2.7. Непараметрические алгоритмы

Непараметрические методы обнаружения и различения характерны тем, что не основываются на знании функционального вида распределений. В широком смысле *непараметрические правила и алгоритмы* обработки предназначены для применения к некоторому непараметризованному множеству допустимых распределений [15]. Например выборочное среднее и дисперсия служат непараметрическими оценками математического ожидания и дисперсии соответственно. Непараметрическими являются алгоритмы формирования спектральных отсчетов: дискретное преобразование Фурье (ДПФ), быстрое преобразование Фурье (БПФ) и др.

Непараметрические алгоритмы позволяют решить общую задачу о *принадлежности* выборки к *тому или иному* распределению (*тесты согласия*), т. е. задачу *различения вида* распределений для одной выборки: гипотетического $F_0(y)$ и альтернативного $F_1(y)$. Частными являются альтернативы *сдвига* $F_1(y) = F_0(y-a)$ или *изменения масштаба* $F_1(y) = F_0(y/a)$, где a – некоторая постоянная, а также нарушения *симметрии* плотности распределения $f_0(y)$ (для симметричной плотности математическое ожидание совпадает с медианой).

В других случаях проверяется *стационарность* и альтернативой служит появление *разладки* (скачкообразного изменения), наличие *тренда* (изменения математического ожидания или иных характеристик) и т. д. Проверка гипотезы о *независимости значений* выборки из произвольного распределения также является непараметрической задачей.

В случае двух выборок обычно проверяется гипотеза *одинаковости* распределений $F(y)$ и $G(x)$ относительно альтернативы общего вида $F \neq G$. Фактически проверяется *однородность составной выборки*.

В число непараметрических алгоритмов входят *знаковые*, *ранговые* и *знаково-ранговые* алгоритмы. Большинство непараметрических статистик являются *инвариантами* по отношению к некоторому мешающему фактору.

Знаковые алгоритмы используют только знаки элементов выборки. Знаковая функция

$$\operatorname{sgn}(y) = \frac{y}{|y|} = \begin{cases} 1, & y \geq 0, \\ -1, & y < 0 \end{cases}$$

осуществляет нелинейное преобразование выборки, инвариантное к абсолютным величинам, в результате которого выборка становится набором из единиц с плюсами и минусами, образуя так называемый «знаковый вектор».

Эквивалентное преобразование осуществляет *функция единичного скачка* (или «степовая» функция)

$$\operatorname{step}(y) = \begin{cases} 1, & y \geq 0, \\ 0, & y < 0, \end{cases}$$

поскольку $\operatorname{sgn}(y) = 2\operatorname{step}(y) - 1$. Очевидно, знаковый вектор $\operatorname{sgn}(y) = (\operatorname{sgn}(y_1), \operatorname{sgn}(y_2), \ldots, \operatorname{sgn}(y_N))^T$ является *инвариантом* к абсолютной величине сигнала.

Ранговые алгоритмы учитывают не только факт, но и степень различия значений. Для определения ранга исходная выборка $y = (y_1, y_2, \ldots, y_N)$ упорядочивается *по возрастанию* (иногда используется упорядочивание по убыванию) значений элементов. Преобразованная выборка представляет так называемый *вариационный ряд* $y_{()} = (y_{(1)}, y_{(2)}, \ldots, y_{(N)})$. В данном случае $y_{(1)} = y_{\min}$, $y_{(N)} = y_{\max}$. Существует множество *порядковых* статистик, использующих вариационный ряд. К ним относится *выборочная медиана* $\bar{y}_{\text{med}}$, которая для нечетного N равна $y_{((N+1)/2)}$, *размах* выборки $R = y_{(N)} - y_{(1)}$ и другие. При четном N выборочная медиана вычисляется как полусумма $y_{(N/2)}$ и $y_{(N/2+1)}$.

Ранг R_i элемента y_i есть *номер этого элемента в вариационном ряду*. Ранг можно вычислить через знаковые или степовые функции следующим образом

$$R_i = \sum_{k=1}^{N} \operatorname{step}(y_i - y_k) + 1 = \frac{1}{2}\sum_{k=1}^{N} \operatorname{sgn}(y_i - y_k) + \frac{N}{2} + 1.$$

Нетрудно видеть, что сумма рангов всех элементов

$$\sum_{i=1}^{N} R_i = \sum_{i=1}^{N} i = \frac{N(N+1)}{2}.$$

Для *однородной независимой* выборки функция правдоподобия инвариантна к группе перестановок аргументов. Отсюда следует, что для указанной выборки все ранговые векторы *равновероятны*, каково бы ни было распределение выборки [14].

Общее число возможных ранговых векторов, соответствующих выборке размера N, равно числу перестановок N чисел, т. е.

равно $N!$. Поэтому для любого распределения однородной независимой выборки вероятность каждого рангового вектора $P(R_i) = 1/N!$. Следовательно, при использовании рангового алгоритма обнаружения сигнала сохраняется неизменной вероятность ложной тревоги для стационарной независимой помехи с произвольным распределением и в этом смысле ранговый алгоритм является *непараметрическим*.

Ранговый вектор $R = (R_1, R_2, ..., R_N)$ инвариантен также к любому нелинейному преобразованию, которое не изменяет относительного расположения элементов выборки.

В некоторых задачах используются вектор абсолютных величин наблюдений $y^+ = (|y_1|, |y_2|, ..., |y_N|)$ и вектор положительных рангов $R^+ = (R_1^+, R_2^+, ..., R_N^+)$, состоящий из порядковых номеров членов вариационного ряда из абсолютных величин.

Ряд *тестов согласия* используют *эмпирическую функцию распределения* [10] $F^*(y) = (1/N)\sum_{k=1}^{N} \mathrm{step}(y - y_{(k)})$. Таким образом, $F^*(y) = k/N$ для $y_{(k)} \le y < y_{(k+1)}$.

При $N \to \infty$ эта функция *сходится по вероятности* к истинной функции распределения $F(y)$. Это означает, что последовательность случайных величин $\zeta_N = \sup_y |F^*(y) - F(y)|$ сходится по вероятности к нулю (к нулю стремятся *вероятности* ненулевых значений последовательности).

Проверяется гипотеза H_0: выборка y имеет распределение $F_0(y)$ против альтернативы H_1: y имеет распределение $F_1(y)$, отличное от $F_0(y)$, т. е. $F_1(y) \ne F_0(y)$. В случае односторонней альтернативы $H_1^{(1)}$: $F_1(y) > F_0(y)$, или $H_1^{(2)}$: $F_1(y) < F_0(y)$. Гипотезы будут простыми, если функции распределения полностью определены, и сложными в других случаях. *Непараметрическими* гипотезы будут тогда, когда соответствующий класс распределений не описывается *конечной* совокупностью параметров.

Пусть решающая статистика $z = \psi\{F^*(y) - F_0(y)\}$ зависит от разности *эмпирической* и *гипотетической* функций распределения. Распределение этой статистики при гипотезе H_0 не зависит от гипотетического распределения $F_0(y)$, т. е. алгоритм имеет непараметрические свойства.

Наряду с одновыборочными тестами широко используются тесты на принадлежность двух выборок одному и тому же распределению. Фактически они проверяют *однородность* составной выборки.

Наиболее известные тесты согласия приведены в табл. 2.1. *Тест Колмогорова–Смирнова* использует верхнюю границу модулей разности $z = \sup_y |F^*(y) - F_0(y)|$.

Таблица 2.1

Тесты согласия	Решающая статистика	Примечания
Тест Колмогорова–Смирнова - односторонний - двусторонний - двухвыборочный односторонний - двухвыборочный двусторонний	$z = \sup_y \{F^*(y) - F_0(y)\}$ $z = \sup_y \lvert F^*(y) - F_0(y)\rvert$ $z = \sup_y \{F^*(y) - G^*(x)\}$ $z = \sup_y \lvert F^*(y) - G^*(x)\rvert$	$F^*(y) = (1/n)\sum_{k=1}^{n} \mathrm{step}(y - y^{(k)})$; $G^*(y) = (1/n)\sum_{k=1}^{n} \mathrm{step}(x - x^{(k)})$ – выборочные функции распределения
Тест Реньи - односторонний - двусторонний	$z = \sup_y \left\{ \frac{F^*(y) - F_0(y)}{F_0(y)} \right\}$ $z = \sup_y \left\lvert \frac{F^*(y) - F_0(y)}{F_0(y)} \right\rvert$	$F_0(y) > 0$ – опорная функция распределения
Тест Крамера–Мизеса–Смирнова	$z = \int_{-\infty}^{\infty} F^*(y) - F_0(y)]^2 \cdot \psi[F_0(y)]dF_0(y)$	$\psi(\cdot)$ – весовая функция
Тест Манна–Уитни	$z = \sum_{j=1}^{N} \sum_{i=1}^{M} \mathrm{step}(x_i - y_j)$	$x = (x_1, \dots, x_M)$ $y = (y_1, \dots, y_N)$
Тест χ^2	$z = \sum_{i=1}^{l} \frac{(\nu_i - np_i)}{np_i}$	ν_i – число выборочных значений в интервале Δ_i, p_i – вероятность попадания в Δ_i, $\sum_{i=1}^{l} \nu_i = n$