Предельное поведение оценки риска оптимальной групповой процедуры классификации выборки из однопараметрического экспоненциального семейства

Бабушкина Е. В., Чичагов В. В.
helvad@yandex.ru
Пермь, Пермский государственный университет

В данной работе продолжены исследования, начатые в работах [1, 2]. В работе [1] было исследовано предельное поведение оценки байесовского риска при групповой классификации с заданной границей. В данной работе также, как и в [2], эти результаты распространяются на случай классификации с использованием квазиоптимального правила. Классификация в этом случае осуществляется по областям со случайными границами. Другая отличительная часть данной работы состоит в том, что направляющая функция экспоненциального семейства распределений наблюдаемой случайной величины, как функция этой величины, имеет гамма-распределение.

Рассматривается решение следующей задачи групповой классификация n_0 объектов по измерениям их характеристик [3].

Классифицируемая выборка $\pi_{00} = \{X_{0,1}, \ldots, X_{0,n_0}\}$ может принадлежать одной из двух совокупностей $\pi_i,\ i = 1, 2$. Имеются две обучающие выборки $\pi_{i0} = \{X_{i,1}, \ldots, X_{i,n_i}\}$, элементы которых являются независимыми случайными величинами, имеющими то же распределение, что и случайная величина η_i с плотностью $f(y, \theta_i),\ i = 1, 2$.

Решение задачи будем искать при следующих предположениях.

(C_1). Распределение вероятностей случайной величины η_i принадлежит однопараметрическому экспоненциальному семейству с плотностью

$$f(y, \theta_i) = h(y) \exp\left\{\theta_i\ T(y) + V(\theta_i)\right\}, \quad y \in G \subset \mathbb{R}, \quad \theta_i \in \Theta \subset \mathbb{R}.$$

Здесь $h(y),\ T(y)$ — известные борелевские функции, $V(\theta)$ — непрерывно дифференцируемая функция параметра θ.

(C_2). Сумма $S_{n_i} = \sum_{j=1}^{n_i} T(X_{i,j})$ является достаточной статистикой параметра θ_i по выборке $\pi_{i0},\ g_m(t, \theta_i)$ — плотность распределения случайной суммы $\sum_{j=1}^{m} T(X_{i,j})$.

(C_3). Случайная величина $Y_i = T(\eta_i),\ i = 1, 2$, имеет гамма-распределение с плотностью

$$f_{Y_i}(t; \sigma_i, \nu) = \frac{t^{\nu-1}}{\sigma_i^{\nu}\Gamma(\nu)} \cdot e^{-t/\sigma_i}, \quad t > 0, \quad \sigma_i = -\frac{1}{\theta_i} > 0, \quad \nu > 0.$$

(C_4). Для случайных событий $B_i = \left\{\left|\sum_{j=1}^{n_0}\left(T(X_{0,j}) - \nu\sigma_i\right)\right| \geqslant n_i^{\gamma_i}\right\}$, определённых при некотором $\gamma_i \in \left(0, \frac{1}{4}\right)$, справедливы соотношения $\lim\limits_{n_i\to\infty} n_i \,\mathsf{P}\,(B_i) = 0,\ \ i = 1, 2.$

Не нарушая общности рассуждений, далее предполагаем, что $\sigma_1 > \sigma_2$, оба параметра неизвестны. Следуя [2, 3], сформулируем утверждение.

Теорема 1. *Если выполнены условия* (C_1), (C_3) *то оптимальное решающее правило групповой классификации* π_{00} *имеет вид:*

$$\pi_{00} \in \pi_1, \quad \textit{если} \ \ q(t) = \ln\frac{\omega_1 g_{n_0}(t,\theta_1)}{\omega_2 g_{n_0}(t,\theta_2)} \geqslant 0, \tag{1}$$

где ω_1, ω_2 — *некоторые заданные числа. Неравенство (1) осуществляет разбиение числовой прямой на два интервала*

$$J_1 = \{t\colon q(t) < 0\} = \{t\colon t < c\}\,, \qquad c = -\frac{\ln\frac{\omega_1}{\omega_2} - n_0\nu\ln\frac{\sigma_1}{\sigma_2}}{\frac{1}{\sigma_2} - \frac{1}{\sigma_1}}.$$
$$J_2 = \{t\colon t \geqslant c\}\,,$$

Основной качественной характеристикой правила (1) является байесовский риск

$$R = \omega_1 \,\mathsf{P}_{\theta_1}\{\pi_{00} \in \pi_2\} + \omega_2 \,\mathsf{P}_{\theta_2}\{\pi_{00} \in \pi_1\} = \sum_{j=1}^{2}\omega_j \int_{J_j} g_{n_0}(t,\theta_j)\,dt.$$

Определим квазиоптимальное решающее правило групповой классификации π_{00}:

$$\pi_{00} \in \pi_1, \quad \text{если} \ \ q(t \mid S_{n_1}, S_{n_2}) = \ln\frac{\omega_1\widehat{g}_{n_0}(t \mid S_{n_1})}{\omega_2\widehat{g}_{n_0}(t \mid S_{n_2})} \geqslant 0, \tag{2}$$

где $\widehat{g}_{n_0}(t|S_{n_j})$ — несмещенная оценка плотности $g_{n_0}(t,\theta_j)$, являющаяся, согласно [4], плотностью бета-распределения.

Неравенство (2) осуществляет разбиение числовой прямой на две области

$$J_1(S_{n_1}, S_{n_2}) = \{t\colon q(t \mid S_{n_1}, S_{n_2}) < 0\}\,,$$
$$J_2(S_{n_1}, S_{n_2}) = \{t\colon q(t \mid S_{n_1}, S_{n_2}) \geqslant 0\}\,.$$

Определим оценку риска квазиоптимального правила (2), также как и в [2], следующим образом

$$\widetilde{R} = \widetilde{R}(S_{n_1}, S_{n_2}) = \sum_{j=1}^{2}\omega_j \int_{J_j(S_{n_1},S_{n_2})} \widehat{g}_{n_0}(t \mid S_{n_j})dt. \tag{3}$$

Подобно теореме 5 из [2], можно доказать следующее утверждение, определяющее предельное поведение оценки риска (3).

Теорема 2. *Пусть выполнены условия* (C_1)–(C_4), $c > 0$, $n_1 \leqslant n_2$, *неравенство правила (2) обращается в равенство в единственной точке* $\widetilde{c}$,

$$[\widetilde{\sigma}_R]^2 = \sum_{j=1}^{2} \frac{[\omega_j \widetilde{\sigma}_j]^2}{n_j}, \quad [\widetilde{\sigma}_j]^2 = \frac{\left[\widetilde{c} \cdot \widehat{g}_{n_0}(\widetilde{c} \mid S_{n_j})\right]^2}{\nu}, \quad \widetilde{c} > 0. \tag{4}$$

Тогда при $n_1 \to \infty$ *последовательность нормированных оценок квазиоптимального риска* $(\widetilde{R} - R)/\widetilde{\sigma}_R$ *сходится по распределению к стандартной нормальной случайной величине.*

Замечание 1. При конечных объемах обучающих выборок для некоторых сочетаний параметров возможно существование двух решений $\widetilde{c}$, или же отсутствие хотя бы одного решения. В первом случае для расчета $[\widetilde{\sigma}_R]^2$ можно либо воспользоваться теоремой 5 из [2], либо видоизменить формулы (4). Возникновение второй ситуации означает вырожденность квазиоптимального правила групповой классификации.

В дальнейшем планируется изучение правомерности применения асимптотических результатов, полученных в данной работе и в [1, 2] для конечных объемов обучающих выборок, а также сравнение изложенных в этих работах подходов к оценке асимптотической дисперсии байесовского риска.

Работа выполнена при поддержке РФФИ, проект № 05-01-00229.

Литература

[1] *Бабушкина Е. В., Чичагов В. В.* Применение несмещенных оценок к оцениванию риска процедуры групповой классификации // Статистические методы оценивания и проверки гипотез. — Перм. ун-т, 2006. — С. 4–11.

[2] *Чичагов В. В.* Построение статистических выводов, основывающихся на несмещенных оценках, по интервалам случайной длины // Статистические методы оценивания и проверки гипотез. — Перм. ун-т, 2007. — С. 59–71.

[3] *Абусев Р. А., Лумельский Я. П.* Статистическая групповая классификация. — Перм. ун-т, 1987. — 92 с.

[4] *Воинов В. Г., Никулин М. С.* Несмещенные оценки и их применения. — М.: Наука, 1989. — 440 с.