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

УДК 519.234.3+519.234.7

БАГАТОВИМІРНА СТАТИСТИКА ВКЛЮЧЕННЯ

В. В. Алєксєєнко

Резюме. Розглядається нова багатовимірна непараметрична статистика, заснована на порядкових статистиках і частотних характеристиках. За основу взято p-статистику і статистику включення. Пропонується новий непараметричний критерій для перевірки гіпотези про однорідність багатовимірних вибірок.

Вступ

Нехай маємо еталонні набори об'єктів, що відповідають відомим класам, і тестові набори об'єктів, належність до класів для яких необхідно визначити. Якщо для таких наборів можна виділити певну числову ознаку, то для класифікації можна застосувати p-статистику, запропоновану Д. А. Клюшиним і Ю. І. Петуніним у роботі [1]. Як альтернативу можна застосувати статистику включення [2], що дозволяє значно зменшити складність обчислення, використовуючи ту саму ідею.

Розглянемо ситуацію, коли для наборів об'єктів можна виділити декілька незалежних числових ознак. Для того щоб застосувати p-статистику або статистику включення можна лінеаризувати вектор ознак, отримавши одну ознаку, але в такому випадку втрачається інформація, яку можна було б використати для класифікації.

Мета роботи — побудувати статистику для порівняння вибірок об'єктів, представлених векторами незалежних числових ознак.

Для спрощення викладок роглянемо об'єкти, для яких можна виділити дві незалежні ознаки.

Багатовимірна статистика включення.

Нехай $x_{1,1}, x_{1,2}, ..., x_{1,n}$ та $x_{2,1}, x_{2,2}, ..., x_{2,n}$ — вибірки, отримані простим випадковим вибором із генеральних сукупностей G_1 і G_2 відповідно, породжені неперервними випадковими величинами із функціями розподілу F_{G_1} і F_{G_2}. Нехай $x_{(1,1)} < x_{(1,2)} < ... < x_{(1,n)}$ та $x_{(2,1)} < x_{(2,2)} < ... < x_{(2,n)}$ — відповідні варіаційні ряди.

Нехай $y_{1,1}, y_{1,2}, ..., y_{1,m}$ та $y_{2,1}, y_{2,2}, ..., y_{2,m}$ — вибірки, отримані простим випадковим вибором із генеральних сукупностей H_1 і H_2 відповідно, породжені неперервними випадковими величинами із функціями розподілу F_{H_1} і F_{H_2}.

Згідно гіпотези Хілла

$$P\left(x_{r,k} \in (x_{(r,i)}, x_{(r,j)})\right) = \frac{j-i}{n+1},\ r = 1, 2. \tag{1}$$

Уведемо позначення:

$$I_0^r = \left(-\infty; x_{(r,1)}\right), I_n^r = \left(x_{(r,n)}, \infty\right),\ I_j^r = \left(x_{(r,j)}, x_{(r,j+1)}\right),\ j = \overline{1, n-1}, \quad (2)$$

$$U_{i,j} = I_i^1 \times I_j^2. \quad (3)$$

Тоді за формулою (1) має місце:

$$p = P(\overline{x_k} \in U_{i,j}) = P(x_{1,k} \in I_i^1, x_{2,k} \in I_j^2) = \quad (4)$$

$$= P(x_{1,k} \in I_i^1) * P(x_{2,k} \in I_j^2) = \frac{1}{(n+1)^2}. \quad (5)$$

Позначимо $l_{i,j}$ — кількість пар $\overline{y_k} = (y_{1,k}, y_{2,k})$, таких що $\overline{y_k} \in U_{i,j}$ $(y_{1,k} \in I_i^1, y_{2,k} \in I_j^2)$ і $f_{i,j} = \frac{l_{i,j}}{m}$ – частоту потрапляння елементів вибірки $\overline{y_k}$ в множину $U_{i,j}$.

Уведемо багатовимірну статистику включення, яка є абсолютною величиною відхилення між частотою і ймовірністю потрапляння векторів вибірки $\overline{y_k}$ в множини $U_{i,j}$:

$$\eta = \sum_{i,j=0}^{n} |p - f_{i,j}|. \quad (6)$$

Уведемо такі позначення:

$\overline{b} = \begin{cases} \left[\frac{m}{(n+1)^2}\right] + 1, & \frac{m}{(n+1)^2} > \left[\frac{m}{(n+1)^2}\right] \\ \frac{m}{(n+1)^2}, & \frac{m}{(n+1)^2} = \left[\frac{m}{(n+1)^2}\right] \end{cases}$,

$\overline{h}$ – кількість (i,j) таких, що $l_{i,j} \geq \overline{b}$,

$\overline{w} = \sum\limits_{l_{i,j} \geq \overline{b}} l_{i,j} - \overline{b}$, $\delta_+ = \overline{b} - \frac{m}{(n+1)^2}$, $\underline{b} = \overline{b} - 1$,

$\underline{h}$ – кількість (i,j) таких, що $l_{i,j} \leq \underline{b}$,

$\underline{w} = \sum\limits_{l_{i,j} \leq \underline{b}} \underline{b} - l_{i,j}$, $\delta_- = \frac{m}{(n+1)^2} - \underline{b}$.

Лема 1. *$\underline{h}$, $\underline{w}$, δ_- можна визначити через $\overline{h}$, $\overline{w}$, δ_+.*

Доведення. $\underline{h} = (n+1)^2 - \overline{h}$.

$$\underline{w} = \sum_{l_{i,j} \leq \underline{b}} \underline{b} - l_{i,j} = \sum_{l_{i,j} \leq \underline{b}} \underline{b} - \sum_{l_{i,j} \leq \underline{b}} l_{i,j} = \underline{hb} - \left(m - \sum_{l_{i,j} \geq \overline{b}} l_{i,j}\right) =$$

$$= ((n+1)^2 - \overline{h})(\overline{b} - 1) - \left(m + \sum_{l_{i,j} \geq \overline{b}} (\overline{b} - l_{i,j}) - \sum_{l_{i,j} \geq \overline{b}} \overline{b}\right) =$$

$$= (n+1)^2\overline{b} - (n+1)^2 - \overline{hb} + \overline{h} - (m + \overline{w} - \overline{hb}) = (n+1)^2\overline{b} + \overline{h} - (n+1)^2 - m - \overline{w}.$$

Отже,

$$\delta_- = \frac{m}{(n+1)^2} - \underline{b} = \frac{m}{(n+1)^2} - \overline{b} + 1 = 1 - \delta_+.$$

□

Теорема 1. *$\overline{h}$ однозначно визначає багатовимірну статистику включення для визначених m і n.*

Доведення.

$$\eta = \sum_{i,j=0}^{n} |p - f_{i,j}| = \sum_{l_{i,j} \geq \overline{b}} \frac{\left| \frac{m}{(n+1)^2} - l_{i,j} \right|}{m} + \sum_{l_{i,j} \leq \underline{b}} \frac{\left| \frac{m}{(n+1)^2} - l_{i,j} \right|}{m}) =$$

$$= \frac{\sum\limits_{l_{i,j} \geq \overline{b}} (l_{i,j} - \frac{m}{(n+1)^2}) + \sum\limits_{l_{i,j} \leq \underline{b}} (\frac{m}{(n+1)^2} - l_{i,j})}{m} =$$

$$\frac{\sum\limits_{l_{i,j} \geq \overline{b}} (l_{i,j} - \overline{b} + \delta_+) + \sum\limits_{l_{i,j} \leq \underline{b}} (\underline{b} - l_{i,j} + \delta_-)}{m} = \frac{\overline{w} + \overline{h}\delta_+ + \underline{w} + \underline{h}\delta_-}{m}.$$

За лемою 1 маємо

$$\eta = \frac{\overline{w} + \overline{h}\delta_+ + (n+1)^2\overline{b} + \overline{h} - (n+1)^2 - m - \overline{w} + ((n+1)^2 - \overline{h})\delta_-}{m} = \eta(\overline{h}).$$

Теореми є слушною, оскільки $\overline{b}$, δ_+ і δ_- обчислюються через m і n. □

Розподіл багатовимірної статистики включення.

Припустимо, що вибірки $x_{1,i}$ та $y_{1,j}$ належать одній генеральній сукупності G_1 ($G_1 = H_1$), так само як і $x_{2,i}$ та $y_{2,j}$ належать одній генеральній сукупності G_2. Цю гіпотезу позначимо як H_0.

Розглянемо об'єднані вибірки $z_{1,1}, z_{1,2}, ..., z_{1,m+n} = x_{1,1}, x_{1,2}, ..., x_{1,n}$, $y_{1,1}, y_{1,2}, ...y_{1,m}$ та $z_{2,1}, z_{2,2}, ..., z_{2,m+n} = x_{2,1}, x_{2,2}, ..., x_{2,n}$, $y_{2,1}, y_{2,2}, ...y_{2,m}$ і варіаційні ряди $z_{(1,1)} < z_{(1,2)} < ... < z_{(1,m+n)}$ та $z_{(2,1)} < z_{(2,2)} < ... < z_{(2,m+n)}$. Фактично така пара варіаційних рядів утворює "варіаційну сітку".

Оскільки елементи вибірок $y_{r,j}$ належать генеральним сукупностям G_r і відповідним варіаційним рядам, то ми можемо вважати, що всі можливі позиції, що займають елементи y_{r_j} у варіаційних рядах, є рівноімовірними. Набір позицій елементів тестової вибірки $(\overline{y_k})$ у варіаційному ряді однозначно визначає матриця $l_{i,j}$ частот потрапляння векторів вибірки $\overline{y_k}$ в множини $U_{i,j}$. Всього різних розподілів $\overline{l_{i,j}}$ існує $C^m_{m+n^2}$ — кількість способів обрати у "варіаційній сітці"m позицій, що відповідають елементам $\overline{y_k}$.

Позначимо як $S(n,k,l)$ кількість способів розмістити k кульок по n урнах так, щоб в кожній було не більше ніж l кульок. Ця величина може бути обчислена за допомогою таких рекурентних формул:

$S(0,k,l) = 0, S(n,0,l) = 1, S(n,1,l) = n, S(1,k,l) = 1, k \leq l,$

$S(n,k,l) = S(n-1,k,l) + S(n-1,k-1,l) + ... + S(n-1,k-l,l).$

Теорема 2. *Для визначених характеристик $(\overline{h}, \overline{w})$ кількість різних матриць $l_{i,j}$, для яких характеристики будуть набувати цих значень, можна обчислити за формулою:*

$$N(\overline{h}, \overline{w}) = C^{\overline{h}}_{(n+1)^2} C^{\overline{w}}_{\overline{h}+\overline{w}-1} S((n+1)^2 - \overline{h}, m - \overline{bh} - \overline{w}, \overline{b} - 1). \qquad (7)$$

Доведення. Кількість способів обрати $\overline{h}$ позицій серед $(n+1)^2$ дорівнює $C^{\overline{h}}_{(n+1)^2}$. Згідно позначень $\overline{h}$ — кількість позицій матриці $l_{i,j}$, вибіркові значення в яких перевищують або дорівнюють $\overline{b}$, $\overline{w}$ — величина, на яку по

цих позиціях сума елементів у інтервалі більша за $\overline{b}$. Кількість матриць, в яких на $\overline{h}$ позиціях знаходяться числа, що перевищують або дорівнюють $\overline{b}$ так, що $\sum\limits_{i,j}(l_{i,j}-\overline{b})=\overline{w}$ — це те саме, що кількість способів спочатку розташувати $\overline{b}$ кульок у $\overline{h}$ урнах, потім ще додати $\overline{w}$ кульок до цих урн. Кулькість способів – $C^{\overline{w}}_{\overline{h}+\overline{w}-1}$. На інших $(n+1)^2-\overline{h}$ позиціях залишається всього $m-\overline{bh}-\overline{w}$ елементів, при чому кількість елементів в кожній множині не більше ніж $\overline{b}-1$. Тобто кількість варіантів дорівнює $S((n+1)^2-\overline{h},m-\overline{bh}-\overline{w},\overline{b}-1)$. Отже, загальна кількість векторів дорівнює $C^{\overline{h}}_{(n+1)^2}C^{\overline{w}}_{\overline{h}+\overline{w}-1}S((n+1)^2-\overline{h},m-\overline{bh}-\overline{w},\overline{b}-1)$. □

Наслідок 1. *Для визначених m і n імовірність того, що матриці розподілу $l_{i,j}$ буде відповідати пара $(\overline{h},\overline{w})$, дорівнює*

$$p(\overline{h},\overline{w})=\frac{C^{\overline{h}}_{(n+1)^2}C^{\overline{w}}_{\overline{h}+\overline{w}-1}S((n+1)^2-\overline{h},m-\overline{bh}-\overline{w},\overline{b}-1)}{C^m_{m+n}}. \tag{8}$$

Критерій прийняття гіпотези H_0

1. Для даних m та n (розмірностей тестової та еталонної вибірки) обчислити ймовірності для всіх можливих пар (h,w).
2. Визначити пару характеристик (h_0,w_0) для заданих тестової і еталонної вибірки і відповідну ймовірність $p(h_0,w_0)$
3. Обчислити α – суму імовірностей менш імовірних пар характеристик (h,w), ніж (h_0,w_0).
4. Нехай бажаний рівень значущості α_0. Тоді критерієм прийняття гіпотези H_0 буде $\alpha>\alpha_0$.

Висновки

На основі існуючих одновимірних статистик запропоновано нову багатовимірну непараметричну статику. За допомогою цієї статистики побудований критерій перевірки гіпотези про еквівалентність вибірок векторів ознак.

Література

1. Клюшин Д. Непараметрический критерий эквивалентности генеральных совокупностей, основанный на мере близости между выборками / Д. Клюшин, Ю. Петунин // Укр. матем. журн. — 2003. — 5(2). — С. 147–163.
2. Алексєєнко В. Статистика включення / В. Алексєєнко // Журнал обчисл. та прикл. математики. — 2012. — 1(107). — С. 105–111.
3. Hill B. Posteriori distribution of percentiles: Bayes' theorem for sampling from a population / B. Hill // J. Amer. Statist. Assoc. — 1968, Oct. — 63. — P. 677–691.

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

Надійшла 23.03.12.