Электронный сетевой политематический журнал «Научные труды КубГТУ». 2024. № 5. С. 111-118
Scientific Works of the Kuban State Technological University. 2024; (5): 111–118.

Научная статья
УДК 51-7
DOI 10.26297/2312-9409.2024.5.9

Анализ вероятностей, связанных с ошибками 1-го и 2-го рода при бинарной классификации

Николай Андреевич Белоусов[1x], Денис Сергеевич Махов[2], Роман Дамирович Махмутов[3], Михаил Юрьевич Захаров[4]

[1,2,3]Краснодарское высшее военное орденов Жукова и Октябрьской Революции Краснознаменное училище имени генерала армии С.М. Штеменко, Краснодар, Россия, [1]loki911112@mail.ru[x]
[4]Кубанский государственный университет, Краснодар, Россия

Аннотация. Проводится анализ вероятностей ошибок 1-го и 2-го рода при бинарной классификации. Проанализирована математическая составляющая указанной проблемы системной оценки достоверности статистических гипотез.

Ключевые слова: системный подход, системный анализ, статистические гипотезы, критическая область, ошибки 1-го и 2-го рода, уровень значимости, мощность критерия, случайное событие, вероятность, условная вероятность, плотность распределения случайной величины, мода случайной величины

Original article

Analysis of probabilities associated with errors of the 1st and 2nd kind in binary classification

Nikolay A. Belousov[1], Denis S. Mahov[2], Roman D. Mahmutov[3], Michail Y. Zaharov[4]

[1,2,3]Krasnodar Higher Military Orders of Zhukov and the October Revolution Red Banner School named after General of the Army S.M.Shtemenko, Krasnodar, Russia, [1]loki911112@mail.ru
[4]Kuban State University, Krasnodar, Russia

Abstract. The analysis of the probabilities of errors of the 1st and 2nd kind in binary classification is carried out. The mathematical component of the specified problem of the systematic assessment of the reliability of statistical hypotheses is analyzed.

Keywords: a systematic approach, system analysis, statistical hypotheses, critical area, errors of the first and second kind, significance level, criterion power, random event, probability, conditional probability, distribution density of a random variable, mode of a random variable.

С ошибками 1-го и 2-го рода при проверке статистических гипотез связаны в соответствии с подходом, изложенном в [1], определенные условные априорные и апостериорные вероятности, а также соответствующие полные

Электронный сетевой политематический журнал «Научные труды КубГТУ». 2024. № 5. С. 111-118
Scientific Works of the Kuban State Technological University. 2024; (5): 111–118.

вероятности. Анализ указанных вероятностей может являться важной частью статистических исследований различных реальных процессов.

В работе анализируются условные распределения статистического критерия проверки гипотез при справедливости основной и конкурирующей гипотез. Строятся соответствующие формулы полной вероятности. С помощью формулы Байеса получены условные апостериорные вероятности ошибок 1-го и 2-го рода.

Актуальность данной работы основывается на том, что применение инструментов указанного анализа позволит проще выполнять корректное моделирование и оптимизацию вероятностей ошибок 1-го и 2-го рода, что в свою очередь позволяет (оптимально) улучшать достоверность бинарной классификации в различных реальных процессах.

Рассмотрим основную (нулевую) гипотезу H_0 и альтернативную гипотезу H_1. Введем также соответствующие события:

$$H_0 = \{\text{гипотеза } H_0 \text{ верна}\}, \quad (1)$$

$$H_1 = \{\text{гипотеза } H_1 \text{ верна}\}. \quad (2)$$

Причем в силу рассмотрения процесса бинарной классификации, будем считать события (1), (2) противоположными друг другу:

$$H_0 = \overline{H_1},\ H_1 = \overline{H_0}. \quad (3)$$

В силу соотношений (3) выполняется:

$$P(H_0) + P(H_1) = 1, \quad (4)$$

причем в (4) будем считать известными обе вероятности в левой части равенства.

Построим следующие условные вероятности (распределения) статистики (статистического критерия) *y* из [1]:

Электронный сетевой политематический журнал «Научные труды КубГТУ». 2024. № 5. С. 111-118
Scientific Works of the Kuban State Technological University. 2024; (5): 111–118.

$$P(a < y < +\infty|H_0) = \int_a^{+\infty} p_0(x)dx = \alpha, \quad (5)$$

где α – вероятность ошибки 1-го рода, a – левая граница критической области для гипотезы H_0 , $p_0(x)$ – плотность вероятности статистики (статистического критерия) при справедливости нулевой гипотезы H_0 [1];

$$P(a < y < +\infty|H_1) = \int_a^{+\infty} p_1(x)dx = 1 - \beta, \quad (6)$$

где β – вероятность ошибки 2-го рода ($1 - \beta$ – мощность статистического критерия), a - аналогично формуле (5) [2], $p_1(x)$ - плотность вероятности статистики при справедливости конкурирующей гипотезы H_1 [1].

Левые части формул (5) и (6) – условные вероятности попадания статистики y в критическую область (для гипотезы H_0) при противоположных гипотезах – событиях H_0 и H_1. Правые части – соответственно, уровень значимости и мощность критерия при проверке гипотезы H_0. Можно говорить, что указанные формулы наглядно иллюстрируют взаимосвязь вероятностных характеристик (α и $1 - \beta$) статистической проверки достоверности гипотез и условного распределения соответствующего статистического критерия (при событиях – условия H_0 и H_1).

Теперь, используя вероятности из (4) и левые части (5) и (6), возможно выписать формулу полной вероятности [3] для попадания статистики y в критическую область (для гипотезы H_0):

$$P(a < y < +\infty) = P(H_0)P(a < y < +\infty|H_0) +$$
$$+P(H_1)P(a < y < +\infty|H_1) = \alpha P(H_0) + (1 - \beta)P(H_1). \quad (7)$$

Электронный сетевой политематический журнал «Научные труды КубГТУ». 2024. № 5. С. 111-118
Scientific Works of the Kuban State Technological University. 2024; (5): 111–118.

Анализируя правую часть формулы (7), можно сделать вывод о (вероятностном) “качестве” критической области (для гипотезы H_0): достоверность критической области можно оценить следующим соотношением:

$$D_{\text{кр.об}} = \frac{(1-\beta)P(H_1)}{\alpha P(H_0) + (1-\beta)P(H_1)}. \quad (8)$$

Трактовка содержательного смысла соотношения (8) такова: полная вероятность попадания статистики y в критическую область $(a; +\infty)$ разбивается на два слагаемых: итоговую вероятность ошибки (1-го рода) $\alpha P(H_0)$ и итоговую вероятность правильного решения $(1-\beta)P(H_1)$, после чего находится доля вероятности правильного решения от указанной полной вероятности. Очевидно, что достоверность критической области (8) существенно зависит от значения $P(H_0) = 1 - P(H_1)$, помимо значений α и β.

Как видно из формул (7), (8) и формулы Байеса [2], введенная выше достоверность критической области является апостериорной вероятностью события (гипотезы) H_1 при условии попадания статистики y в критическую область $(a; +\infty)$:

$$D_{\text{кр.об}} = P(H_1 | a < y < +\infty). \quad (9)$$

Достоверность критической области в (8), положив $P = P(H_0)$ и учитывая (4), можно считать функцией переменных p, α, β $D_{\text{кр.об}} = D_{\text{кр.об}}(p, \alpha, \beta)$. Нетрудно показать, что производные указанной функции

Электронный сетевой политематический журнал «Научные труды КубГТУ». 2024. № 5. С. 111-118
Scientific Works of the Kuban State Technological University. 2024; (5): 111–118.

по каждой из переменных отрицательны (для реальных значений p, α, β: $0 < p, \alpha, \beta < 1$). Что говорит об убывании достоверности критической области с ростом любой из вероятностей: верность гипотезы H_0 ошибки 1-го рода, ошибки 2-го рода.

По аналогии с выводом формулы (9) можно получить соотношение для ошибочности критической области:

$$O_{\text{кр.об}} = P(H_0 | a < y < +\infty). \quad (10)$$

Из соотношений (3), (9) и (10) следует, что

$$D_{\text{кр.об}} + O_{\text{кр.об}} = 1. \quad (11)$$

Из (11) очевидно, что $O_{\text{кр.об}} = 1 - D_{\text{кр.об}}$, а значит ошибочность критической области также является функцией переменных p, α, β.

$$O_{\text{кр.об}} = O_{\text{кр.об}}(p, \alpha, \beta) = 1 - D_{\text{кр.об}}(p, \alpha, \beta). \quad (12)$$

Из соотношения (12) и того, что производные функции $D_{\text{кр.об}}(p, \alpha, \beta)$ по каждой из переменных p, α, β отрицательны, следует, что соответствующие производные функции $O_{\text{кр.об}}(p, \alpha, \beta)$, наоборот, положительны. Что свидетельствует о росте ошибочности критической области с ростом любой из вероятностей p, α, β.

Построим теперь формулы, аналогичные (7), (9), и (10), для произвольного интервала статистики *y*: $c < y < d$. С помощью формул полной вероятности и Байеса получим следующие соотношения:

Электронный сетевой политематический журнал «Научные труды КубГТУ». 2024. № 5. С. 111-118
Scientific Works of the Kuban State Technological University. 2024; (5): 111–118.

$$P(c < y < d) = P(H_0)P(c < y < d|H_0) + P(H_1)P(c < y < d|H_1) \quad (13)$$

$$= P(H_0)\int_c^d p_0(y)\,dy + P(H_1)\int_c^d p_1(y)\,dy;$$

$$P(H_0|c < y < d) = \frac{P(H_0)P(c < y < d|H_0)}{P(c < y < d)}; \quad (14)$$

$$P(H_1|c < y < d) = \frac{P(H_1)P(c < y < d|H_1)}{P(c < y < d)}. \quad (15)$$

Формулы (13) – (15) позволяют анализировать распределение статистики *y*, учитывая обе плотности $p_0(x)$ и $p_1(x)$, а также оценивать и сравнивать вероятности верности обеих гипотез H_0 и H_1 для произвольных интервалов статистики *y*.

Перейдем теперь к построению и анализу формул для расчета вероятностей ошибок 1-го и 2-го рода, при условии попадания статистики *y* в произвольный интервал (c; d). Обозначим эти вероятности следующим образом:

$$P(O_{1p}|c < y < d) = \alpha_{(c;d)}, \quad (16)$$

$$P(O_{2p}|c < y < d) = \beta_{(c;d)}, \quad (17),$$

где O_{1p} и O_{2p}- ошибки 1-го и 2-го рода соответственно.

Будем строить формулы для вероятностей (16), (17), исходя из следующих соображений:

1) ошибки 1-го и 2-го рода могут иметь место только при верности гипотез H_0 либо H_1соответственно.

Электронный сетевой политематический журнал «Научные труды КубГТУ». 2024. № 5. С. 111-118
Scientific Works of the Kuban State Technological University. 2024; (5): 111–118.

2) интервал (c; d) должен пересекаться с критической областью, либо с областью ложного принятия гипотезы H_0.

3) надо использовать вероятность условия, то есть $P(c < y < d)$.

Основываясь на вышеприведенных соображениях (п. п. 1-3), построим выражения для вероятностей (16), (17):

$$\alpha_{(c;d)} = \frac{P(H_0)\int_{max(c;a)}^{max(d;a)} p_0(y)dy}{P(c < y < d)}; \quad (18)$$

$$\beta_{(c;d)} = \frac{P(H_1)\int_{min(c;a)}^{min(d;a)} p_1(y)dy}{P(c < y < d)}. \quad (19)$$

Вероятности $\alpha_{(c;d)}$ и $\beta_{(c;d)}$, определяемые выражениями (18), (19), можно (условно) назвать апостериорными вероятностями ошибок 1-го и 2-го рода соответственно.

Формулы (18), (19) позволяют находить вероятности ошибок 1-го и 2-го рода при конкретных интервалах наблюдаемых значений статистики *y*.

Проведенный анализ вероятностей, связанных с ошибками 1-го и 2-го рода, построенные формулы для соответствующих вычислений дают дополнительные инструменты, позволяющие проще выполнять корректное моделирование и оптимизацию вероятностей ошибок 1-го и 2-го рода. Что позволяет оптимально улучшать достоверность бинарной классификации.

Представляется целесообразным продолжить исследование в направлениях, затронутых в данной статье.

Электронный сетевой политематический журнал «Научные труды КубГТУ». 2024. № 5. С. 111-118
Scientific Works of the Kuban State Technological University. 2024; (5): 111–118.

Список источников

1. Белоусов Н.А., Махов Д.С., Захаров М.Ю. О методах нахождения вероятностей ошибок 1-го и 2-го рода при бинарной классификации: URL: https://ntk.kubstu.ru/data/mc/0105/4826.pdf. Дата вхождения 1.11.2024.

2. Корн Г., Корн Т. Справочник по математике для научных работников и инженеров. М., 1973. 832 с.

3. Гмурман В.Е. Теория вероятностей и математическая статистика. М., 2023. 479 с.

References

1. Belousov N.A., Makhov D.S., Zakharov M.Yu. On methods for finding the probabilities of errors of the 1st and 2nd kind in binary classification: URL: https://ntk.kubstu.ru/data/mc/0105/4826.pdf. Date of entry: 1.11.2024.

2. Korn G., Korn T. *Spravochnik po matematike dlya nauchnyh rabotnikov i ingenerov* [Handbook of Mathematics for researchers and engineers]. Moscow, 1973. 832 c.

3. Gmurman V.E. *Teoria veroyatnostei I matematicheskaya statistika* [Probability theory and mathematical statistics]. Moscow, 2023. 479 c.