С.И. Гуров

ОЦЕНКИ НАДЁЖНОСТИ АЛГОРИТМОВ КЛАССИФИКАЦИИ. III. ИНТЕГРАЛЬНЫЕ ОЦЕНКИ[1]

Данная статья завершает рассмотрение оценок надежности классифицирующих алгоритмов, начатое в [6] и [7].

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.

Первый 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.

[1]Работа выполнена при финансовой поддержке РФФИ (коды проектов 01-01-00885 и 04-01-00161).

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, если среди m прецедентов имеется m_w неправильно распознанных построенным р.п. Доверительный интервал оценивания записывает в виде $J = (p_-, p_+)$.

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

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

и *функцией выживания*

$$\mathrm{P}_p\{m_w > t\} = 1 - P\,(t-1) = \sum_{k=t}^{m} \binom{m}{k} p^k (1-p)^{m-k}\,. \quad (2)$$

В этих формулах $p \in (0,1)$ и $t = 0, 1, \ldots, m$ и должно выполняться условие

$$\mathrm{P}\,\{\, p^* \notin J \,\} = \mathrm{P}_{p_+}\{\, m_w \leqslant t_1\} + \mathrm{P}_{p_-}\{\, m_w > t_2\} \leqslant 1 - \eta = \alpha\,, \quad (3)$$

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 значений биномиальных вероятностей или функции распределения (1) является весьма трудоемкой процедурой. Поэтому во всех случаях, когда 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 [9]. Замена базируется на том факте, что первая производная логарифма функции правдоподобия L распределена асимптотически нормально со средним, равном нулю и дисперсией

$$\mathbf{D}\left(\frac{\partial \ln L}{\partial p}\right) = \mathbf{M}\left\{\left(\frac{\partial \ln L}{\partial p}\right)^2\right\} = -\mathbf{M}\left\{\frac{\partial^2 \ln L}{\partial p^2}\right\}. \tag{4}$$

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

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

имеет асимптотически стандартное нормальное распределение $\mathcal{N}(0,1)$ т.е.

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

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

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

Таким образом приходим к уравнению для границ искомого интервала

$$p^2\left(1 + \frac{z_\eta^2}{m}\right) - p\left(2\hat{p} + \frac{z_\eta^2}{m}\right) + \hat{p}^2 = 0. \tag{7}$$

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 $c = \frac{1}{1+z_\eta^2/m}$[2].

[2]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 (9)).

Решая квадратное уравнение (7), получаем

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

где величина z_η находится из уравнения

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

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

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

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

$$p_+ = 1 - \sqrt[m]{\alpha}, \tag{8}$$

соответственно, для полного события $\hat{p} = 1$ точная нижняя граница есть $p_- = \sqrt[m]{\alpha}$. Для $\alpha = 0.95$, $m > 50$ 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 $Po(k;\lambda) = \lambda^k \exp(-\lambda)/k!$ с $\lambda = mp$:

$$\mathrm{P}\{m_w \leqslant t \,|\, m, p\} \approx \sum_{k=0}^{t} \frac{(m\,p)^k}{k!} e^{-mp}. \tag{9}$$

Далее можно воспользоваться методами доверительного (одностороннего $(0, \lambda^s_{\eta,+})$ или двустороннего $(\lambda_{\eta,-}, \lambda_{\eta,+})$) оценивания пуассоновского параметра λ при достоверности η [15], [18]и затем определить интервал для p : $(J = (0,\ \lambda^{s}_{\eta,+}/m)$ или $J = (\lambda_{\eta,-}/m,\ \lambda_{\eta,+}/m))$ соответственно.

При невозможности использования аппроксимационных формул (подробный перечень предположений для их применения дан в [5], а для (6) – и в [9]) 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)

$$\alpha_1 = \mathrm{P}_{p_-}\{m_w \leqslant t_1\} \geqslant 0, \quad \alpha_2 = \mathrm{P}_{p_+}\{m_w > t_2\} \geqslant 0,$$
$$\alpha_1 + \alpha_2 = \alpha = 1 - \eta,$$

можно получать различные доверительные интервалы. При $\alpha_1 = \alpha_2 = \alpha/2$ соответствующий интервал J называется *центральным*[3].

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 [10], [11], [14] 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 доверительный интервал даже укорачивается. Это объясняется тем, что потери от новой неопределенности, связанной с введением случайной величины оказываются меньше устраненных потерь от неопределенности, связанной с неравенством.

В [24], [27] предложено использовать центральные интервалы, т.е. разделить вероятность интервальной недооценки и переоценки поровну и находить p_+, t_1 и p_-, t_2 из условий

$$\begin{cases} t_1 = \arg\max\limits_{0 \leqslant t \leqslant m}\left\{\mathrm{P}_{p_+}\{m_w \leqslant t\}\right\} = \dfrac{1-\eta}{2}, \\[2ex] t_2 = \arg\min\limits_{0 \leqslant t \leqslant m}\left\{\mathrm{P}_{p_-}\{m_w > t\}\right\} = \dfrac{1-\eta}{2} \end{cases} \tag{10}$$

[3]Заметим, что центральность интервала не означает, что его границы равноотстоят от выборочной статистики.

соответственно. Здесь полученные значения t_1 и 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 (10) известны под названием формул Клоппера-Пирсона. Прямой метод их решения основанный на переборе значений $t = 0,\ 1,\ \ldots$ кратко описан в [22] и [11]. Однако соответствующий алгоритм, очевидно, достаточно трудоемок. Поэтому более удобно воспользоваться другим методом решения (10), основанным на использовании известной связи между величиной $B(t,m,p) = \mathrm{P}_p\{m_w \leqslant t\}$ в (1) и функцией F-распределения случайной величины U_{ν_1,ν_2} с ν_1 и ν_2 степенями свободы [15]:

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

С помощью указанных соотношений определяются точные формулы для границ центрального доверительного интервала. Они имеют следующий вид [9]:

$$\begin{cases} p_- = \dfrac{m_w}{m_w + (m - m_w + 1)F_{\nu_1^-,\nu_2^-}}, \\[2ex] p_+ = \dfrac{(m_w+1)F_{\nu_1^+,\nu_2^+}}{m - m_w + (m_w+1)F_{\nu_1^+,\nu_2^+}}, \end{cases} \tag{11}$$

где $F_{\nu_1^-,\nu_2^-}$ и $F_{\nu_1^+,\nu_2^+}$ – квантили F-распределения с $\nu_1^- = 2(m - m_w + 1)$, $\nu_2^- = 2m_w$ и $\nu_1^+ = 2(m_w + 1)$, $\nu_2^+ = 2(m - m_w)$ 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

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

будет равняться $\frac{1-\eta}{2}$ и $\frac{1+\eta}{2}$ соответственно.

Для быстрого приближенного решения уравнений (10) со значениями достоверности $\alpha = 0.1;\ 0.05$ и объемов выборки

$m = 10, \ldots, 1000$ построены графические зависимости между наблюдаемыми значениями $\hat{p}$ и относительными частотами генеральной совокупности, определяющими доверительный интервал (см. например, [9], [13], [21]). 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 (8).

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 (11), определенная для доверительной вероятности $1 - \eta$. Заметим, что при $\hat{p} = 0$ полученная оценка совпадет с (8).

Наиболее селективные интервалы. Дж. Нейман предложил метод построения доверительных интервалов, которые также назвал "кратчайшими" [26], [16]. В тоже время они построены на совершенно иной идее. Чтобы отличать их от рассмотренных кратчайших доверительных интервалов, неймановские интервалы J_N в [11] предложено называть ***наиболее селективными*** (там же см. обсуждение различий между кратчайшими доверительными и наиболее селективными интервалами). Последние, в отличие от ранее рассмотренных кратчайших доверительных интервалов J с данной достоверностью η минимизируют не свою длину в условии $\theta^* \in J$, а вероятность $\theta \in J_N$, $\theta \neq \theta^*$. В силу этого ясно, что они не обязательно являются кратчайшими в прямом смысле этого слова. Однако оказалось, что "селективный" подход, связанный с желанием исключить из интервала J_N как можно больше ложных значений θ, так чтобы ошибка принятия неверного значения была бы минимальной, оказался в общем случае значительно более простым и удобным. Кроме того, именно такой подход используется в теории проверки статистических гипотез (где указанная выше ошибка является *ошибкой второго рода*).

Наиболее селективные интервалы рассмотрены подробно в [14]. Там же описан 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

$$G(T, \theta) = \begin{cases} 1 - P\,, \\ P\,. \end{cases} \tag{12}$$

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 [26], [14], [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 $Bi\,(m, p)$ в (1)

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

Функцией распределения вероятностей ***B***-распределения является неполная ***B***-функция, обозначаемая $I_p(a,\, b)$, $(0 \leqslant p \leqslant 1,\ a > 0,\, b > 0)$:

$$I_p\,(a,\, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int\limits_0^p x^{a-1}(1-x)^{b-1}dx\,.$$

Неполная ***B***-функция обладает свойством

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

а для целых a и b имеют место замечательные равенства

$$\begin{cases} I_p\,(a,\, b) = \displaystyle\sum_{k=a}^{a+b-1} \binom{a+b-1}{k} p^k (1-p)^{a+b-1-k}\,, \\[2ex] I_{1-p}\,(b,\, a) = \displaystyle\sum_{k=b}^{\infty} \binom{a+k-1}{a-1} p^a (1-p)^k\,. \end{cases}$$

Используя данную зависимость показать, что

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

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

Ясно, что пользоваться хорошо изученной и затабулированной неполной $\boldsymbol{B}$-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 (12), 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 $m_w + 1$[4]. При этом верхняя граница оказывается несколько завышенной, что, естественно, компенсируется большей вероятностью накрытия истинного значения p^*.

В результате [5] 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

$$\begin{cases} I_{p_-}(m_w,\, m - m_w + 1) = 1 - P = \dfrac{1-\eta}{2}\,, \\[2ex] I_{p_+}(m_w + 1,\, m - m_w) = P = \dfrac{1+\eta}{2}\,. \end{cases} \tag{13}$$

Для решения уравнений (13) можно воспользоваться таблицами $\boldsymbol{B}$-распределения (см., например, [17], [15]). В [23], [5], [15]

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

границы доверительного интервала J_N для биномиального распределения табулированы. Заметим, что метод остается пригодным и когда неизвестный параметр рассматривается как случайный.

Фидуциальные интервалы. Теория фидуциальных интервалов Р. Фишера [25], [12] базируется на следующих соображениях.

Пусть x – случайная величина с функцией распределения $G(x, \theta)$, где θ --- 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[5]. "Фидуциальное распределение не является распределением вероятности в смысле частотной теории. Это новое понятие, выражающее интенсивность нашей веры в различные возможные значения параметра" [11]. Вопрос заключается в том, когда $G(x, \theta)$ действительно можно рассматривать как распределение вероятности в частотном смысле. Фишер считал, что всегда. Однако оказалось, что такое утверждение является ошибочным. Подробный разбор условий, при выполнении которых на указанный вопрос можно все же ответить положительно см., например в [2], [1], [3]. При этом оказывается, что данные условия не являются слишком жесткими, в силу чего доверительные фишеровские и неймановские распределения, как правило, совпадают, и отличие (принципиальное!) их исходных концепций заключается лишь в истолковании результатов. Основными из упомянутых условий являются непрерывность распределения $G(x, \theta)$ по x и (строго) монотонная зависимость $G(x, \theta)$ от θ. Ясно также, что вместо x можно рассматривать некоторую статистику T, рассматриваемую в качестве точечной оценки θ. В таком случае пределы интегрирования $G(x, \theta)$ по θ могут быть приняты за границы доверительных интервалов, а значения этих интегралов – за достоверность соответствующей интервальной оценки и теория фидуциального оценивания приводит к тем же результатам, что и теория доверительного оценивания.

Для нашей задачи первое из упомянутых условий можно обойти, используя, например, рандомизацию или переход от m_w к $m_w + 1$ (см.

[5]Функцию $G(x, \theta)$ Фишер и назвал "fiducial distribution" – фидуциальным (точнее "фидьюшиальным"), т.е. "доверительным" распределением. Поэтому правильнее будет говорить о *доверительных по Фишеру* в отличие от *доверительных по Нейману* распределениях и интервалах.

выше). Также имеет место (строго убывающая) монотонная зависимость функции биномиального распределения вероятностей (1) от p (второе условие). Таким образом, фишеровские интервалы в нашем случае (как и в подавляющем числе практически интересных случаев) совпадают с неймановскими.

Метод доверительных интервалов Л.Н. Большева [3] совмещает фишеровский фидуциальный подход (в тех случаях, когда он применим) с построением наиболее 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 каждому отдельному параметру p_k (несмотря на то, что мультиномиальное распределение $M(m; p_1, p_2, \ldots, p_v)$ 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. Действительно, величина T^2 из (5) подчиняется распределению 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 $\widehat{\bar{p}} = (\hat{p}_1, \ldots, \hat{p}_v)$ из $\mathbb{R}^v_{\geqslant 0}$ для которых

$$\sum_{k=1}^{v} \frac{(m_k - m\hat{p}_k)^2}{m\hat{p}_k} < \chi^2_\eta, \qquad (14)$$

где χ^2_η – квантиль уровня η распределения χ^2 с $v-1$ степенями свободы [20], [18]. 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 (14) крайне неудобна для практического использования.

Заметим, что распределение χ^2 допускает аппроксимацию многомерным нормальным распределением [19].

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

$$J_S = (\hat{p}_1 \pm \varepsilon_1, \ldots, \hat{p}_v \pm \varepsilon_v)\,. \tag{15}$$

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

Обозначим $E^i = (\varepsilon_1^i, \ldots, \varepsilon_v^i) \in (0, 1)^v$. Согласно подходу Р. Фишера распределение вероятностей будет являться распределением Дирихле и интервал указанного вида будет (фидуциальным) доверительным с достоверностью $\eta = \eta(E^i)$, если

$$\frac{(m+v-1)!}{m_1! \ldots m_v!} \int\limits_{\hat{p}_1-\varepsilon_1}^{\hat{p}_1+\varepsilon_1} \ldots \int\limits_{\hat{p}_v-\varepsilon_1}^{\hat{p}_v+\varepsilon_v} x_1^{m_1} \ldots x_v^{m_v}\, dx_1 \ldots dx_v = \eta(E^i)\,. \tag{16}$$

Здесь, естественно, выполняется условия нормировки $\sum_{k=1}^{v} m_k = m$ и $(x_1, x_2, \ldots, x_v) \in S_{v-1}$, где

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

Конечно, такие интервалы не будут являться кратчайшими ни с какой точки зрения, однако они исключительно удобны в использовании на практике. Интеграл в (16) может быть вычислен численно для разных наборов E^i, $\varepsilon_k^{i+1} \leqslant \varepsilon_k^i$, $k = \overline{1, v}$, $i = 1, 2, \ldots$ 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 $p_{1,-}, \ldots, p_{v,-}$ и $p_{1,+}, \ldots, p_{v,+}$ численно решая соответственно первое и второе уравнение системы

$$\frac{(m+v-1)!}{m_1!\dots m_v!}\int\limits_0^{p_1}\dots\int\limits_0^{p_v} x_1^{m_1}\dots x_v^{m_v}\,dx_1\dots dx_v \;=\; \begin{cases} 1-P=\dfrac{1-\eta}{2}\,, \\ P=\dfrac{1+\eta}{2}\,. \end{cases}$$

Здесь, конечно, $(x_1,\,x_2,\,\dots\,,\,x_v)\in S_{v-1}$.

Доверительный интервал достоверности η будет при этом иметь вид

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

Доверительная теория "достигает общности ценой того, что оказывается неспособной включать априорные" знания в свои утверждения [11].

6.2 Байесовский подход

Рассмотрим сразу многомерный случай.

Интервальное байесовское оценивание тесно связано с фидуциальными распределениями [11]. В частности, получаем (см. [7]), что при априорном распределении $Di(d_1,\,\dots\,,\,d_v)$ байесовский доверительный интервал (15) достоверности $\eta(E^i)$ при точечных оценках $(\hat p_1,\,\dots\,,\,\hat p_v)$ должен удовлетворять соотношению

$$\frac{\Gamma(d_1+\ldots+d_v+m)}{\Gamma(d_1+m_1)\ldots\Gamma(d_v+m_v)}\int\limits_{\hat p_1-\varepsilon_1}^{\hat p_1+\varepsilon_1}\dots\int\limits_{\hat p_v-\varepsilon_v}^{\hat p_v+\varepsilon_v} x_1^{d_1+m_1-1}\dots \tag{17}$$
$$\dots x_v^{d_v+m_v-1}\,dx_1\dots dx_v \;=\; \eta(E^i)\,.$$

В случае равномерного априорного распределения получим формулу (16).

При неравных весах прецедентов, повторяя рассуждения последнего раздела [7], вместо (17) получим формулу

$$\frac{\Gamma(d_1+\ldots+d_v+M)}{\Gamma(d_1+\mu_1)\ldots\Gamma(d_v+\mu_v)}\int\limits_{\hat p_1-\varepsilon_1}^{\hat p_1+\varepsilon_1}\dots\int\limits_{\hat p_v-\varepsilon_v}^{\hat p_v+\varepsilon_v} x_1^{d_1+\mu_1-1}\dots \tag{18}$$
$$\dots x_v^{d_v+\mu_v-1}\,dx_1\dots dx_v \;=\; \eta(E^i)\,.$$

где μ_k, M, и $\hat{p}_k$, $k = \overline{1,\, v}$ определяются по соответствующим формулам последнего раздела [7].

Для равномерного априорного распределения формула (18) превращается в (16) с заменой m_k на μ_k и m на M, $k = \overline{1,\, v}$. Это, фактически, распространение метода Большева на многомерный случай.

В уравнениях (17) – (18) выполняется естественное ограничение $(x_1,\, x_2,\, \ldots\,,\, x_v) \in S_{v-1}$; их можно решать численно тем же методом, что и (16).

Одним из направлений байесовского подхода является т.н. *эмпирический байесовский метод* рассматривающий построение оценок в условиях неизвестного априорного распределения [4]. В рамках указанного метода можно предложить следующий "комплексный" метод доверительного оценивания.

Как указывалось в [7], в одномерном случае при малых p^* в качестве априорного распределения может быть использовано распределение $Be(1,\, b)$ с большим b приводяшее к байесовской оценке

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

Как обоснованно 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 (см. [8]) 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

$$I_{p_+}\left(m_w + 1,\, m_r + m/m_w - 1\right) \;=\; \eta\,. \qquad (19)$$

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

Данное уравнение можно решать используя таблицы неполной $\boldsymbol{B}$-функции или используя связь $\boldsymbol{B}$-функции с $\boldsymbol{F}$-распределением. В последнем случае используя (11) получим

$$p_+ \;=\; \frac{(m_w + 1)F_{\nu_1,\,\nu_2}}{m_r + m/m_w - 1 + (m_w + 1)F_{\nu_1,\,\nu_2}}\,,$$

где F_{ν_1,ν_2} – квантиль F-распределения со степенями свободы $\nu_1 = 2(m_w + 1)$, $\nu_2 = 2(m_r + m/m_w - 1)$ для доверительной вероятности ошибки $1 - \eta$.

В заключение автор выражает глубокую признательность академику РАН Ю.И. Журавлеву за понимание и поддержку, а также проф. В.Е. Бенингу и к.ф.-м.н. К.В. Воронцову, прочитавших всю работу и сделавших ценные замечания по ее улучшению.

Литература

1. Бернштейн С.Н. О "доверительных" вероятностях Фишера /С.Н. Бернштейн. Собрание сочинений. Том IV. Теория вероятностей и математическая статистика (1911-1946). – М.: Наука, 1964. – С. 386--393

2. Большев Л.Н. Комментарий к работе С.Н. Бернштейна "О "доверительных" вероятностях Фишера" / С.Н. Бернштейн. Собрание сочинений. Том IV. Теория вероятностей и математическая статистика (1911-1946). – М.: Наука, 1964. – С. 566--569.

3. Большев Л.Н. О построении доверительных пределов // Теория вер.-ти и ее применен. 1965, том X, вып. 1. – С. 197-192.

4. Большев Л.Н. Приложения эмпирического байесовского подхода / Международный конгресс математиков в Ницце 1970. Доклады советских математиков. М., 1972, С. 48-55.

5. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. – М.: Наука, Гл. ред. физ.-мат. лит., 1983.

6. Гуров С.И. Оценки надежности алгоритмов классификации I. Введение. Точечные оценки.

7. Гуров С.И. Оценки надежности алгоритмов классификации II. Точечные оценки (продолжение).

8. Гуров С.И. Принцип согласованности и бейесовское интервальное оценивание // Таврический вестник информатики и математики. Вып. 3, 2003. – Симферополь: КНЦ НАН Украины (в печати).

9. Закс Л. Статистическое оценивание / Пер. с нем. под ред. Ю.П. Адлера, В.Г. Горского. – М.: Статистика, 1976.

10. Закс Ш. Теория статистических выводов: Пер. с англ./ Под ред. Ю.К. Беляева. – М.: Мир, 1975.

11. Кендал М., Стюарт А. Статистические выводы и связи / Пер. с англ. – М.: Наука, 1973.

12. Климов Г.П. Инвариантные выводы в статистике. – М.: Наука, Гл. ред. физ.-мат. лит., 1973.

13. Кремер Н.Ш. Теория вероятностей и математическая статистика. – М.: ЮНИТИ-ДАНА, 2000.

14. Леман Э. Проверка статистических гипотез. – М.: Наука, Гл. ред. физ.-мат. лит., 1979.

15. Мюллер П., Нойман П., Шторм Р. Таблицы по математической статистике /Пер. с нем. и предисл. В.М. Ивановой. – М.: Финансы и статистика, 1982.

16. Нейман Ю. Статистическая оценка как проблема классической теории вероятностей // Успехи матем. наук., т. 10, 1944, с. 207-229.

17. Оуэн Д.Б. Сборник статистических таблиц. – М.: ВЦ РАН, 1966.

18. Поллард Дж. Справочник по вычислительным методам статистики /Пер. с англ. В.С. Зандворова; Под. ред. и с прдисл. Е.М. Четыркина. – М.: Финансы и статистика, 1982.

19. Тихомирова М.И., Гостяков В.П. Нормальное приблежение многомерного χ^2 распределения // Труды по дискретной математике. Том 4. М.: Физматлит, 2001. – С. 259--272.

20. Уилкс С. Математическая статистика /Пер. с англ. – М.: Наука, 1967.

21. Фукунага К. Введение в статистическую теорию распознавания образов /Пер. с англ. – М.: Наука, Гл. ред. физ.-мат. лит., 1979.

22. Шметтер Л. Введение в математическиую статистику. – М.: Наука, 1976.

23. Янко Я. Математико-статистические таблицы. – М.: Госстатиздат, 1961.

24. Clopper C.J., Pearson E.S. The use of confidence or fiducial limits illustrated in the case of the binominal //Biometrika 26 (1934), 404-413.

25. Fisher R.A. The fiducial argument in statistical inference // Annals of Eugenics, vol. 5, 1935. 391-398.

26. Neyman J. Outline of a theory of statistical estimation based on the classical theory of probability // Philos. Trans. Roy. Soc. London. Ser. A., 1937, v. 236, p. 333-380.

27. Pearson E.S., Hartlay H.O. Biometrika tables for ststisticians, I, II. – Cambridge, 1966/72.