УДК 519.234.3+519.234.7 519.71

ЕФЕКТИВНИЙ АЛГОРИТМ ОБЧИСЛЕННЯ P-СТАТИСТИКИ

І. П. Сіренко, Д. А. Клюшин

Резюме. Ю. І. Петунін зробив значний внесок до статистичного розпізнавання образів. Один із інструментів, які були ним винайдені для розв'язання багатьох задач, є p-статистика. В статті пропонується новий ефективний спосіб її обчислення.

Вступ

Вперше p-статистику як міру близькості між двома вибірками Ю. І. Петунін запропонував у 1984 році в статті [1], що була присвячена розв'язанню задачі розпізнавання раку шлунку. Ця статистика виявила добрі практичні властивості, але довгий час не мала теоретичного підгрунтя. Її теоретичні властивості (рівень значущості, межа чутливості, довірчий інтервал) були досліджені в 2003 році в статті [2]. Втім, обчислення p-статистики для двох вибірок розміру n має складність $O\left(n^3\right)$. Для вибірок великого розміру ця проблема може стати важливою, тому спроби оптимізувати обчислення досі є актуальними.

1. P-статистика — міра близькості між вибірками

Нехай $x=(x_1,\ldots,x_n)\in G_1$ та $y=(y_1,\ldots,y_m)\in G_2$, $x_{(1)}\leq\cdots\leq x_{(n)}$, $y_{(1)}\leq\cdots\leq y_{(m)}$ — порядкові статистики, де генеральні сукупності G_1 та G_2 мають абсолютно неперервні функції розподілу $F_1(u)$ та $F_2(u)$ відповідно. Позначимо через $A_{ij}^{(k,n)}$, $k=1,2,...,n$, випадкову подію

$$A_{ij}^{(k,n)}=\left\{y_k\in\left(x_{(i)},x_{(j)}\right)\right\},\ i=1,\ldots,n-1,\ j=2,\ldots,n,\ k=1,\ldots,m.$$

Якщо функції розподілу $F_1(u)$ та $F_2(u)$ є однаковими, то ймовірність

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

Покладемо

$$p_{ij}^{(n,1)}=\frac{h_{ij}^{(n)}m+\frac{g^2}{2}-g\sqrt{h_{ij}^{(n)}(1-h_{ij}^{(n)})m+\frac{g^2}{4}}}{m+g^2}, \tag{2}$$

$$p_{ij}^{(n,1)}=\frac{h_{ij}^{(n)}m+\frac{g^2}{2}+g\sqrt{h_{ij}^{(n)}(1-h_{ij}^{(n)})m+\frac{g^2}{4}}}{m+g^2}, \tag{3}$$

де $h_{ij}^{(n)}$ — частота появи події $A_{ij}^{(k,n)}$ в m випробуваннях, а g – корінь рівняння $\Phi\left(g\right)=1-2\beta$. Тут $\Phi\left(u\right)$ — функція нормованого нормального розподілу,

β — рівень значущості. Якщо n мале, то у відповідності з правилом "3σ" $g = 3$.

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

$$h^{(N)} = \rho(x, y) = \frac{L}{N}. \tag{4}$$

Статистика $h^{(N)}$ є мірою близькості $\rho(x, y)$ між вибірками x та y, вона має назву p-статистика.

2. Ідея алгоритму обчислення p-статистики для двох вибірок

Алгоритм обчислення p-статистики для двох вибірок досить очевидний. Організовуються два вкладених цикла, в яких перебираються інтервали $I_{ij}^{(n)}$ і перевіряються умови $p_{ij}^{(n)} \in I_{ij}^{(n)}$. Проте для обчислення меж інтервалів $I_{ij}^{(n)}$ необхідно мати значення $h_{ij}^{(n)}$. Обчислення цього значення напряму вимагає ще одного цикла для перебору вибірки y. Обчислювальна складність такого алгоритму буде $O(n^2 \times m)$.

Подія $\{y_k \in x_{(i)}, x_{(j)}\}$ еквівалентна виконанню умови $\{x_{(i)} < y_{(k)} \leq x_{(j)}\}$. Отже, кількість елементів вибірки $y_{(k)}$, що задовольняють цю умову, дорівнює $h_{ij}^{(n)}$.

Відштовхуючись від ідеї, яка закладена у формулювання критерію Вілсона [4], та враховуючи упорядкованість вибірок x та y за зростанням, обчислимо новий масив

$$z_i = \max_{k=\overline{1,m}}(k : y_{(k)} \leq x_{(i)}),\ i = 1, \ldots, n,$$

тобто z_i — це число елементів вибірки y, що не більші i-того елемента вибірки x. Тоді

$$h_{ij}^{(n)} = (z_j - z_i)/m. \tag{5}$$

Отже, якщо вирахувати значення масиву z_i наперед, можна зекономити на обчисленнях — ми фактично виносимо обчислення $h_{ij}^{(n)}$ з-під вкладених циклів. Тому обчислювальна складність алгоритму стане $O(n^2 + m)$.

Крім того, необхідно відмітити, що в загальному випадку p-статистика не є симетричною, тобто $p(x, y) \neq p(y, x)$. Тому у випадку, коли потрібна міра близькості з властивостями відстані, необхідно використовувати симетричний варіант p-статистики

$$p'(x, y) = (p(x, y) + p(y, x))/2.$$

3. Асимптотичний рівень значущості довірчого інтервалу для p-статистики

Теорема 1. *Якщо* $x = (x_1, ..., x_n) \in G_1$ *і* $y = (y_1, ..., y_n) \in G_2$ *та* $F_1(u) \neq F_2(u)$, *то асимптотичний рівень значущості інтервалу* $I^{(n)} = \left(p^{(1)}, p^{(2)}\right)$ *при* $g = 3$ *не перевищує 0,05.*

Доведення. Враховуючи результати роботи [2], розглянемо довірчі інтервали

$$I_{ij}^{(n)} = \left(p_{ij}^{(1)},\ p_{ij}^{(2)}\right),\ i < j,$$

що побудовані за частотою $h_{ij}^{(n)}$ для випадкової події $\left\{y_k \in \left(x_{(i)}, x_{(j)}\right)\right\}$ за допомогою правила 3σ. Оскільки $F_{G_1}(u) \neq F_{G_2}(u)$, ймовірність $P\left(B_{ij}^{(n)}\right)$ випадкової події $B_{ij}^{(n)} = \left\{p_{ij}^{(n)} = \frac{j-i}{n+1} \in I_{ij}^{(n)}\right\}$ може залежати від індексів i та j. Введемо випадкові величини

$$\delta_{ij}^{(n)} = \begin{cases} 1, \text{якщо } p_{ij}^{(n)} \in I_{ij}, \\ 0, \text{якщо } p_{ij}^{(n)} \notin I_{ij}, \end{cases} \quad z^{(n)} = \frac{1}{N}\sum_{i<j} \delta_{ij}^{(n)} = h^{(n)},\ N = \frac{n\,(n-1)}{2}.$$

Тоді

$$p^{(n)} = E\left(z^{(n)}\right) = \frac{1}{N}\sum_{i<j} p\left(B_{ij}^{(n)}\right). \tag{6}$$

Пронумеруємо випадкові величини $\delta_{ij}^{(n)}$ довільним чином і позначимо отриману множину як $X = (X_1, X_2, ..., X_N)$, а сумісну функцію розподілу елементів вибірки X — як $F\,(u_1, u_2, ..., u_N)$. Виберемо елементи з множини X без повертання. Багатовимірна випадкова величина $\gamma^{(N)} = (\gamma_1, \gamma_2, ..., \gamma_N)$, яка отримана таким чином, називається індукованою вибіркою, отриманою за урновою моделлю. Як показано в роботі [3], функція сумісного розподілу $F_{\gamma^{(N)}}\,(u_1, u_2, ..., u_N)$ індукованої вибірки $\gamma^{(N)}$ має вигляд

$$F_{\gamma^{(N)}}\,(u_1, u_2, ..., u_N) = \frac{1}{N!}\sum_{(i_1, i_2, ..., i_N)} F\,(u_{i_1}, u_{i_2}, ..., u_{i_N}),$$

де G_N — група перестановок чисел $(1, 2, ..., N)$. Ця функція є перестановочною, отже елементи $\gamma_k,\ k = 1, 2, ..., N$, індукованої вибірки є симетрично залежними випадковими величинами з маргінальною функцією розподілу

$$F_{\gamma_k}\,(u) = \tfrac{1}{N}\left[F_1\,(u) + ... + F_N\,(u)\right].$$

Звідки випливає, що маргінальні функції розподілу $F_{\gamma_k}\,(u)$ є однаковими і мають вигляд

$$F_{\gamma_k}\,(u) = \begin{cases} 0, \text{якщо } u \leq 0, \\ q^{(n)} = 1 - p^{(n)}, \text{якщо } 0 < u < 1, \\ p^{(n)}, \text{якщо } u \geq 1, \end{cases}$$

де $p^{(n)} = E\,(z)$ визначається за формулою (6). Позначимо як B випадкову подію, що полягає у тому, що ймовірність $p_{ij}^{(n)} = \frac{j-1}{n+1}$ належить довірчому інтервалу $I_{ij}^{(n)}$ при випадковому виборі цього інтервала. Очевидно, що подія B відбувається тоді і лише тоді, коли $\delta_{ij}^{(n)} = \gamma_l = 1,\ k = 1, 2, ..., N$, тому індуковану вибірку $\gamma^{(N)}$ можна інтерпретувати як схему випробувань. В

результаті подія B може відбутися з імовірністю $p(B) = p^{(n)}$. Легко бачити, що частота h_B цієї події збігається з $h_1 = \rho(x, y)$, тому [2]

$$h_1 = \frac{1}{N}\sum_{k=1}^{N}\gamma_k,\ E(h_1) = \frac{1}{N}\sum_{k=1}^{N}E(\gamma_k) = p^{(n)},$$

$$D(h_1) = \frac{1}{N^2}\left(\sum_{k=1}^{N}D(\gamma_k) + \sum_{k\neq s}K(\gamma_k, \gamma_s)\right).$$

Тоді

$$D(h_1) \leq 2\frac{h_1(1-h_1)}{N}.$$

Аналогічно, для $h_2 = \rho(y, x)$, мають місце рівності

$$h_2 = \frac{1}{N}\sum_{k=1}^{N}\delta_k,\ E(h_2) = \frac{1}{N}\sum_{k=1}^{N}E(\delta_k) = p^{(n)},$$

$$D(h_2) = \frac{1}{N^2}\left(\sum_{k=1}^{N}D(\delta_k) + \sum_{k\neq s}K(\delta_k, \delta_s)\right).$$

Тоді

$$D(h_2) \leq 2\frac{h_2(1-h_2)}{N},$$

де δ_k — елементи індукованої вибірки, отриманої аналогічно елементам γ_k, коли вибірки x і y при обчисленні p-статистики міняються місцями. Звідси випливає, що для $h = \frac{1}{2}(\rho(x, y) + \rho(y, x))$ має місце

$$\lim_{N\to\infty} D(h) = 0.$$

Це означає, що асимптотичний рівень значущості інтервалу $I^{(n)} = \left(p^{(1)}, p^{(2)}\right)$ при $g = 3$ не перевищує 0,05. □

4. Алгоритм обчислення p-статистики

На вхід алгоритму подається два масиви $x[i]$, $i = \overline{0, n-1}$ та $y[j]$, $j = \overline{0, m-1}$. Загальна схема алгоритму обчислення p-статистики така:

1. Упорядковуємо за зростанням масиви x та y методом QuickSort [5], що має складність $O(n \ln n)$.
2. Обчислюємо значення елементів масиву $z[i], i = \overline{0, n-1}$ за алгоритмом, який подаємо як фрагмент програми на Pascal:

```
begin
  i1:=0; j1:=0; z[0]:=0;
  repeat
    if x[i1]>=y[j1] then
    begin
      z[i1]:=z[i1]+1; j1:=j1+1;
    end
    else
```

```
      begin
        i1:=i1+1;
        if i1<=(n-1) then z[i1]:=z[i1-1];
      end;
    until (j1>(m-1)) or (i1>(n-1));
    if i1<(n-1) then for i:=i1+1 to n-1 do z[i]:=z[i-1];
  end;
```

3. Покладемо $L = 0$. У двох вкладених циклах по $i = \overline{0, n-2}$ та $j = \overline{i+1, n-1}$ обчислюємо:
   - $h_{ij}^{(n)}$ за формулою (5);
   - межі кожного довірчого інтервалу $I_{ij}^{(n)}$ за формулами (2)–(3) ;
   - виконуємо $L = L + 1$ за виконання умови $\left\{p_{ij}^{(n)} \in I_{ij}^{(n)}\right\}$.
4. Обчислюємо p-статистику формулою (4).

5. Висновки

Запропонований алгоритм обчислення p-статистики має складність $O\left(n^2\right)$. Його практичне використання продемонстриване в роботі [6].

Література

1. Ганина К. П. Количественные характеристики ядерного полиморфизма эпителиальных клеток при фиброаденоматозе и раке молочной железы / К .П. Ганина, Ю. И. Петунин, Я. Г. Тимошенко // Докл. АН УССР. — 1984. — 35, № 12. — С. 1414–1449.
2. Клюшин Д. А. Непараметрический критерий эквивалентности генеральных совокупностей, основанный на мере близости между выборками / Д. А. Клюшин, Ю. І. Петунин // Укр. матем. журн. — 2003. — Т. 5, № 2. — С. 147–163.
3. Петунин Ю. И. Случайные точечные процессы с независимым маркированием / Ю. И. Петунин, Н. Г. Семейко // ДАН СССР. — 1986. — Т. 288, № 4. — С. 823–827.
4. Ван дер Варден. Математическая статистика. / Ван дер Варден— М.: Иностранная литература, 1960. — 436 с.
5. Кормен Т. Алгоритмы: построение и анализ. 2-е изд. / Т. Кормен, Ч. Лейзерсон, Р. Ривест, К. Штайн — М: Вильямс, 2005. — 1296 с.
6. Розпізнавання зображень ядер клітин за допомогою міри близькості / І. П. Сіренко // Журнал обчисл. та приклад. матем. — 2007. — №1 (94). — С. 87–90.

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

Надійшла 19.11.12