С.И. Гуров

ОЦЕНКИ НАДЁЖНОСТИ АЛГОРИТМОВ КЛАССИФИКАЦИИ. II. ТОЧЕЧНЫЕ БАЙЕСОВСКИЕ ОЦЕНКИ[1]

В данной статье продолжается рассмотрение точечных оценок надёжности классифицирующих алгоритмов, начатое в [4].

4.2 Байесовский подход

Байесовские точечные оценки $\widehat{\bar{p}}_W$ получаются как решения задачи минимизации функционала среднего риска записываемой как

$$\int\limits_{S_{v-1}(\bar{p})} W(\bar{p},\, \bar{q}) f(\bar{p}\,|\, m_1, m_2, \dots m_v)\, d\bar{p} \;=\; R(\bar{q})\,,$$

$$\widehat{\bar{p}}_W \;=\; \arg \min_{\bar{q} \in S_{v-1}(\bar{x})} R(\bar{q})\,.$$

Здесь и далее

- $S_{v-1}(\bar{x}) = \{(x_1,\, x_2,\, \dots\,,\, x_v) : x_k \geqslant 0,\; k = \overline{1, v};$ $\sum_{k=1}^{v} x_k = 1\}$ – $(v-1)$-мерный симплекс в пространстве $\mathbb{R}^v$;
- $\bar{p}$, $\bar{q}$, $\widehat{\bar{p}}_W$ – векторы из $S_{v-1}(\bar{x})$, причем последний – вектор оценок вероятностей при данной функции потерь W;
- $W(\bar{p},\, \bar{q}) : S_{v-1}(\bar{x}) \times S_{v-1}(\bar{x}) \to \mathbb{R}_{\geqslant 0}$ – функция потерь для выбранных значений $\bar{q}$, когда $\bar{p}$ суть истинные значения искомых вероятностей;
- $f(\bar{p}\,|\, m_1, m_2, \dots m_v)$ – апостериорная плотность вероятности вектора $\bar{p}$ при наблюдённых значениях $m_1, m_2, \dots m_v$ попадания прецедентов в соответствующие области пространства образов.

Решение данной задачи в значительной мере определяется видом функции потерь.

Как отмечалось, "простая" функция потерь приводит к методу максимизации апостериорной вероятности, которая при использовании принципа неопределённости Лапласа даёт, как мы видели, полученную ранее в рамках частотного подхода МП-оценку.

[1]Работа выполнена при финансовой поддержке РФФИ (код проекта 04-01-00161)

Практически используют либо квадратичную

$$W(\bar{p}, \bar{q}) = c(\bar{p}) \parallel \bar{p} - \bar{q} \parallel^2,$$

либо нормированную квадратичную функцию потерь

$$W(\bar{p}, \bar{q}) = c(\bar{p}) \frac{\|\bar{p} - \bar{q}\|^2}{\prod_{k=1}^{v} p_k},$$

где $c(\bar{p})$ - весовая функция вектора вероятностей $c(\bar{p})$; обычно полагают $c(\bar{p}) = \mathbf{1}$.

Отметим, что в общем случае получить байесовскую функцию оценки для произвольной функции потерь, как правило, нелегко. Однако общепринято, что наиболее адекватные результаты получаются при использовании именно квадратичной функции потерь (см., например [6], [8]). Тот же результат – математическое ожидание апостериорной плотности вероятности искомого параметра (апостериорное среднее) – получается для широкого класса апостериорных распределений и при использовании любой другой выпуклой симметричной функции потерь [12][2].

4.2.1.Одномерный случай

Рассмотрим для простоты сначала случай $v = 2$, который соответствует разбиению пространства образов на две подобласти: правильных и неправильных классификаций.

Пусть полученное р.п. из имеющихся m прецедентов m_r распознает правильно, а на остальных $m_w = m - m_r$ - ошибается. Частота, как известно, является достаточной статистикой и условное распределение наблюдений при фиксированной статистике, не зависит, следовательно, от распределения наблюдений (чередования правильно и неправильно распознанных прецедентов).

Построим байесовские точечные функции оценки $\hat{p}$ неизвестной вероятности $p^* = 1 - \nu$ ошибочной классификации при различном задании функции потерь.

[2]Единственное существенное возражение против применения квадратичной функции потерь состоит в том, что она "подчеркивает хвосты" распределений, приписывая слишком большой вес редким, вообще говоря, значениям параметра. Однако для задачи оценки вероятностей это возражение снимается, поскольку область изменения параметра в этом случае конечна.

Формула Байеса в нашем случае имеет вид

$$f(p\,|\,m_w,\,m_r\,) \;=\; \frac{f(p)f(\,m_w,\,m_r\,|\,p)}{\int\limits_0^1 f(p)f(\,m_w,\,m_r\,|\,p)\,dp}\,. \tag{1}$$

Здесь $f(\,m_w,\,m_r\,|\,p\,) = p^{m_w}(1-p)^{m_r}$ - правдоподобие.

В качестве априорного распределения $f(p)$ мы будем использовать бетта-распределение (***B***) $Be\,(a,\,b)$ с параметрами $a > 0,\ b > 0$, плотность которого равна

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

B-распределение очень удобно для наших целей, поскольку в этом случае вычисления апостериорного распределения наиболее просто. С другой стороны, формы кривых плотностей $Be\,(a,\,b)$ при различных $a > 0,\ b > 0$ весьма разнообразны. Заметим здесь, что математическое ожидание и дисперсия ***B***-распределения равны

$$\mu_\beta \;=\; \frac{a}{a+b}\,, \quad \sigma_\beta^2 \;=\; \frac{ab}{(a+b)^2(a+b+1)}\,.$$

С учётом сделанного выбора плотность вероятности апостериорного распределения будет

$$\begin{aligned} f(\,p\,|\,m_w,\,m_r\,) \;=\; &\frac{\Gamma(a+b+m)}{\Gamma(m_w+a)\Gamma(m_r+b)} \\ &p^{m_w+a-1}(1-p)^{m_r+b-1}, \;\; p \in (0,1)\,, \end{aligned} \tag{2}$$

т.е. $Be(m_w+a,\,m_r+b)$.

Укажем, что для вычисления знаменателя (1) и подобных выражений используют формулу Лиувилля [5], [11]:

$$\int\limits_{S_{v-1}(\bar{x})} \prod_{i=1}^{n} x_i^{m_i}\,dx_1\ldots dx_n \;=\; \frac{m_1!\ldots m_n!}{(\sum\limits_{i=1}^{n} m_i + n - 1)!}\,,$$

где $m_1,\,m_2,\,\ldots\,,\,m_n$ - натуральные числа.

При $v = 2$ и учетом $p_1 + p_2 = 1,\ p_1 = p$ сформулированная в начале п. 4.2 задача минимизации принимает вид

$$\int\limits_0^1 W(\,p,\,q\,)f(\,p\,|\,m_w,\,m_r\,)\,dp \;=\; R(q) \;\rightarrow \min\,, \quad q \in S_{v\text{-}1}(x)\,.$$

Как указывалось выше, при квадратичной

$$W_1(p, q) = (p-q)^2$$

функции потерь байесовская оценка совпадает с математическим ожиданием апостериорного распределения. Математическое ожидание μ апостериорного распределения (2) есть

$$\mu = \frac{m_w + a}{m + a + b}.$$

Полученная оценка может рассматриваться как модификация МП-оценки с учётом априорной информации относительно p^* или как модификация априорной оценки $a/(a+b)$ с учётом наблюдённых величин m_w и m_r.

При отсутствии какой-либо информации о значениях вероятности p ($\gamma_i = 1$, $i = \overline{1,m}$) по принципу неопределённости Лапласа полагаем, что априорная вероятность имеет равномерное на $(0,1)$ распределение. Равномерное распределение – это $\boldsymbol{B}$-распределение с параметрами $a = b = 1$. Тогда получаем апостериорную плотность в виде

$$f(p\,|\,m_r,\, m_w) = \frac{\Gamma(m+2)}{\Gamma(m_w+1)\Gamma(m_r+1)}\, p^{m_w}(1-p)^{m_r},$$

т.е. плотность $\boldsymbol{B}$-распределения $Be(m_r + 1, m_w + 1)$ у которого $\mu = (m_w+1)/(m+2)$. Таким образом получена точечная функция оценка $\hat{p}_{W_1} = \hat{p}_W$ вероятности ошибки распознавания $1-\nu$:

$$\hat{p}_W = \frac{m_w + 1}{m + 2}. \tag{3}$$

Найдем теперь функцию оценки $\hat{p}_{W_2} = \hat{p}_W$ при нормированной функции потерь W_2. Имеем [1]:

$$R(q) = \int_0^1 \frac{(p-q)^2}{p(1-p)} \frac{(m+1)!}{m_r!\, m_w!}\, p^{m_w}(1-p)^{m_r} dp =$$

$$= \int_0^1 (p-q)^2 \frac{(m-1)!\, m\,(m+1)}{(m_r-1)!\, m_r\,(m_w-1)!\, m_v}\, p^{m_w-1}(1-p)^{m_r-1} dp =$$

$$= \frac{m\,(m+1)}{m_r m_w} \int_0^1 (p-q)^2 \frac{(m-1)!}{(m_r-1)!\,(m_w-1)!}\, p^{m_w-1}(1-p)^{m_r-1}\, dp =$$

$$= \frac{m\,(m+1)}{m_r m_w} \int_0^1 (p-q)^2 f(p\,|\,m_w-1,\, m_r-1)\, dp.$$

Минимум значения интеграла в последнем выражении будет достигаться при $q = \hat{p}_{W_2} = m_w/m$, и, таким образом, мы снова получаем оценку максимального правдоподобия.

Возвратимся к оценке (3). Ясно, что она является *смещённой*: если $\hat{p}$ - МП-оценка, то

$$\hat{p}_W = \frac{m}{m+2}\hat{p} + \frac{1}{m+2},$$

и с учетом свойств $\hat{p}$, приведённых в предыдущем разделе,

$$\mathbf{M}\{\hat{p}_W\} = \mathbf{M}\left\{\frac{m}{m+2}\hat{p} + \frac{1}{m+2}\right\} = \frac{mp^* + 1}{m+2} \neq p^*.$$

Также ясно, что оценка $\hat{p}_W$ *несмещена асимптотически.*

Дисперсия $\mathbf{D}\{\hat{p}_W\}$ полученной оценки равна

$$\mathbf{D}\{\hat{p}_W\} = \mathbf{D}\left\{\frac{m}{m+2}\hat{p} + \frac{1}{m+2}\right\} =$$
$$= \left(\frac{m}{m+2}\right)^2 \mathbf{D}\{\hat{p}\} = \frac{m\,p^*\,(1-p^*)}{(m+2)^2},$$

и оценка, очевидно, *состоятельна.*

Легко видеть, что несмещённая оценка $\overline{\mathbf{D}\{\hat{p}_W\}}$ дисперсии полученной оценки равна

$$\overline{\mathbf{D}\{\hat{p}_W\}} = \frac{m_w\,(m - m_w)}{(m+2)^2(m-1)}.$$

Имеем $\mathbf{D}\{\hat{p}_W\} < \mathbf{D}\{\hat{p}\}$ и дисперсия оценки $\mathbf{D}\{\hat{p}_W\}$ в $(m+2)^2/m^2$ раз меньше минимальной граничной по неравенству Крамера-Рао.

Указанное обстоятельство объясняется тем, что полученная байесовская оценка есть оценка смещённая и понизить дисперсию оценки удалось именно за счет выхода из класса несмещённых (для которых и выведено неравенство Крамера-Рао). Естественно, тот же результат получится, если сразу воспользоваться формулой для нижней границы смещённой оценки [10][3]. Ясно, что выигрыш в дисперсии оценки

[3]Дисперсии смещённых $\mathbf{D}_d$ и несмещённых $\mathbf{D}$ оценок параметра p связаны формулой $\mathbf{D}_d = (1 + b'_m(p))^2\,\mathbf{D}$, где $b_m(p)$ – смещение. В нашем случае

$$\hat{p} = p + \frac{1-2p}{m+2},\ b_m(p) = \frac{1-2p}{m+2},$$
$$b'_m(p) = -\frac{2}{m+2},\ (1 + b'_m(p))^2 = \left(\frac{m}{m+2}\right)^2.$$

будет особенно существенным при малых выборках. Следует, однако, иметь в виду, что для смещённой оценки дисперсия служит мерой близости не к оцениваемому параметру, а к математическому ожиданию оценки. Поэтому важное значение приобретает вопрос об "истинном" виде распределения вероятности p.

4.2.2. Обсуждение полученных оценок. Другие точечные оценки

С общей точки зрения нет никаких оснований, кроме удобства математических свойств (а также традиции практиков), выделять равенство истинному значению именно математического ожидания оценки в качестве критерия несмещённости. Вместо математического ожидания могут также быть выбраны медиана распределения или его мода (т.н. медианная несмещённость или несмещённость по моде[4]). В нашем случае мы столкнулись с ситуацией, когда смещённая оценка имеет дисперсию меньше, чем несмещённая, а значит и большую эффективность[5]. Мы считаем это достаточным основанием для того, чтобы отказаться от рассмотрения лишь класса несмещённых оценок. Кроме того, обоснованность использования байесовских оценок подтверждается и проведённым стохастическим моделированием (см. ниже).

Во-первых, полученная оценка обладает свойством асимптотической несмещённости, а само смещение невелико.

Во-вторых, представляется ясным, что для случая малых выборок, именно эффективность является основным критерием качества оценки (ср. [3]). Наличие у оценок последнего нерассмотренного основного свойства – состоятельности – имеет ценность всё же в основном при теоретических исследованиях, являясь асимптотическим свойством.

И, наконец, в третьих, МП-оценки, как правило, получаются неустойчивыми [14], а иногда и "катастрофически неустойчивыми"[6] к малым отклонениям от закона распределения. Поэтому такая оценка неудобна и с точки зрения *робастности* (устойчивости по отношению к постулируемым распределениям).

Заметим, что, неформально рассуждая, принятие МП-оценки (по моде) будет приводить к ошибкам, вообще говоря, редким, но, возможно, значительным, а байесовская оценка (по математическому ожиданию)

---

[4]См., например, [6], [9].

[5]Оценку с меньшей дисперсией мы считаем более эффективной.

[6]См. Tukey, J.W. A survey of sampling from contaminated distribution / Contributions to Prob. and Stat. Ed. I. Olkin et al. Stanford: Stanford Univ. Press, 1960, p. 446-486.

повлечет, как правило, ошибки частые, но небольшие. Представляется, что данные оценки в силу указанных свойств являются в своём роде граничными, и исходя из специфики конкретных задач Z в качестве точечной оценки искомой вероятности p^* можно выбрать любое значение между модой и математическим ожиданием полученного ***B***-распределения. Можно показать, что, например, его медиана $x_{(\beta)1/2}$ всегда расположена в указанном диапазоне и за оценку вероятности принять именно медиану. Такая оценка будет обладать свойством равновероятной недооценки и переоценки p^*, что может оказаться удобным для некоторых приложений. Кстати, она будет являться байесовской с функцией штрафа $W_3(p,q) = |p - q|$ [7].

В [6] для малых p^* предлагается в качестве априорного распределения брать $Be(1,\ b)$ с большим b. Тогда байесовской функцией оценки будет

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

Для нашей задачи можно попытаться использовать т.н. W-*минимаксную* оценку, при которой максимальные потери для некоторой выбранной функции потерь W минимальны по $p^* \in (0,1)$. Понятие W-минимаксности вводится независимо от задания какого-либо априорного распределения и поэтому, вообще говоря, может рассматриваться в рамках частотного подхода. Иногда оказывается возможным подобрать априорное распределение, при котором полученная минимаксная оценка оказывается также равной и соответствующей байесовской. Такое априорное распределение называют *наименее благоприятным*.

Если выбрать функцию потерь квадратичной ($W = W_1$), то минимаксная оценка параметра p биномиального распределения будет иметь вид [2], [13]

$$\hat{p} = \frac{\sqrt{m}}{1+\sqrt{m}}\frac{m_1}{m} + \frac{1}{1+\sqrt{m}}\frac{1}{2}.$$

Представляется, однако, что использование полученной функции оценки в нашем случае недостаточно оправдано с точки зрения "физики" задачи. Действительно, для вышеуказанной оценки наименее благоприятным распределением оказывается ***B***-распределение $Be(\sqrt{m}/2,\ \sqrt{m}/2)$. Неясно, как параметры этого распределения могут быть обоснованы в рамках задачи Z.

Если же выбрать нормированную квадратичную функцию потерь ($W = W_2$), то W_2-минимаксными оценками искомых вероятностей будут являться относительные частоты. При этом наименее благоприятном распределением оказывается равномерное. Неприемлемость же точечных оценок в виде относительных частот для случая малых выборок обсуждалась выше.

Для выяснения вопроса: *Какая из возможных точечных оценок наиболее адекватна реальным практическим ситуациям?* был проведен численный эксперимент. Для разных значений $p \in [0, 1]$ появления условного события A генерировались выборки объёма $n = 1, 2, \dots, 20$ и фиксировалось количество r наблюдённых событий. Затем вычислялось наиболее вероятное (среднее) значение p для которой при данном n наблюдается r появлений события A, т.е. определялась стохастическая оценка $\breve{p}$ вероятности $p(A)$ появления события A. Она сравнивалась с МП $\hat{p}_{ML} = r/n$ и байесовской $\hat{p}_B = (r+1)/(n+2)$ оценками по формуле

$$\breve{p} = \lambda \cdot \hat{p}_{ML} + (1 - \lambda) \cdot \hat{p}_B$$

(для чётных n и $r = n/2$ указанные оценки совпадают и значение λ не определено).

В результате оказалось, что полученные стохастические оценки, как правило, очень близки к соответствующим байесовским ($\lambda \approx 0$). Наибольшие относительные отклонения значений λ наблюдались когда r было равно $\frac{n\pm1}{2}$ для нечётных или, соответственно, $\frac{n}{2} \pm 1$ для чётных n, где рассматриваемые оценки мало различаются и величина λ плохо обусловлена. В интересующей нас области малых n и r значения стохастической и байесовской оценок совпадали с большой точностью (для приблизительно 10000 наблюдений значений r при данном n величина λ составляла порядка нескольких процентов). Таким образом целесообразность использования байесовские оценок, особенно в случае малых выборок, можно считать подтвержденным стохастическим моделированием[7].

[7]Программа стохастического моделирования написана А. Лапшиным в среде Delphi 5.0 для ПК. Для генерации случайной величины r имеющей биномиальное распределение использовался метод "браковки". В программе моделировалось 10000 экспериментов соответствующих каждому p при данном n. Время счета при этом не превосходило трех минут (процессор Pentium-III).

4.2.3. Многомерный случай

Пусть теперь $v > 2$. В многомерном случае формула Байеса имеет вид

$$f(\bar{p}\,|\,m_1, m_1 \ldots m_v) = \frac{f(\bar{p})\, f(\,m_1, m_1 \ldots m_v\,|\,\bar{p}\,)}{\int\limits_{S_{v-1}(\bar{p})} f(\bar{p})\, f(\,m_1, m_1 \ldots m_v\,|\,\bar{p}\,)\, d\bar{p}}\,. \tag{4}$$

Здесь

$$f(\,m_1,\, m_2,\, \ldots\,,\, m_v\,|\,\bar{p}\,) = \prod_{k=1}^{v} p_k^{m_k}$$

является функцией правдоподобия и, естественно, выполняется условие нормировки

$$\sum_{k=1}^{v} p_k = 1\,. \tag{5}$$

Как отмечалось в предыдущей статье, искомые вероятности $\bar{p} = \{p_k\}_{k=1}^{v}$ подчиняются полиномиальному распределению.

В качестве априорного распределения $f(\bar{p})$ мы будем использовать $(v\ -\ 1)$-мерное распределение Дирихле $Di\,(d_1,\, d_2,\, \ldots\,,\, d_{v-1};\, d_v)$ с параметрами $d_1,\, d_2,\, \ldots\,,\, d_v$, имеющее плотность

$$f(\,\bar{p}\,|\,d_1,\, d_2,\, \ldots\,,\, d_v\,) = \frac{\Gamma(d_1 + d_2 + \ldots + d_v)}{\Gamma(d_1)\Gamma(d_2)\ldots\Gamma(d_v)} \prod_{k=1}^{v} p_k^{d_k - 1} \tag{6}$$

в любой точке симплекса $S_{v-1}(\bar{x})$ и равную нулю в других точках $\mathbb{R}^v$. Здесь все $d_1,\, d_2,\, \ldots\,,\, d_v$ - вещественные положительные числа. При $v = 2$ $\;Di(d_1;\, d_2)$ сводится к $Be(a,\, b)$.

С помощью формулы Лиувилля легко установить, что среднее, дисперсия и ковариация $(v\ -\ 1)$-мерного распределения Дирихле выражаются формулами

$$\mu_{Di}(x_k) = \frac{d_k}{d}\,, \quad \sigma^2_{Di}(x_k) = \frac{d_k(d - d_k)}{d^2(d+1)}\,,$$

$$\sigma_{Di}(x_i,\, x_j) = \frac{d_i d_j}{d^2(d+1)}\,,$$

где $\quad k = \overline{1, v}\,, \quad i \neq j, \quad i, j = \overline{1, v}\,, \quad d = \sum_{k=1}^{v} d_k.$

Из (4) и (6) следует, что плотность вероятности апостериорного распределения есть

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

т.е. будет являться плотностью $(v-1)$-мерного распределения Дирихле

$$Di\,(m_1+d_1,\,\dots\,,\,m_{v-1}+d_{v-1};\,m_v+d_v)\,.$$

Для квадратичной функции потерь

$$W(\bar{p},\,\bar{q}\,) = \|\,\bar{p}-\bar{q}\,\|^2$$

байесовскими оценками $\hat{p}_i$ вероятностей p_i^* будут являться компоненты вектора μ_k апостериорного среднего $\bar{\mu}=(\mu_1,\,\dots\,\mu_v)^T$, равные

$$\hat{p}_k \;=\; \mu_k \;=\; \frac{m_k+d_k}{m+\sum_{j=1}^{v} d_j}\,,\quad k=\overline{1,v}\,.$$

Заметим, что при $d_k=m_k,\ k=\overline{1,v}$, байесовские оценки будут совпадать с МП-оценками.

В условиях отсутствия информации о весах прецедентов принимаем в качестве распределения $\bar{p}$ равномерное. Равномерное распределение есть распределение Дирихле $Di\,(1,\,\dots\,,\,1;\,1)$. Получаем отсюда, что апостериорная плотность вероятностей имеет вид

$$f(\bar{p}\,|\,m_1,m_2\ldots m_v) \;=\; \frac{\Gamma(m+v)}{\Gamma(m_1+1)\ldots\Gamma(m_v+1)}\prod_{k=1}^{v} p_k^{m_k} \;=$$

$$=\; \frac{(m+v-1)!}{m_1!\ldots m_v!}\,p_1^{m_1}p_2^{m_2}\ldots\,p_v^{m_v}\,,$$

где $\bar{p}\in S_{v-1}(\bar{x})$, т.е. является плотностью $(v-1)$-мерного распределения Дирихле

$$Di\,(m_1+1,\,\dots\,,\,m_{v-1}+1;\,m_v+1)\,,$$

а байесовскими оценками $\hat{p}_k$ вероятностей p_k^* будут являться величины

$$\hat{p}_k \;=\; \mu_k \;=\; \frac{m_k+1}{m+v}\,,\quad k=\overline{1,v}\,. \tag{7}$$

Заметим, что если формально положить $m\;=\;0$ (отсутствие прецедентов) получаем

$$\hat{p}_1 \;=\; \hat{p}_2 \;=\; \ldots \;=\; \hat{p}_v \;=\; 1/v$$

– принцип неопределенности Лапласа, использованный нами при выводе (7).

Легко показать, что применение нормированной многомерная функция потерь

$$W(\bar{p}, \bar{q}) = \frac{\|\bar{p} - \bar{q}\|^2}{\prod\limits_{k=1}^{v} p_k}$$

приводит к оценкам $\hat{p}_k = m_k/m,\ k = \overline{1, v}$, совпадающим в этом случае МП-оценками.

Аналогично одномерному случаю, используя свойство воспроизводимости[8] по m полиномиального распределения $M(m; \cdot)$ и свойства распределения Дирихле получим, что компоненты вектора дисперсий оценок (7) суть

$$\mathbf{D}\{\hat{p}_k\} = \frac{p_k^*(1 - p_k^*)\, m}{(m + v)^2},$$

а их несмещённые оценки –

$$\overline{\mathbf{D}\{\hat{p}_k\}} = \frac{m_k(m - m_k)}{(m - 1)(m + v)^2}, \quad k = \overline{1, v}.$$

4.2.4. Случай неравных весов прецедентов

Перейдем теперь к рассмотрению случая, когда прецедентная информация включает в себя вектор весов экзаменационных элементов $\left\{\gamma_i = \gamma(x_i)\right\}_{i=m}^{m} = \bar{\gamma}_m$, где не все компоненты равны.

Значение γ_i показывает "важность" или частоту встречаемости прецедента x_i. Часто заказчик, готовя исходные данные для решения задачи распознавания и желая дать как можно более полное и компактное описание пространства образов, намеренно или вынужденно[9] предоставляет разработчику список прецедентов более-менее равномерно распределённых по пространству образов, указывая большую или меньшую "типичность" данного прецедента с помощью приписывания ему соответствующего веса. Этот приём может существенно понизить

---

[8]Параметрическая с параметром θ функция распределения $P(u, \theta)$ случайной величины u называется *воспроизводящей по* θ, если для независимых случайных величин u_1 и u_2, которые имеют функции распределения $P(u_1, \theta_1)$ и $P(u_2, \theta_2)$ соответственно, величина $u_1 + u_2$ распределена по $P(u_1 + u_2, \theta_1 + \theta_2)$ (см. [10]). Если в (4) $f(\bar{p})$ и $f(m_1, m_1 \ldots m_v \,|\, \bar{p})$ принадлежат к одному типу воспроизводящих плотностей, то и плотность $f(\bar{p} \,|\, m_1, m_1 \ldots m_v)$ будет относится к тому же типу распределений.

[9]например, из-за отсутствия соответствующих данных.

объём предоставляемой прецедентной информации без потери её репрезентативности.

Заметим, что "важность" или "типичность" $\gamma_i \geqslant 1$ данного прецедента x_i можно трактовать как задание "дополнительных прецедентов" вблизи x_i с аналогичными признаками, и так, что дополнительные прецеденты всегда классифицируются также, как и x_i. *Указанные "дополнительные прецеденты" назовем квазипрецедентами.* Для точного соответствия с информацией, заложенной в весах, их число не обязано быть целым. Действительно, в этом случае та или иная классификация x_i приведет к соответствующему увеличению оценки вероятности p_i, что повысит её вклад в величину среднего риска и отразит, таким образом, значимость данного прецедента. Заметим, что возможность такого представления информации о весах вытекает из гипотезы компактности.

Ясно, однако, что в рассматриваемом случае при остающейся верной гипотезе представительности, её форма в виде "Гипотеза 1" (см. [4]) уже становится недостаточной. Поэтому для обоснования определения надежности выбранного р.п. данную гипотезу нужно дополнить предположениями относительно имеющегося вида прецедентной информации.

Наше основное предположение состоит в том, что веса образов γ_i через количества квазипрецедентов описывают вероятности появления образов в окрестностях x_i с тем же значением истинного классификатора $f^*(x_i)$. Точнее, мы считаем, что веса γ_i образов x_i линейно и аддитивно связаны с вероятностями появления в процессе классификации на практике новых образов в окрестностях x_i с тем же значением истинного классификатора $f^*(x_i)$, $i = 1, 2, \ldots, m$. Конкретно, мы дополняем Гипотезу 1 нижеследующей Гипотезой 2.

Гипотеза 2. *При неравных весах $\gamma_i, \neq const$, $i = \overline{1,m}$, набор прецедентов $\{x_i\}_{i=1}^m$ не является реализацией независимой выборки m случайных величин из генеральной совокупности с распределением $P(X)$ на $\mathcal{X}$, однако веса прецедентов $\{\gamma_1, \gamma_2, \ldots, \gamma_m\}$ отражают априорную информацию о распределении $P(X)$.*

Поскольку мы трактуем веса как информацию о количестве квазипрецедентов в окрестности x_i, естественно считать, что $\gamma_i, \geqslant 1$, $i = \overline{1,m}$, (для чего, при необходимости, поделим все веса на $\min \gamma_i$). Количество дополнительных квазипрецедентов будет описываться величинами $\gamma_i - 1$, т.к. в окрестности x_i уже есть один прецедент – сам x_i. Обозначим $\gamma'_i = \gamma_i - 1,\ i = \overline{1,m}$.

Естественно считать, что априорный вес μ'_k области X_k аддитивен и пропорционален весам попавших в него квазипрецедентов, т.е.

$$\mu'_k = \sum_{i:\, x_i \in X_k} \gamma'_i,\ k = \overline{1,v}\,.$$

Введём обозначение

$$\sum_{i:\, x_i \in X_k} \gamma_i, = \mu_k\,. \qquad (8)$$

Понятно, что

$$\mu'_k = \mu_k - m_k \geqslant 0,\ k = \overline{1,v},\ \text{поскольку}\ m_k = \sum_{i:\, x_i \in X_k} 1\,.$$

В качестве априорного распределения вероятностей на $\{X_k\}_{k=1}^{v}$ примем распределение Дирихле

$$Di\,(\mu'_1 + 1,\, \mu'_2 + 1,\, \ldots\,,\, \mu'_{v-1} + 1;\, \mu'_v + 1)\,.$$

Представляется, что такая трактовка весов прецедентов достаточно адекватно отражает рассматриваемую ситуацию.

Обозначим

$$M = \sum_{k=1}^{v} \mu_k\,. \qquad (9)$$

Используя формулу Байеса (4) и вышеприведённые зависимости получим апостериорное распределение вектора вероятностей $\bar{p} = \{p_1,\, p_2,\, \ldots\,,\, p_v\},\ p_k \in (0,1),\ k = \overline{1,v}$:

$$f(\,\bar{p}\,|\,m_1, m_2 \ldots m_v) = \frac{\Gamma(m + v + \sum\limits_{k=1}^{v} \mu'_k)}{\prod\limits_{k=1}^{v} \Gamma(m_k + \mu'_k + 1)} \prod_{k=1}^{v} p_k^{m_k + \mu'_k} =$$

$$= \frac{\Gamma(M + v)}{\prod\limits_{k=1}^{v} \Gamma(\mu_k + 1)} \prod_{k=1}^{v} p_k^{\mu_k} = \frac{(M + v - 1)!}{\mu_1!\,\mu_2!\,\ldots\,\mu_v!}\, p_1^{\mu_1} p_2^{\mu_2} \ldots\, p_v^{\mu_v},$$

которое является плотностью $(v-1)$-мерного распределения Дирихле

$$Di\,(\mu_1 + 1,\, \mu_2 + 1,\, \ldots\,,\, \mu_{v-1} + 1;\, \mu_v + 1)\,.$$

Байесовской оценкой искомых вероятностей при квадратичной функции потерь будет вектор апостериорного среднего с компонентами

$$\hat{p}_k = \frac{\mu_k + 1}{M + v}, \quad k = \overline{1, v}, \tag{10}$$

где μ_k и M вычисляются по (8) и (9) соответственно. Эти значения и предлагается использовать в качестве точечных оценок вероятностей событий $x \in X_k$ в общем случае задачи Z (легко проверить, что при $\gamma_i = const$, $i = \overline{1, m}$, формула (10) превращается в (7)).

Ясно также, что в рамках частотного подхода формула (10) примет вид

$$\hat{p}_k = \frac{\mu_k}{M}, \quad k = \overline{1, v}.$$

Литература

1. *Беляев Ю.К., Носко В.П.* Основные понятия и задачи математической статистики: Учеб. пособие. – М.: Изд-во МГУ, ЧеРо, 1998.
2. *Боровков А.А.* Математическая статистика. – М.: Наука, 1984.
3. *Гасканов Д.В., Шаповалов В.И.* Малая выборка. – М.: Статистика, 1978.
4. *Гуров С.И.* Оценки надёжности алгоритмов классификации I. Введение. Точечные оценки.
5. *Интегралы* и ряды. Элементарные функции /Прудников А.И., Брычков Ю.А., Маричев О.И. – М.: Наука, 1981.
6. *Леман Э.* Теория точечного оценивания /Пер. с англ. – М.: Наука, Гл. ред. физ.-мат. лит., 1991.
7. *Леман Э.* Проверка статистических гипотез. – М.: Наука, Гл. ред. физ.-мат. лит., 1979.
8. *Патрик Э.* Основы теории распознавания образов /Пер. с англ. Под. ред. Б.Р. Левина. – М.: Сов. радио, 1980.

9. *Рао С.Р.* Линейные статистические методы и их применение /Пер. с англ. – М.: Накуа, 1968.
10. *Уилкс С.* Математическая статистика /Пер. с англ. – М.: Наука, 1967.
11. *Фихтенгольц Г.М.* Курс дифференциального и интегрального исчисления. Т. 3. – М.: Наука, 1966.
12. *Фукунага К.* Введение в статистическую теорию распознавания образов /Пер. с англ. – М.: Наука, Гл. ред. физ.-мат. лит., 1979.
13. *Чибисов Д.М., Пагурова В.И.* Задачи по математической статистике. – М.: Изд-во Моск. ун-та, 1990.
14. *Шурыгин А.М.* Прикладная стохастика: робастность, оценивание, прогноз. – М.: Финансы и статистика, 2000.