ՀԱՅԿԱԿԱՆ ՍՍՀ ԳԻՏՈՒԹՅՈՒՆՆԵՐԻ ԱԿԱԴԵՄԻԱՅԻ ԶԵԿՈՒՅՑՆԵՐ
ДОКЛАДЫ АКАДЕМИИ НАУК АРМЯНСКОЙ ССР

LVII 1973 3

УДК 519.2

МАТЕМАТИКА

Д. Г. Асатрян

Эмпирическое байесовское сравнение двух совокупностей

(Представлено чл.-корр. АН Армянской ССР Р. А. Александряном 19/IV 1973)

Эмпирический байесовский подход к задачам математической статистики предложен Г. Роббинсом ([1]). В ([2,3]) эмпирический байесовский подход применен в задаче проверки гипотез относительно параметра некоторых распределений из экспоненциального семейства при полиномиальной функции потерь.

В настоящей заметке результаты работ ([1-3]) обобщаются на задачу эмпирического байесовского сравнения двух совокупностей.

1°. Рассмотрим сначала байесовскую задачу сравнения. Пусть независимые случайные величины X и Y имеют плотности распределения вероятностей (относительно некоторой σ—конечной меры μ) из семейств $P=\{p_\theta(x):\theta\in\Theta\}$ и $Q=\{q_\omega(y):\omega\in\Omega\}$ соответственно, определенные на выборочных пространствах X и Y, где Θ и Ω—действительные параметрические пространства, а область, на которой $p_\theta(x)>0$ ($q_\omega(y)>0$) не зависит от θ (ω). Предположим, что параметры θ и ω являются независимыми случайными величинами, имеющими (априорные) функции распределения из множеств $A_1=\{G_1(\theta)\}$ и $A_2=\{G_2(\omega)\}$, определенные над Θ и Ω соответственно и пусть

$$f_{G_1}(x)=\int_\Theta p_\theta(x)\,dG_1(\theta), \tag{1}$$

$$f_{G_2}(y)=\int_\Omega q_\omega(y)\,dG_2(\omega) \tag{2}$$

безусловные плотности распределения X и Y (относительно меры μ), и $A=\{G=G_1G_2:\ G_1\in A_1,\ G_2\in A_2\}$.

Пусть H_0 и H_1 обозначают гипотезы относительно пары (θ,ω). Требуется указать процедуру, позволяющую по результату наблюде-

ния (x, y) над (X, Y) принимать одну из заданных гипотез так, чтобы средние ожидаемые потери были минимальными.

Пусть событие a_j обозначает выбор гипотезы H_j $(j=0,1)$. Будем предполагать, что функция потерь $L(a_j; \theta, \omega) \geqslant 0$ известна и удовлетворяет условию

$$E_G|| L(a_0; \theta, \omega) - L(a_1; \theta, \omega)|| < \infty, \qquad (3)$$

где E_G—символ математического ожидания относительно распределения G.

Пусть $t = t(x,y)$ решающая функция, которую будем предполагать равной вероятности выбора гипотезы H_1, когда результатом наблюдения является (x,y). Тогда средние потери, соответствующие решающей функции t, равны

$$R(t; \theta, \omega) = \int_X \int_Y |t(x, y)\ L(a_1; \theta, \omega) +$$

$$+(1 - t(x, y))\ L(a_0; \theta, \omega)]\ p_\theta(x) q_\omega(y) d_\mu(x) d_\mu(y), \qquad (4)$$

следовательно, полные средние потери, соответствующие априорному распределению G, равны

$$R(t; G) = \int_\Theta \int_\Omega R(t; \theta, \omega)\ dG_1(\theta)\ dG_2(\omega). \qquad (5)$$

Предположим, что существует решающая функция $t_G(x, y)$ такая, что $R(G) = R(t_G; G) = \min_t R(t; G)$.

Подставляя (4) в (5) можем написать

$$R(t; G) = E[\ t(x, y)\ L(a_1; \theta, \omega) + (1 - t(x, y))\ L(a_0; \theta, \omega)], \qquad (6)$$

где $E|\cdot|$ —математическое ожидание относительно совместного распределения всех случайных величин, входящих в $|\ |$.
Обозначая

$$\Delta_G(x, y) = E_G|L(a_0; \theta, \omega) - L(a_1; \theta, \omega)|x, y|, \qquad (7)$$

легко привести (6) к виду

$$R(t; G) = E_G[L(a_0; \theta, \omega)] - E|t(x, y)\ \Delta_G(x, y)|, \qquad (8)$$

откуда следует, что при фиксированном G решающая функция

$$t_G(x, y) = 1 \text{ при } \Delta_G(x, y) > 0,$$
$$= 0 \text{ при } \Delta_G(x, y) \leqslant 0 \qquad (9)$$

минимизирует (8) и, если $P\{\Delta_G(x, y)\} = 0$, то (9) единственна с точностью до множества нулевой вероятности.

Таким образом, если априорное распределение G известно, то

решающая функция (9) вместе с (7) дает решение поставленной задачи.

2°. Теперь сформулируем эмпирическую байесовскую задачу сравнения двух совокупностей. Предположим, что задача сравнения, рассмотренная выше, встречается многократно и независимо, всегда с тем же самым, но неизвестным априорным распределением $G \in A$. Пусть

$$(x_1; \theta_1), \ldots, (x_n; \theta_n); \quad (y_1; \omega_1), \ldots, (y_n; \omega_n) \tag{10}$$

результаты независимых наблюдений над X и Y с плотностью распределения (1) и (2) и соответствующие им значения параметров θ и ω, которые имеют одно и то же распределение G и не наблюдаются. Пусть (x, y) является результатом $n+1$-го наблюдения над (X, Y). Обозначим $\overline{x_n} = (x_1, \ldots, x_n)$, $\overline{y_n} = (y_1, \ldots, y_n)$ и предположим, что (X, Y) не зависит от $(\overline{x_n}, \overline{y_n})$. Требуется построить такую решающую функцию $t_n(x, y) = t(x, y; \overline{x_n}, \overline{y_n})$, чтобы полные ожидаемые потери по мере возрастания n стремились к минимально возможному значению $R(G)$, достижимому лишь в случае полностью известного G, т. е. чтобы

$$\lim_{n \to \infty} E|R(t_n; G)| = R(G). \tag{11}$$

Решающая функция t_n, удовлетворяющая (11), называется асимптотически оптимальной относительно G ([2]).

Решение этой задачи основано на результатах ([1–3]), сформулированных в терминах эмпирической байесовской задачи сравнения. В частности, верна

Теорема. *Пусть* $\Delta_n(x, y) = \Delta(x, y; \overline{x_n}, \overline{y_n})$ *— последовательность (случайных) функций таких, что для каждого* $x \in X$ *и* $y \in Y$

$$p \lim_{n \to \infty} \Delta_n(x, y) = \Delta_G(x, y) \tag{12}$$

(относительно распределения $(\overline{x_n}, \overline{y_n})$*). Предположим, что функция потерь* $L(a_j; \theta, \omega)$ *такова, что при* $G \in A$ *выполняется* (3) *и* $P|\Delta_G(x, y) = 0| = 0$. *Решающую функцию* t_n *определим следующим образом*

$$t_n(x, y) = 1 \text{ при } \Delta_n(x, y) > 0, \tag{13}$$

$$= 0 \text{ при } \Delta_n(x, y) \leqslant 0.$$

Тогда решающая функция t_n *асимптотически оптимальна относительно каждой пары* $G_1 \in A_1$ *и* $G_2 \in A_2$.

Теперь построим последовательность Δ_n, удовлетворяющую (12) в предположениях, что функция потерь имеет вид

$$L(a_0; \theta, \omega) - L(a_1; \theta, \omega) = \sum_{k=0}^{s_1} \sum_{m=0}^{s_2} c_{km} \theta^k \omega^m, \tag{14}$$

а плотности из семейств P и Q представимы в виде

$$p_\theta(x) = \theta^x c_1(x) h_1(\theta) \quad (x \in X)$$

$$q_\omega(y) = \omega^y c_2(y) h_2(\omega) \quad (y \in Y). \tag{15}$$

Очевидно, для выполнения (3) в предположениях (14) и (15) достаточно, чтобы абсолютный момент s_r-го порядка распределения G_r $(r = 1,2)$ был конечен.

Подставляя (14) и (15) в (7) и учитывая (1) и (2), получим

$$\Delta_G(x, y) = \sum_{k=0}^{s_1} \sum_{m=0}^{s_2} c_{km} U_k(x) V_m(y), \tag{16}$$

где

$$U_k(x) = c_1(x) f_{G_1}(x+k) / c_2(x+k) f_{G_1}(x), \tag{17}$$

$$V_m(y) = c_2(y) f_{G_2}(y+m) / c_2(y+m) f_{G_1}(y) \tag{18}$$

при $x+k \in X$ $(k=0,1,\ldots,s_1)$ и $y+m \in Y$ $(m=0,1,\ldots,s_2)$.

Отметим, что в частном случае, когда распределение G_2 сосредоточено в некоторой точке $\omega_0 \in \Omega$, имеем $V_m(y) = \text{const}$ и $\Delta_G(x, y) = \Delta_{G_1}(x) = \sum_{k=0}^{s_1} c_k U_k(x)$, т. е. мы приходим к однопараметрической задаче, рассмотренной в [2,3].

Как видно из (16)—(18), для построения последовательности Δ_n, удовлетворяющей (12), достаточно в этих формулах $f_{G_r}(\cdot)$ заменить своей состоятельной (непараметрической) оценкой $f_{r,n}(\cdot)$, построенной по данным (10). В частности, в случае дискретных f_{G_r} можно положить

$$f_{r,n}(z) = n^{-1} \text{ (число тех } i, \text{ для которых } z_i = z).$$

Таким образом, если обозначить

$$U_{k,n}(x) = c_1(x) f_{1,n}(x+k) / c_1(x+k) f_{1,n}(x),$$

$$V_{m,n}(y) = c_2(y) f_{2,n}(y+m) / c_2(y+m) f_{2,n}(y),$$

то искомая последовательность будет иметь вид

$$\Delta_n(x, y) = \sum_{k=0}^{s_1} \sum_{m=0}^{s_2} c_{km} U_{k,n}(x) V_{m,n}(y).$$

3°. В качестве примера рассмотрим эмпирическое байесовское сравнение двух пуассоновских совокупностей.

Пусть $X = Y = \{0,1,2,\ldots\}$, μ —считающая мера на X и Y,

$$p_\theta(x) = \theta^x (x!)^{-1} e^{-\theta} (\theta > 0), \tag{19}$$

$$q_\omega(y) = \omega^y (y!)^{-1} e^{-\omega} (\omega > 0).$$

Рассмотрим следующую пару гипотез: H_0: $\theta-\omega\leqslant l_0$, H_1: $\theta-\omega>l_0$. Функцию потерь определим следующим образом

$$L(a_j;\ \theta,\ \omega)=\max(0,\ \theta-\omega-l_0),\ i=0,$$

$$=-\min(0,\ \theta-\omega-l_0),\ j=1.$$

Очевидно, в этом случае имеем $L(a_0;\ \theta,\ \omega)-L(a_1;\ \theta,\ \omega)=\theta-\omega-l_0$, т. е. условия (14) и (15) выполняются. Предположим, что A_1 и A_2 содержат лишь распределения с конечным математическим ожиданием. Положив $\delta_1(x)=(x!)^{-1}$, $\delta_2(y)=(y!)^{-1}$ и предположив, что $f_{1,n}(x)f_{2,n}(y)\neq 0$*, получим

$$U_{k,n}(x)=(x+k)!\,f_{1,n}(x+k)/x!\ f_{1,n}(x)\quad(k=0,\ 1),$$

$$V_{m,n}(y)=(y+m)!f_{2,n}(y+m)/y!\ f_{2,n}(y)\quad(m=0,\ 1).$$

Критерий $\Delta_n(x,\ y)$ в этом случае принимает вид

$$\Delta_n(x,\ y)=U_{1,n}(x)-V_{1,n}(y)-l_0=$$

$$=(x+1)f_{1,n}(x+1)/f_{1,n}(x)-(y+1)f_{2,n}(y+1)/f_{2,n}(y)-l_0. \qquad (20)$$

Согласно теореме, решающая функция (13), определенная с помощью (20), асимптотически оптимальна относительно всех априорных распределений с конечным математическим ожиданием.

Аналогичную решающую функцию можно построить и в случае двухсторонних гипотез типа H_0: $|\theta-\omega|\leqslant l_0$, H_1: $|\theta-\omega|>l_0$, если функция потерь определена, например, следующим образом

$$L(a_j,\ \theta,\ \omega)=\max|0,\ (\theta-\omega)^2-l_0^2|,\quad j=0,$$

$$=-\min|0,\ (\theta-\omega)^2-l_0^2|,\quad j=1.$$

4°. В заключение отметим, что семейства распределений P и Q, определенные согласно (15) включают, кроме пуассоновского, еще ряд хорошо известных распределений из экспоненциального семейства, например, биномиальное, геометрическое, нормальное, гамма и другие распределения. Некоторые из этих распределений сводятся к виду (15) с помощью замены неизвестного параметра некоторой функцией от него. Так, для биномиального распределения это преобразование имеет вид $\theta=p/(1-p)$ $(0<p<1)$. В таких случаях, разумеется, гипотезы и соответствующие им функции потерь формулируются в терминах преобразованных параметров. Если функция потерь не является полиномиальной относительно параметров θ и ω, то следует ограничиваться ее аппроксимацией полиномами вида (14).

Вычислительный центр
Министерства легкой промышленности Армянской ССР

---

* Вероятность этого события стремится к 1 при $n\to\infty$

Գ. Գ. ԱՍԱՏՐՅԱՆ

Երկու համախմբությունների էմպիրիկ բայեսյան համեմատում

Դիտարկվում է էքսպոնենցիալ ընտանիքին պատկանող հավանականությունների բաշխման երկու ֆունկցիաների պարամետրերի համեմատման խնդիրը, երբ վերջիններս միմյանցից անկախ պատահական մեծություններ են՝ բաշխված ըստ անհայտ օրենքի։ Առաջարկվում է այդ խնդրի ասիմպտոտիկորեն օպտիմալ լուծում բազմանդամի տեսք ունեցող կորուստների ֆունկցիայի և մի քանի վերջավոր բացարձակ մոմենտներ ունեցող ապրիորբաշխման ֆունկցիայի համար։

ЛИТЕРАТУРА — ԳՐԱԿԱՆՈՒԹՅՈՒՆ

[1] *H. Robbins*, Proc. Third Berkeley Symp. on Math. Stat. and Prob., Univ. of Calif. Press, v 1, 1956 (Русск. пер. в сб. „Математика", 8:2. 1964). [2] *H. Robbins*, Ann. Math. Statist., v 35, 1964 (Русск. пер. в. сб. „Математика", 10:5, 1966). [3] *E. Samuel*, Ann. Math. Statist., v 34, 1963.