскравості зображення ядра; 5%-процентіль розподілу яскравості зображення ядра; 95%-процентіль розподілу яскравості зображення ядра. Для усіх показників при їх визначенні маються на увазі вибіркові числові характеристики випадкової величини.

Використовуючи показники сканограм пацієнтів та міри близькості показників пацієнта P, захворювання якого ми бажаємо визначити, до груп G_1, G_2, G_3 знаходимо $\rho(P, G_1)$, $\rho(P, G_2)$, $\rho(P, G_3)$. Крім того, використовуючи правило $3s_1$, обчислюються довірчі інтервали $(\alpha_k(P), \beta_k(P))$, $(k = 1, 2, 3)$ для усереднених p-статистик $\rho(P, G_k)$, $(k = 1, 2, 3)$.

Критерій класифікації для розпізнавання раку формулюється таким чином:

1) якщо $\rho(P, G_1) < \rho(P, G_2)$, а $\rho(P, G_1)$ не належить інтервалу $(\alpha_2(P), \beta_2(P))$ або $\rho(P, G_2)$ не належить інтервалу $(\alpha_1(P), \beta_1(P))$, то пацієнта P відносимо до групи хворих на РМЗ;

2) якщо $\rho(P, G_1) > \rho(P, G_2)$, а $\rho(P, G_2)$ не належить інтервалу $(\alpha_1(P), \beta_1(P))$ або $\rho(P, G_1)$ не належить інтервалу $(\alpha_2(P), \beta_2(P))$, то пацієнта P відносимо до групи хворих на ФАМ;

3) в протилежному випадку ніякого рішення не приймається.

Група G_3 застосовується для контролю. При підрахунку міри близькості для пацієнта $P \in G_1$ до групи G_1 або $P \in G_2$ до групи G_2 використовується добре відомий метод: пацієнт P вилучається з його групи, а далі підраховується міра близькості між P та групою пацієнтів, що залишаються після його вилучення [15].

Описаний метод розпізнавання раку молочної залози має високий степінь точності: похибка при розпізнаванні раку дорівнює $7,9\%$, а фіброаденоматозу $18,2\%$; і на відміну від більшості існуючих методів є абсолютно неінвазивним.

Висновки

Наукова робота Ю. І. Петуніна та його школи має суттєве значення для статистичного розпізнавання образів. Отримані результати розширили і збагатили теорію перевірки статистичних гіпотез, і більше того мають незаперечне прикладне значення, яке підтверджується їх використанням в біології та медицині: побудований критерій для перевірки гіпотези про приналежність двох вибірок до однієї генеральної сукупності та побудуваний критерій класифікації для розпізнавання раку молочної залози.

Література

1. Spearman C. The proof and measurement of association between two things / C. Spearman // Amer. J. Psychol. — 1904. — V. 15. — P. 72–101.
2. Kendall M. A New Measure of Rank Correlation / M. Kendall // Biometrika. — 1938. — V. 30. — P. 81–89.
3. Смирнов Н. В. Оценка расхождения между эмпирическими кривыми распределения в двух независимых выборках / Н. В. Смирнов // Бюллетень Московского университета. — 1939. — Т. 2, № 2. — С. 3–14.

4. Wilcoxon F. Individual Comparisons by Ranking Methods / F. Wilcoxon// Biometrics. — 1945. — Bul. 1. — P. 80–83.
5. Mann H. B. On a test of whether one of two random variables is stochastically larger than the other / H. B. Mann, D. R. Whitney // Ann. Math. Statist. — 1947. — V. 18. — P. 50–60.
6. Kruskal W. H. Use of ranks in one criterion variance analysis / W. H. Kruskal, A. Wallis // J. Amer. Statist. Assoc. — 1952. — V. 47. — P. 583–621.
7. Frideman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance / M. Frideman // J. Amer. Statist. Assoc. — 1937. — V. 32. — P. 675–701.
8. Высочанский Д. Ф. Обоснование правила 3σ для одномодальных распределений / Д. Ф. Высочанский, Ю. И. Петунин // Теор. вероятн. и матем. статистика. — 1980. — № 21. — С. 25–36.
9. Петунин Ю. И. Приложение теории случайных процессов в биологии и медицине / Ю. И. Петунин. — К. : Наук. думка, 1981. — 319 с.
10. Клюшин Д. А. Доказательная медицина / Д. А. Клюшин, Ю. И. Петунин. — М. : И. Д. Вильямс, 2008. — 320 с.
11. Голубєва К.М. Непараметричні методи перевірки гіпотези про однорідність двох вибірок / К. М. Голубєва, Д. А. Клюшин, Ю. І. Петунін // Вісник Київського національного університету імені Тараса Шевченка Серія: кібернетика. — 2011. — № 11. — С. 4–9.
12. Andrushkiw A. Diagnosis of Breast Cancer Using Averaged Proximity Measure between Samples / R. Andrushkiw, E. Golubeva, D. Klyushin, Yu. Petunin, N. Boroday // BIOCOMP'11'. — 2011. — P. 387–393.
13. Голубева Е. Н. Непараметрический критерий эквивалентности генеральных совокупностей, основанный на усредненной мере близости между выборками / Е. Н. Голубева, Д. А. Клюшин, Ю. И. Петунин // Проблемы управления и информатики. — 2010. — № 5. — С. 89–94.
14. Andrushkiw R. I. Computer-Aided Cytogenetic Method of Cancer Diagnosis / R. I. Andrushkiw, N. V. Boroday, D. A. Klyushin, Yu. I. Petunin. — New-York: Nova-Science Publisher, 2007. — 303 p.
15. Бородай Н. В. Комп'ютерний цитогенетичний метод діагностики раку молочної залози, що заснований на усередненні p-статистик / Н. В. Бородай, К. М. Голубєва, Д. А. Клюшин, Л. Ю. Ковальова, Ю. Лозовська, Ю. І. Петунін // Журнал обчисл. та прикл. матем. — 2009. — № 3 (99). — С. 24–29.

Факультет кібернетики, Київський національный університет імені Тараса Шевченка, вул. Володимирська, 64, Київ, 01601, Україна.

Надійшла 19.11.12