I. ИНФОРМАТИКА

УДК 519.68: 681.513.7

КАК ОЦЕНИТЬ НАДЕЖНОСТЬ АЛГОРИТМА КЛАССИФИКАЦИИ. II. ИНТЕРВАЛЬНЫЕ ОЦЕНКИ

С.И. Гуров

факультет ВМиК МГУ им. Ломоносова, г.Москва, Россия
e-mail: *sgur@cs.msu.su, gurov@ccas.ru*

Работа выполнена при поддержке гранта РФФИ № $01-01-00885_{-1}$

Abstract

Investigation on the estimates of trustworthy of classification algorithms is preceded in the paper. Existing methods are considered and new methods of derivation of interval estimate useful for case of small number of precedents are offered.

Данная статья является продолжением предыдущей [6]. Для удобства ссылок обе статья имеют сквозную нумерацию разделов.

6. Интервальные оценки

Обычно в математической статистике пользуются интервальными оценками, имеющими достоверность $\eta = 0.9; 0.95; 0.98; 0.99$ и т.д. представляется, что для задач оценки надежности распознающих алгоритмов в большом числе случаев надежность $\eta = 0.95$ или даже $\eta = 0.9$ будет достаточной.

6.1. **Частотный подход.** В рамках частотного подхода для получения интервальных оценок параметров распределений используются следующие методы:

- метод кратчайших доверительных интервалов;
- метод наиболее селективных интервалов;
- метод фидуциальных интервалов;
- метод Большева.

Метод базируется на элементарных свойствах функций распределений. Второй и третий методы предложены, соответственно, Дж.Нейманом и Р.Фишером. Как будет видно из дальнейшего, для нашей задачи представляют интерес первые два метода. рассмотрим их применение сначала в одномерном, а затем в многомерном случае. Замечания относительно остальных методов см. в конце нижеследующего раздела.

6.1.1. **Одномерный случай.** При $v = 2$ наша задача состоит в том, чтобы построить доверительный интервал для вероятности ошибочного распознавания $p^*_w = p^*$ с надежностью η, среди m прецедентов имеется m_w неправильно распознанных построенным р.п. Доверительный интервал оценивания записывается в виде $J = (p_-, p_+)$.

Кратчайшие доверительные интервалы. Рассмотрим построение кратчайших доверительных интервалов. Достаточная статистика m_w имеет биномиальное распределение $B_i(m,p)$ с функцией распределения

$$P_p\{m_w \leqslant t\} = P\{m_w \leqslant t|m,p\} = P(t) = \sum_{k=0}^{t} \binom{m}{k} p^k (1-p)^{m-k} \tag{1}$$

и функцией выживания

$$P_p\{m_w > t\} = 1 - P(t-1) = \sum_{k=t}^{m} \binom{m}{k} p^k (1-p)^{m-k}. \tag{2}$$

В этих формулах $p \in (0,1)$ и $t = 0,1,...,m$. Для $J = (p_-, p_+)$ должно выполняться условие

$$P\{p^* \notin J\} = P_{p+}\{m_w \leqslant t_1\} + P_{p-}\{m_w > t_2\} \leqslant 1 - \eta = \alpha, \tag{3}$$

где t_1 и $t_2, t_1 \leqslant t_2-$ целые значения t в (1) и (2) при подстановке в указанные зависимости p_+ и p_- соответственно.

Выражение (3) означает, что с достоверностью не меньше, чем η выполняются двойные неравенства $t_1 < m_w \leqslant t_2$ и $p_- < p^* < p_+$. Здесь и выше m_w рассматривается как случайная величина, а не как конкретное ее значение.

Вычисление значения биномиальных вероятностей $B_i(m,p)$

$$p(m_1) = \binom{m}{m}_1 p^{m_1}(1-p)^{m-m_1}; p \in (0,1) \tag{4}$$

или функции распределения (1) является весьма трудоемкой процедурой. Поэтому во всех случаях, когда это это возможно $m \gg 1$, прибегают к аппроксимации биномиального распределения.

В случае **больших выборок** и не слишком малых p^*, точнее, если одновременно mp^* и $m(1-p^*) > 5$ для вычисления границ доверительного интервала можно воспользоваться аппроксимацией биномиального распределения нормальным [7]. Замена базируется на том факте, что первая производная логарифма функции правдоподобия L распределена асимптотически нормально со средним, равном нулю и дисперсией

$$\mathbf{D}\left(\frac{\partial \mathrm{lnL}}{\partial \mathrm{p}}\right) = \mathbf{M}\left\{\left(\frac{\partial \mathrm{lnL}}{\partial \mathrm{p}}\right)^2\right\} = -\mathbf{M}\left\{\frac{\partial^2 \mathrm{lnL}}{\partial \mathrm{p}^2}\right\}.$$

Для нашего случая биномиального распределения и $p \in (0,1)$ функции правдоподобия есть

$$L(p) = p^{m_w}(1-p)^{m-m_w},$$

и получаем, что величина

$$T = \frac{m_w - mp}{\sqrt{mp(1-p)}} \tag{5}$$

имеет асимптотически стандартное нормальное распределение $\mathcal{N}(\prime, \infty)$ т.е.

$$P\{m_w \leqslant t | m, p\} \approx \Phi_0(T), \tag{6}$$

где $\Phi_0(\cdot)-$ функция стандартного (нормированного и центрированного) нормального распределения. В силу этого

$$P\{-z_\eta < T < z_\eta\} \approx \tfrac{1}{\sqrt{2\pi}} \int\limits_{-z_\eta}^{z_\eta} e^{-\frac{x^2}{2}} dx = 2\Phi_0(z_\eta) = \eta.$$

Таким образом приходим к квадратному уравнению для границ искомого интервала

$$p^2\Big(1 + \tfrac{z_\eta^2}{m}\Big) - p\Big(2\hat{p} + \tfrac{z_\eta^2}{m}\Big) + \hat{p}^2 = 0.$$

Здесь $\hat{p} = m_w/m-$ несмещенная оценка вероятности p.

Легко показать, что этому уравнению в координатах p и $\hat{p}$ соответствует эллипс,вписанный в полосу $0 \leqslant p \leqslant 1$ и пересекающий единичный квадрат в точках (0, 0), (1 -c, 0), (1, 1), (c, 1), где $c = \frac{1}{1+z_\eta^2/m}$.[1]

Решая вышеприведённо квадратное уравнение, получаем

$$\begin{cases} p_- = \frac{m}{m+z_\eta^2}\left[\hat{p} + \frac{z_\eta^2}{2m} - z_\eta\sqrt{\frac{\hat{p}(1-\hat{p})}{m} + \left(\frac{z_\eta}{2m}\right)^2}\right], \\ p_+ = \frac{m}{m+z_\eta^2}\left[\hat{p} + \frac{z_\eta^2}{2m} + z_\eta\sqrt{\frac{\hat{p}(1-\hat{p})}{m} + \left(\frac{z_\eta}{2m}\right)^2}\right], \end{cases}$$

где величина z_η находится из уравнения

$$\Phi_0(z_\eta) = \tfrac{\eta}{2}$$

при помощи таблиц функции $\Phi_0(\cdot)$.

При значениях m порядка сотен можно пренебречь малыми значениями отношений $z^2/2m,\ z^2/4m^2,\ z^2/m$ и пользоваться более грубыми оценками

$$\begin{cases} p_- = \hat{p} - z_\eta\sqrt{\frac{\hat{p}(1-\hat{p})}{m}}, \\ p_+ = \hat{p} + z_\eta\sqrt{\frac{\hat{p}(1-\hat{p})}{m}}. \end{cases}$$

[1]Выход эллипса за полосу $0 \leqslant \hat{p} \leqslant 1$ связан с тем, что при p^* близких к 0 или 1 аппроксимация (6)некорректна и надо использовать пуассоновскую аппроксимацию (для p^* или $1 - p^*$, см. далее (8))

Особый случай представляет $\hat{p} = 0$(0 - событие). Для нас это случай корректного алгоритма с $m_w = 0$. Здесь точную верхнюю границу вычисляют по формуле

$$p_+ = 1 - \sqrt[m]{\alpha}, \tag{7}$$

соответственно, для полного события $\hat{p} = 1$ точная нижняя граница есть $p_- = \sqrt[m]{\alpha}$. Для $\alpha = 0.95$, $n > 50$ и $m_w = 0$ ($m_w = 1$) справедливо приближение $p_+ \simeq 3/m (p_- \simeq 1 - 3/m)$.

В случае больших m, но таких, что mp^* не слишком велико, биномиальное распределение можно аппроксимировать распределением Пуассона $P_0(k;\lambda) = \lambda^k \exp(-\lambda)/\mathrm{k!}$ с $\lambda = mp$:

$$p\{m_w \leqslant t | m, p\} \approx \sum_{k=0}^{t} \frac{(mp)^k}{k!} e^{-m_p}. \tag{8}$$

Далее можно воспользоваться методами доверительного (одностороннего $(0, \lambda^s_{\eta,+})$) или двустороннего $(\lambda_{\eta,-}, \lambda_{\eta,+})$) оценивания пуассоновского параметра λ при достоверности η [12], [15] и затем определить интервал для p : $(J = (0, \lambda^s_{\eta,+}/m) \text{или} J = (\lambda_{\eta,+}/m))$ соответственно. При невозможности использования аппроксимационных формул (подробный перечень предположений для их применения дан в [2], а для (6) - и в [7]) можно говорить, что имеет место **малая выборка.** В этом случае необходимо перейти к прямому решению уравнений (1) - (3).

Ясно, что задавая различные величины α_1 и α_2 в (3)

$\alpha_1 \triangleq P_{p-}\{m_w \leqslant t_1\} \geqslant, \qquad \alpha_2 \triangleq P_{p+}\{m_w > t_2\} \geqslant 0, \qquad \alpha_1 + \alpha_2 = \alpha = 1 - \eta,$

можно получать различные доверительные интервалы. При $\alpha_1 = \alpha_2 = \alpha/2$ соответствующий интервал J называется *центральным.*

Желание получить интервал наименьшей длины приводит к требованию максимальности t_1 и минимальности t_2 в (3).

Если условие (3) выполняется со знаком равенства, то данное требование приводит к однозначному определению p_-, p_+, t_1, t_2. К сожалению, это является скорее исключением, в силу чего границы p_-, p_+ доверительного интервала J по (3), как правило, не устанавливаются однозначно. Для разрешения указанного затруднения были предложены различные подходы.

Существуют [8], [9], [11] способы избежать неоднозначности определения границ доверительного интервала основанные на идее модификации выражения (3) с помощью введения дополнительных случайных величин. Такая процедура называется *рандомизацией.* При этом оказывается, что укорачивается. Это объясняется тем, что потери от новой неопределенности, связанной с введением случайной величины

оказываются меньше устраненных потерь от неопределенности, связанной с неравенством.

В [20], [23] предложено использовать центральные интервалы, т.е. разделить вероятность интервальной недоооценки и переоценки поровну и находить p_+, t_1 и p_-, t_2 из условий

$$\begin{cases} t_1 = \arg\max\limits_{0\leqslant t\leqslant m}\Big\{\mathrm{P_{p+}}\{\mathrm{m_w}\leqslant \mathrm{t}\}\Big\} = \dfrac{1-\eta}{2}, \\ t_2 = \arg\max\limits_{0\leqslant t\leqslant m}\Big\{\mathrm{P_{p-}}\{\mathrm{m_w} > \mathrm{t}\}\Big\} = \dfrac{1-\eta}{2} \end{cases} \tag{9}$$

соответственно. Здесь полученные значения t_1b t_2 однозначно определяют границы p_+ и p_- доверительного интервала. Следует только иметь в виду, что $p_+ = 1$ при $t = m$ и $p_- = 0$ при $t = 0$.

Уравнения (9) известны под названием формул Клоппера-Пирсона. Прямой метод их решения основанный на переборе значений $t = 0, 1, ...$ кратко описан в [18] и [9]. Однако соответствующий алгоритм, очевидно, достаточно трудоемок. Поэтому более удобно воспользоваться другим методом решения (9), основанным на использовании известной связи между величиной $B(t, m, p) \triangleq p_p\{m_w \leqslant t\}$ в (1) и функцией **F**-распределения случайной величины U_{v_1,v_2} с v_1, v_2 степенями свободы [12]:

$$B(t,m,p) = P\Big\{U_{2(m-t),2(t+1)} < \frac{(t+1)(1-p)}{(m-t)p}\Big\} = \\ = P\Big\{U_{2(t+1),2(m-t)} > \tfrac{(m-t)p}{(t+1)(1-p)}\Big\}.$$

С помощью указанных соотношений определяются точные формулы для границ центрального доверительного интервала. Они имеют следующий вид [7]:

$$\begin{cases} p_- = \dfrac{m_w}{m_w + (m - m_w + 1)F_{v_1^-,v_2^-}}, \\ p_+ = \dfrac{(m_w+1)F_{v_1^+,v_2^+}}{m - m_w + (m_w+1)F_{v_1^+,v_2^+}}, \end{cases} \tag{10}$$

где $F_{v_1^-,v_2^-}$ и $F_{v_1^+,v_2^+}-$ квантили **F** - распределения с $v_1^- = 2(m-m_w+1), v_2^- = 2m_w$ и $v_1^+ = 2(m_w+1), v_2^+ = 2(m-m_w)$ степенями свободы соответственно для доверительной вероятности ошибки $\alpha/2$.

Для решения (3) можно также воспользоваться таблицами биномиального распределения. Границы p_-, p_+ доверительного интервала будут тогда определяться как значения вероятности p, при которой величина (4) будет равняться $\frac{1-\eta}{2}$ и $\frac{1+\eta}{2}$ соответственно.

Для быстрого приближенного решения уравнения (9) со значениями достоверности $\alpha = 0.1; 0.05$ и объемов выборки $m = 10, ..., 1000$ построены графические зависимости между наблюдаемыми значениями $\hat{p}$ и относительными частотами генеральной совокупности, определяющими доверительный интервал (см. например [7], [10], [17]). Представляется, что точность данного графического метода в большинстве случаев достаточна для задач оценки надежности алгоритмов распознавания образов. Следует только иметь в виду, что на указанных графиках не учтен особый случай 0-события, когда нужно пользоваться формулой (7).

Скажем здесь, что с нашей точки зрения применение центральных интервалов для оценки вероятности ошибки $p^* = p_w^*$ алгоритма распознавания, вообще говоря, не является оправданным. Действительно, ошибка $\hat{p}$, как правило, мала (а для корректных алгоритмов вообще равна нулю), и мы хотим быть уверены, что ее величина не превзойдет некоторого значения. Поэтому ошибиться мы имеем право скорее в большую сторону. В силу этого для оценки p_w^* более адекватным представляется использование нецентральных, а для достаточно малых значений $\hat{p}$ и односторонних интервалов $J(0, p_+)$. В последнем случае в качестве p_+ берется соответствующая величина из (10), определенная для доверительной вероятности $1 - \eta$. Заметим, что при $\hat{p} = 0$ полученная оценка совпадет с (7).

Наиболее селективные интервалы. Дж.Нейман предложил метод построения доверительных интервалов, которые также назвал «кратчайшими» [22], [13]. Чтобы отличать интервалы J_N в [9] предложено называть **наиболее селективными** (там же см. обсуждение различий между кратчайшими доверительными и наиболее селективными интервалами). Наиболее селективные интервалы рассмотрены в [11]. Там же метод их построения, основанный на лемме Неймана-Пирсона.

Согласно неймановскому методу границы θ_-, θ_+ доверительного интервала $J_N = (\theta_-, \theta_+)$ с коэффициентом доверия $\eta = 2P - 1$, где $0.5 \leqslant P < 1$ для неизвестной величины θ^* определяются как решения соответственно первого и второго уравнений

$$G(T, \Theta) = \begin{cases} 1 - P, \\ P. \end{cases} \tag{11}$$

Здесь $G(T, \Theta)-$ непрерывная функция распределения статистики T, используемой в качестве точечной оценки θ^* и называемая *(неймановским) доверительным распределением* T. При условии выполнения некоторых условий регулярности [22], [11], [3], которые выполняются почти во всех интересных для практики случаях, вышеприведенные уравнения имеют единственные решения θ_-, θ_+.

В нашем случае $T = m_w, \theta = p \in (0,1)$ и $G(m_w, p)-$ функция распределения $B_i(m,p)$ в (1)

$$G(m_w, p) = P\{m_w \leqslant t|m, p\}.$$

Функцией распределения вероятностей **B**-распределения

$$f(p) = f(p|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{\alpha-1}(1-p)^{b-1}, \qquad p \in (0,1)$$

неполная **F**- функция, обозначаемая $I_p(a,b), \quad (0 \leqslant p \leqslant, a > 0, b > 0)$:

$$I_p(a,b) \triangleq \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^p x^{\alpha-1}(1-x)^{b-1} dx.$$

Неполная **B**-функция обладает свойством

$$I_p(a,b) \equiv 1 - I_{1-p}(a,b),$$

а для целых a и b имеют место замечательные равенства

$$\begin{cases} I_p(a,b) = \sum\limits_{k=a}^{a+b-1} \binom{a+b-1}{k} p^k (1-p)^{a+b-1-k}, \\ I_{1-p}(b,a) = \sum\limits_{k=b}^{\infty} \binom{a+k-1}{a-1} p^a (1-p)^k. \end{cases}$$

Используя эти равенства, легко показать, что

$$P\{m_w \leqslant t|m,p\} = \sum_{k=0}^{t} \binom{m}{k} p^k (1-p)^{m-k} = \\ = 1 - I_p(t+1, m-t) = I_{1-p}(m-t, t+1).$$

Значения функции биномиального распределения $G(m_w, p) = P\{m_w \leqslant t|m,p\}$ совпадают, следовательно, со значениями функции **B**-распределения $I_{1-p}(b,a)$ в целочисленных точках $b = m - t$ и $a = t + 1, t = 0, 1, ..., m$.

Ясно, что пользоваться хорошо изученной и затабулированной неполной **B**-функцией намного удобнее, чем с биномиальными суммами (1) и (2). Однако при замене $G(m_w, p)$ на $I_{1-p}(m-t, t+1)$возникает следующая трудность. Поскольку m_w подчиняется биномиальному закону и имеет дискретное распределение, функция $G(m_w, p)$ не будет непрерывной. Это, в свою очередь, приведет к неравенствам в формулах для определения границ интервала (11). А при использовании там равенств – к тому, что либо m_w должно быть нецелым числом, либо границы интервала не будут определяться однозначно. В качестве выхода из данной ситуации можно прибегнуть к рандомизации, рассмотрев новую статистику $T = m_w + U$, где $U-$ случайная величина, равномерно распределенная на (0,1). На практике же, чтобы избежать рандомизации и установить границы, не зависящие от дополнительной

величины U обычно величину T при определении верхней границы доверительного интервала заменяют величиной $m_w + 1$ [2]. При этом верхняя граница оказывается несколько завышенной, что, естественно, компенсируется большей вероятностью накрытия истинного значения p^*.

В результате [2] границы неймановского доверительного интервала $J_N = (p_-, p_+)$с коэффициентом доверия $\eta = 2P - 1$, где $0.5 \leqslant P < 1$ могут быть определены как решения уравнений

$$\begin{cases} I_{p-}(m_w, m - m_w + 1) = 1 - P = \dfrac{1-\eta}{2} \\ I_{p+}(m_w + 1, m - m_w) = P = \dfrac{1+\eta}{2}. \end{cases} \tag{12}$$

Для решения уравнений (12) можно воспользоваться таблицами **B**- распределения (см., например, [14], [12]). Границы доверительного интервала J_N для биномиального распределения табулированы в [19], [2], [12]. Заметим, что метод остается пригодным и когда неизвестный параметр рассматривается как случайный.

Фидуциальные интервалы. Пусть x − случайная величина с функцией распределения $G(x, \theta)$, где $\theta-$ некоторый скалярный параметр. При фиксированном x во многих случаях $G(x, \theta)$ с точностью до мультипликативной константы формально представляет собой функцию распределения вероятностей или функцию выживания θ, рассматриваемой как случайная величина.[3]

«Фидуциальное распределение не является распределением вероятности в смысле частотной теории. Это новое понятие, выражающее интенсивность нашей веры в различные возможныезначения параметра» [9]. Вопрос заключается в том, когда $G(x, \theta)$ действительно можно рассматривать как распределение вероятности в «частном» смысле. Ответ оказывается положительным при выполнении не слишком жестких условий [4], [1], [3]. При этом доверительные фишеровские и неймановские интервалы, как правило, совпадают. Это же имеет место и в нашем случае.

Метод доверительных интервалов Л.Н.Большева [3] совмещает фишеровский фидуциальный подход (в тех случаях, когда он применим) с построением селективных интервалов по Нейману. В силу этого в нашей задаче он не приведет к новым результатам.

6.1.2. **Многомерный случай.** Отметим сначала технические и математические сложности работы с многомерными доверительными интервалами и некорректность применения прямых методов построения доверительного минимального интервала к

[2]конечно, если $m_w < m$, иначе имеем случай полного события

[3]Функцию $G(x, \theta)$ Фишер [21] и назвал «$fiducialdistribution$» - фидуциальным (точнее «фидьюшиальным»), т.е. «доверительным» распределением. Поэтому правильнее будет говорить о *доверительных по Фишеру* в отличие от *доверительных по Нейману* распределениях и интервалах.

каждому отдельному параметру p_k(несмотря на то, что мультиномиальное распределение $M(m; p_1, p_2, ..., p_v)$является воспроизводящим по m.)

В случае **больших выборок** можно обобщить результаты аппроксимации нормальным распределением биномиального на мультиномиальное. Легко показать, что в рассматриваемом случае мультиномиальное распределение аппроксимируется распределением χ^2 с одной степенью свободы. Кроме того, $\frac{1}{mp(1-p)} = \frac{1}{mp} + \frac{1}{m(1-p).}$ Таким образом в многомерном случае мы будем иметь v величин

$$T_k^2 = \frac{(m_k - mp_k)^2}{mp_k}$$

таких, что $T_k \sim \mathcal{N}(0,1),\ k = \overline{1,v}$ и получим, что кратчайший многомерный доверительный интервал с надежностью η для оценивания величины $\bar{p}^* \in S_{v-1}$ будет представлять собой множество векторов $\hat{\bar{p}} = (\hat{p}_1, ..., \hat{p}_v)$ из $\mathbb{R}^v_{\geqslant 0}$ для которых

$$\sum_{k=1}^{v} \frac{(m_k - m\hat{p}_k)}{m\hat{p}_k} < \chi^2_\eta \tag{13}$$

где $\chi^2_\eta -$ квантиль уровня η распределения χ^2 с $v-1$ степенями свободы [16], [15]. Данная формула считается достаточно точной при $m\hat{p}_k > 5,\ k = \overline{1,v}$ или $m\hat{p}_k > 1$, когда доля таких $\hat{p}_k$ не менее $1/5$.

Приведенной зависимостью исчерпываются результаты по построению доверительных интервалов мультиномиально распределенной величины. С другой стороны ясно, что формула (13) крайне неудобна для практического использования.

Можно предложить пригодный для **малых выборок** численный метод построения симметричных относительно некоторой точечной оценки $\hat{p}_1, ..., \hat{p}_v$ интервалов вида

$$J_S = (\hat{p}_1 \pm \varepsilon_1, ..., \hat{p}_v \pm \varepsilon_v). \tag{14}$$

Здесь $(\varepsilon_1, ..., \varepsilon_v) \in (0,1)^v -$ точности доверительного определения соответствующих вероятностей.

Обозначим $E^i \triangleq (\varepsilon_1^i, ..., \varepsilon_v^i) \in (0,1)^v$. Согласно подходу Р.Фишера распределение вероятностей будет являться распределением Дирихле имеющее плотность

$$f(\bar{p}|d_1, d_2, ..., d_v) = \frac{\Gamma(d_1 + d_2 + ... + d_v)}{\Gamma(d_1)\Gamma(d_2)...\Gamma(d_v)} \prod_{k=1}^{v} p_k^{d_k - 1}$$

в любой точке симплекса $S_{v-1}(\bar{x})$ и равную нулю в других точках $\mathbb{R}^v$. Здесь все $d_1, d_2, ..., d_v -$ вещественные положительные числа. Тогда интервал указанного вида будет (фидуциальным) доверительным с достоверностью $\eta = \eta(E^i)$, если

$$\frac{(m+v-1)!}{m_1!...m_v!}\int\limits_{\hat{p}_1-\varepsilon_1}^{\hat{p}_1+\varepsilon_v}...\int\limits_{\hat{p}_v-\varepsilon_1}^{\hat{p}_v+\varepsilon_v} x_1^{m_1}...x_v^{m_v}dx_1...dx_v=\eta(E^i). \tag{15}$$

Здесь, естественно, $\sum_{k=1}^{v} m_k = m$ и

$$x\in S_{v-1}(\bar{x})=\left\{(x_1,x_2,...,x_v): x_k\geqslant 0, k=\overline{1,v}; \sum_{k=1}^{v} x_k=1\right\}$$

(см. п.5.2. в [6]).

Конечно, такие интервалы не будут являться кратчайшими ни с какой точки зрения, однако они исключительно удобны в использовании на практике. Интеграл в (15) может быть вычислен численно для разных наборов $E^i, \varepsilon_k^{i+1}\leqslant\varepsilon_k^i,\ k=\overline{1,v},\ i=1,2,....$ При этом значение достоверности будет уменьшаться. Можно остановиться на значении $\eta(E^i)$ не меньшем некоторого выбранного.

Распространяя метод Неймана на многомерный случай можно предложить находить величины $p_{1,-},...,p_{v,-}$ и $p_{1,+},...,p_{v,+}$ численно решая соответственно первое и второе уравнение системы

$$\frac{(m+v-1)!}{m_1!...m_v!}\int\limits_{0}^{p_1}...\int\limits_{0}^{p_v} x_1^{m_1}...x_v^{m_v}dx_1...dx_v=\begin{cases}1-P=\dfrac{1-\eta}{2},\\ P=\dfrac{1+\eta}{2}.\end{cases}$$

Здесь также $x\in S_{v-1}(\bar{x})$.

Доверительный интервал достоверности η будет при этом иметь вид

$$J_N=(p_{1,-}\leqslant p_1\leqslant p_{1,+}...,p_{v,-}\leqslant p_v\leqslant p_{v,+}).$$

6.2. **Байесовский подход.** Рассмотрим сразу **многомерный случай.**

Интервальное байесовское оценивание тесно связано с фидуциальными распределениями [9], что, впрочем, можно заключить из определения фидуциального распределения в п.6.1.1. В частности, используя апостериорную плотность

$$f(\bar{p}|d_1,...,d_v)=\frac{\Gamma(d_1+...+d_v+m)}{\Gamma(d_1+m_1)...\Gamma(d_v+m_v)}\prod_{k=1}^{v}p_k^{d_k+m_k-1},$$

получаем, что при априорном распределении $D_i(d_1,...,d_v)$ байесовский доверительный интервал (14) достоверности $\eta(E^i)$ при точечных оценках $\hat{p}_1,...,\hat{p}_v$ должен удовлетворять соотношению

$$\frac{\Gamma(d_1... + d_v + m)}{\Gamma(d_1 + m_1)...\Gamma(d_v + m_v)} \int\limits_{\hat{p}_1-\varepsilon_v}^{\hat{p}_1+\varepsilon_1} ... \int\limits_{\hat{p}_v-\varepsilon_v}^{\hat{p}_v+\varepsilon_v} x_1^{d_1+m_1-1}...x_v^{d_v+m_v-1}dx_1...dx_v = \eta(E^i). \qquad (16)$$

В случае равномерного априорного распределения получим формулу (15).

При неравных весах прецедентов, повторяя рассуждения п.5.2.4 из [6], посвященному случаю неравных весов прецедентов, вместо (16) получим формулу

$$\frac{\Gamma(d_1... + d_v + M)}{\Gamma(d_1 + \mu_1)...\Gamma(d_v + \mu_v)} \int\limits_{\hat{p}_1-\varepsilon_v}^{\hat{p}_1+\varepsilon_1} ... \int\limits_{\hat{p}_v-\varepsilon_v}^{\hat{p}_v+\varepsilon_v} x_1^{d_1+\mu_1-1}...x_v^{d_v+\mu_v-1}dx_1...dx_v = \eta(E^i). \qquad (17)$$

где μ_k, M и $\hat{p}_k, k = \overline{1,v}$ определяются по формулам

$$\sum_{i:x_i\in X_k} \gamma_i = \mu_k, \qquad M = \sum_{k=1}^{v} \mu_k, \qquad \hat{p}_k = \frac{\mu_k + 1}{M + v}, \qquad k = \overline{1,v}.$$

Для равномерного априорного распределения формула (17) превращается в (15) с заменой m_k на μ_k и m на $M, k = \overline{1,v}$. Это, фактически, распространение метода Большева на многомерный случай.

Уравнения (16) и (17) можно решать численно тем же методом, что и (15).

Одним из направлений байесовского подхода является т.н. *эмпирический байесовский метод* рассматривающий построение оценок в условиях неизвестного априорного распределения [5]. В рамках указанного метода можно предложить следующий <<комплексный>> метод доверительного оценивания.

В [6] указывалось, что в одномерном случае при малых p^* в качестве априорного распределения может быть использовано распределение $Be(1, b)$ с большим b приводящее к байесовской оценке

$$\hat{p} = \frac{m_w + 1}{m + b + 1}.$$

Как обосновано выбрать значение b? Заметим, что при $m_w \neq 0$ и $b = m_r/m_w = m/m_w - 1$ данная оценка будет совпадать с МП-оценкой, а при $m > 1 - b \geqslant 1$. Это дает основание в указанных условиях принять за априорное распределение $Be(1, m/m_w - 1)$. Тогда можно предложить определять верхнюю границу одностороннего доверительного интервала $J = (0, p_+)$ достоверности η из условия

$$\frac{\Gamma(m + m/m_w)}{\Gamma(m_w + 1)\Gamma(m_r + m/m_w - 1)} \int\limits_0^{p_+} x^{m_w}(1 - x)^{m_r+m/m_w-2}dx = \eta,$$

что эквивалентно

$$I_{p+}(m_w + 1, m_r + m/m_w - 1) = \eta. \tag{18}$$

Заметим, что здесь второй параметр неполной бетта-функции I_{p+} не обязательно целочисленный.

Данное уравнение можно решать используя таблицы неполной **B** - функции или используя связь **B** - функции с **F** - распределением. В последнем случае получим

$$p_+ = \frac{(m_w + 1)F_{v_1,v_2}}{m_r + m/m_w - 1 + (m_w + 1)F_{v_1,v_2}},$$

где $F_{v_1,v_2}-$ квантиль **F** - распределения $v_1 = 2(m_w + 1), v_2 = 2(m_r + m/m_w - 1)$ степенями свободы для доверительной вероятности ошибки $1 - \eta$.

7. Сравнение оценок, полученных различными методами

В качестве примера приведем оценки вероятностей событий, полученные различными методами для одномерного случая. При этом возьмем значения m, при которых неприменимы аппраксимационные методы.

Пусть $m_1 = 8, m_2 = 10, m_3 = 15$ и $m_w = 1$ во всех трех случаях. Номер варианта будем обозначать верхним индексом соответствующей оценки.

Точечные оценки. МП-оценки равны

$$\hat{p}^1_{ML} = \frac{1}{8} = 0.125, \qquad \hat{p}^2_{ML} = \frac{1}{10} = 0.100, \qquad \hat{p}^3_{ML} = \frac{1}{15} \approx 0.067.$$

Байксовские оценки суть

$$\hat{p}^1_B = \frac{2}{10} = 0.2, \qquad \hat{p}^2_B \approx 0.167, \qquad \hat{p}^3_B \approx 0.118.$$

Медианные оценки можно определить по Таблице 3.4 [2] квантилей уровня $P = 0.5$ **B** - распределения:

$$\hat{p}^1_m \approx 0.190, \qquad \hat{p}^2_m \approx 0.148, \qquad \hat{p}^3_m \approx 0.103.$$

Мы видим, что медианные оценки ближе к байесовским, чем к МП-оценкам.

Интервальные оценки. В [12] затабулированы экстремальные решения неравенства (3). По таблице 1. В находим:

$$J^1 = (0.006, 0.500), \qquad J^2 = (0.005, 0.446), \qquad J^3 = (0.003, 0.302) \text{ для } \eta = 0.95.$$

Решение (3) использующее замену биномиального распределения на **F** - распределение (см. [15]) дает

$$J^1 = (0.104, 0.526), \qquad J^2 = (0.025, 0.445), \qquad J^3 = (0.105, 0.319) \qquad \text{для } \eta = 0.95.$$

Решая уравнения (9) Клоппера-Пирсона по формулам (10) (заменой биномиального на **F** - распределение), получим

$$J^1 = (0.125, 0.526), \qquad J^2 = (0.100, 0.445), J^3 = (0.007, 0.319) \text{ для } \eta = 0.95,$$

что очень близко к предыдущим решениям.

Графическим методом решения уравнений (9) Клоппера-Пирсона более-менее уверенно можно определить лишь оценку для второго случая:

$$J^2 = (0.02, 0.32) \text{ для } \eta = 0.9 \text{ и } J^2 = (0.01, 0.46) \text{ для } \eta = 0.95.$$

Неймановские оценки найдем по таблице 5.2 [2] Доверительных пределов для параметра p биномиального распределения:

$$J_N^1 = (0.006, 0.417), \qquad J_N^2 = (0.005, 0.394), \qquad J_N^3 = (0.003, 0.279)$$

и

$J_N^1 = (0.003, 0.527), \qquad J_N^2 = (0.003, 0.445), \qquad J_N^3 = (0.002, 0.319)$

для $\eta = 0.9 \quad (P = 0.95)$ и $\eta = 0.95 \quad (P = 0.975)$ соответственно. Видно, что неймановские наиболее селективные интервалы не суть кратчайшие доверительные.

В заключение уместно привести слова из фундаментальной монографии [9]: если результаты различных подходов не совпадают, то «основная причина различия не в том, что тот или иной поход не верен, а в том, что они, сознательно или не сознательно, либо отвечают на разные вопросы, либо основываются на разных постулатах».

Список литературы

1. Бернштейн С.Н. О «доверительных» вероятностях Фишера / С.Н.Бернштейн. Собрание сочинений. - М.: Наука, 1964. - Т. IV: Теория вероятностей и математическая статистика (1911 - 1946). - С.. 386-393.
2. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. - М.: Наука, 1983. - 416 с.
3. Большев Л.Н. О построении доверительных пределов // Теория вер.-ти и ее применен. - 1965. - Т.X, вып.1. - С.197-192.
4. Большев Л.Н. Комментарий к работе С.Н.Бернштейна «О доверительных вероятностях Фишера»/С.Н.Бернштейн. Собрание сочинений. - М.: Наука, 1964. - Т. IV: Теория вероятностей и математическая статистика (1911-1946). - С.566-569.
5. Большев Л.Н. Приложения эмпирического байесовского подхода / Международный конгресс математиков в Ницце 1970. Доклады советских математиков. М. - 1972. - С. 48 - 55.
6. Гуров С.И. Как оценить надежность алгоритма классификации // Таврический вестник информатики и математики. - Симферополь: КНЦ НАН Украины. - 2002. - №1. - С.27 - 56.
7. Закс Л. Статистическое оценивание - М.: Статистика, 1976. - 560 с.
8. Закс Ш. Теория статистических выводов - М.: Мир, 1975. - 776 с.
9. Кендал М., Стюарт А. Статистические выводы и связи - М.: Наука, 1973. - 900 с.
10. Кремер Н.Ш. Теория вероятностей и математическая статистика. - М.: ЮНИТИ - ДАНА, 2000. - 543 с.
11. Леман Э. Проверка статистических гипотез. - М.: Наука, 1970. - 408 с.
12. Мюллер П., Нойман П., Шторм Р. Таблицы по математической статистике - М.: Финансы и статистика, 1982 - 278 с.

13. Нейман Ю. Статистическая оценка как проблема классической теории вероятностей // Успехи матем. наук. - 1944. - Т.10. - С. 207-229.
14. Оуэн Д.Б. Сборник статистических таблиц. - М.: ВЦ РАН, 1966. - 186 с.
15. Поллард Дж. Справоченик по вычислительным методам статистики - М. : Финансы и статистика, 1982. - 344 с.
16. Уилкс С. Математическая статистика - М.: Наука, 1967. - 632 с.
17. Фукунага К. Введение в статистическую теорию распознавания образов - М.: Наука, 1979. - 368 с.
18. Shmetter Л. Введение в математиескую статистику. - М.:Наука, 1976. - 520 с.
19. Янко Я. Математико-статистические таблицы. - М.:Госстатиздат, 1961. - 552 с.
20. Clopper C.J., Pearson E.S. The use of confidence or fiducial limits illustrated in the case of the binomial // Biometrika. - 1934. - Vol.26. - P. 404-413.
21. Fisher R.A. The fiducial argument in statistical inference // Annals of Eugenics. - 1935. - Vol.5. - P. 391 - 398.
22. Neyman J. Outline of a theory of statistical estimation based on the classical theory of probability // Philos. Trans. Roy. Soc. London. Ser. A. - 1937. - Vol. 236. - P.333-380.
23. Pearson E.S., Hartlay H.O. Biometrika tables for ststisticians, I, II. - Cambridge. - 1966/72.