УДК 519.2

М.С. АБРАМОВИЧ, М.Н. МИЦКЕВИЧ

КРИТЕРИИ ОБНАРУЖЕНИЯ РАЗЛАДОК БИНАРНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ, ОСНОВАННЫЕ НА ВЕЙВЛЕТ – ПРЕОБРАЗОВАНИИ

The tests for detecting disorders in a binary random sequences using the wavelet transform are proposed and their performance is investigated using the statistical modelling.

Введение

Обнаружение разладок («скачкообразного изменения» вероятностных свойств) бинарных последовательностей является актуальной прикладной задачей [1]. В последнее десятилетие для выявления локальных особенностей последовательностей и, в частности, разладок широко применяется вейвлет – анализ [2]. Подход, основанный на применении вейвлет-анализа, для обнаружения разладки, заключается в том, что, согласно [2], абсолютные значения вейвлет-коэффициентов в силу свойства локальности вейвлет – преобразования имеют максимальные значения в момент разладки.

В настоящей работе рассматриваются критерии обнаружения разладок бинарных последовательностей, основанные на использовании вейвлет-преобразования.

Математическая модель

Рассмотрим бинарную последовательность независимых случайных величин:

$$X(t) = \left\{ x_t \mid t = \overline{0, T-1} \right\},\ x_t \in \{\pm 1\}, T = 2^M, M \in N. \qquad (1)$$

Последовательность (1) имеет разладку в неизвестный момент времени $\tau \in \{\tau_-, \tau_- + 1, \dots, \tau_+\}$, где $0 < \tau_- < \tau_+ < T - 1 -$ некоторые априорно заданные граничные значения, и состоит из фрагментов бинарных последовательностей:

$$X(t) = \begin{cases} X_1(t), & 0 \le t \le \tau - 1, \\ X_2(t), & \tau \le t \le T - 1, \end{cases}$$

$$X_1(t):\ P(x_t = 1) = p_1,\ X_2(t): P(x_t = 1) = p_2, \text{ где } p_1, p_2 \in [0,1].$$

Сформулируем нулевую H_0 и альтернативную H_1 гипотезы о распределении вероятностей последовательности (1) следующим образом: $H_0: p_1 = p_2 = 0.5$, $H_1: p_1 \neq p_2$.

Определим дискретное диадное вейвлет-преобразование последовательности (1)[3]:

$$d_{j,k}^{(\psi)} = \sum_{t=0}^{T-1} x_t \psi_{j,k}(t),\ j = \overline{1,M},\ k = \overline{0, 2^{M-j}-1}, \tag{2}$$

где $\psi_{j,k}(t) = 2^{-\frac{j}{2}} \psi\left(2^{-j}t - k\right)$, $\psi(t)$ - материнский вейвлет, j - параметр масштаба (уровень разрешения), k - параметр сдвига.

В качестве материнского вейвлета будем использовать вейвлет Хаара, который в [3] рекомендуется применять для выявления разладок последовательностей:

$$\psi(t) = \begin{cases} 1,\ 0 \le t < 0.5, \\ -1,\ 0.5 \le t < 1, \\ 0,\ t < 0,\ t \ge 1. \end{cases}$$

Отметим, что коэффициенты дискретного вейвлет-преобразования (2) последовательности (1) при использовании вейвлета Хаара принимают значения из следующего множества:

$$\left\{ d_j(i) ::= \pm 2^{1-\frac{j}{2}} i,\ j = \overline{1,M}, i = \overline{0, 2^{j-1}} \right\}. \tag{3}$$

Статистические свойства коэффициентов дискретного вейвлет-преобразования

Теорема 1. Если верна гипотеза H_0, то коэффициенты вейвлет-преобразования (2) последовательности (1), построенные с использованием вейвлета Хаара, на каждом уровне разрешения $j \in \{1, \ldots, M\}$ независимы и имеют следующее распределение вероятностей:

$$P\left\{ d_{j,k}^{(\psi)} = d_j(i) \right\} = C_{2^j}^{2^{j-1}-i} \Big/ 2^{2^j} = \frac{(2^j)!}{2^{2^j}(2^{j-1}-i)!(2^j - 2^{j-1}+i)!}. \tag{4}$$

Доказательство. Коэффициенты вейвлет-преобразования (2), построенные с использованием вейвлета Хаара, можно представить в виде:

$$d_{j,k}^{(\psi)} = \sum_{t=0}^{T-1} x_t \psi_{j,k}(t) = 2^{-\frac{j}{2}} \left(\sum_{t=2^j k}^{2^j\left(k+\frac{1}{2}\right)-1} x_t - \sum_{t=2^j\left(k+\frac{1}{2}\right)}^{2^j(k+1)-1} x_t \right) = 2^{-\frac{j}{2}} \left(S_m^{(1)} - S_m^{(2)} \right), \qquad (5)$$

где $m = 2^{j-1}$, $S_m^{(1)}$ и $S_m^{(2)}$ – частичные суммы симметричного случайного блуждания [4].

Из (5) следует независимость вейвлет-коэффициентов $d_{j,k}^{(\psi)}$ и $d_{j,l}^{(\psi)}$, $l \neq k$ на каждом уровне разрешения j, так как интервалы, по которым вычисляются вейвлет-коэффициенты, не пересекаются.

В случае выполнения нулевой гипотезы H_0 выражение (5) для вейвлет-коэффициентов на каждом уровне разрешения представляет собой разность двух частичных сумм симметричного случайного блуждания и, очевидно, также является частичной суммой симметричного случайного блуждания:

$$2^{\frac{j}{2}} d_{j,k}^{(\psi)} = S_m^{(1)} - S_m^{(2)} = S_{2m}. \qquad (6)$$

Для частичных сумм симметричного случайного блуждания справедливо следующее распределение вероятностей [4]:

$$p_{n,r} = P\{S_n = r\} = C_n^{\frac{n+r}{2}} 2^{-n}, \ r \in \{-n, -n+1, \ldots, n-1, n\}.$$

Отсюда для частичной суммы (6) имеем:

$$P\{S_{2m} = r\} = C_{2m}^{\frac{2m+r}{2}} 2^{-2m}, \ r \in \{-2m, -2m+1, \ldots, 2m-1, 2m\}.$$

Подставляем в эту формулу $m = 2^{j-1}$ и получаем:

$$P\left\{ 2^{\frac{j}{2}} d_{j,k}^{(\psi)} = r \right\} = \frac{C_{2^j}^{\frac{2^j + r}{2}}}{2^{2^j}}, \quad r \in \left\{ -2^j, -2^j + 1, \ldots, 2^j - 1, 2^j \right\}.$$

Обозначив $r = -2i$, имеем (4).

Следствие. Если верна гипотеза H_0, то коэффициенты вейвлет-преобразования (2) при $T \to \infty$ на каждом уровне разрешения $j = \overline{1, M}$ имеют асимптотически нормальное стандартное распределение.

Доказательство. Частичные суммы симметричного случайного блуждания (6) при $j \to \infty$ ($T \to \infty$) имеют асимптотически нормальное стандартное распределение [4]: $S_{2^j} \Big/ \sqrt{2^j} \xrightarrow{D} N(0,1)$.

Поэтому для коэффициентов вейвлет-преобразования (5) имеем: $2^{\frac{j}{2}} d_{j,k}^{(\psi)} \Big/ 2^{\frac{j}{2}} = d_{j,k}^{(\psi)} \xrightarrow{D} N(0,1)$ при $T \to \infty$.

Отметим, что как следует из (5), распределение вейвлет-коэффициентов (4) более точно аппроксимируется нормальным распределением на уровнях разрешения $j > M - j_0$, $j_0 < [M/2]$.

На основании следствия из теоремы 1 исходные гипотезы H_0 и H_1 о распределении вероятностей последовательности (1) можно заменить набором гипотез H_{0j} и H_{1j} о распределении коэффициентов вейвлет-преобразования на каждом уровне разрешения $j = \overline{1, M}$:

$$H_{0j} : d_{j,k}^{(\psi)} \sim N(0,1),$$

H_{1j} : иначе.

Критерий, основанный на максимуме вейвлет-периодограммы

В соответствии с [3] для каждого уровня разрешения определим вейвлет - периодограмму последовательности (1):

$$I_{j,k}^{(\psi)} = \left(d_{j,k}^{(\psi)} \right)^2, k = \overline{0, 2^{M-j} - 1}. \tag{7}$$

На основании следствия из теоремы 1 вейвлет-периодограмма (7) при $T \to \infty$ имеет асимптотически хи-квадрат распределение с одной степенью свободы:

$$I_{j,k}^{(\psi)} \xrightarrow{D} \chi_1^2. \tag{8}$$

Определим максимум вейвлет-периодограмм на каждом уровне разрешения:

$$I_{j,\max}^{(\psi)} = \max_k I_{j,k}^{(\psi)}. \tag{9}$$

Найдем асимптотическое распределение вероятностей статистики (9). Пусть $F_{I_{j,\max}^{(\psi)}}(x)$ - функция распределения максимума вейвлет-периодограммы на уровне разрешения j, $F_{\chi_1^2}(x)$ - функция распределения хи-квадрат с одной степенью свободы, $L_j = 2^{M-j}$.

Теорема 2. Если верна гипотеза H_0, то статистика (9) при $T \to \infty$ имеет асимптотическое распределение вероятностей:

$$(F_{I^{(\psi)}_{j,\max}}(x) - F^{L_j}_{\chi^2_1}(x)) \xrightarrow{D} 0, \; x \ge 0.$$

Доказательство. Вейвлет-периодограммы на каждом уровне разрешения являются независимыми, как борелевские функции от коэффициентов вейвлет-преобразования $d^{(\psi)}_{j,k}$, которые являются независимыми и одинаково распределенными согласно теореме 1. Поэтому, учитывая (8) и определение функции распределения имеем:

$$F_{I^{(\psi)}_{j,\max}}(x) = P\left\{I^{(\psi)}_{j,\max} \le x\right\} = P\left\{I^{(\psi)}_{j,0} \le x,\; I^{(\psi)}_{j,1} \le x, \dots, I^{(\psi)}_{j,L_j-1} \le x\right\} =$$

$$= \prod_{i=0}^{L_j-1} P\left\{I^{(\psi)}_{j,i} \le x\right\} = F^{L_j}_{\chi^2_1}(x).$$

Критерий для обнаружения разладки последовательности (1), основанный на максимуме вейвлет-периодограммы, строится следующим образом.

На каждом уровне разрешения $j = \overline{1,M}$ проверяются гипотезы H_{0j} и H_{1j}. Решающее правило состоит в следующем:

$$\textit{принимается гипотеза} \begin{cases} H_{0j}, \; \textit{если} \;\; P_j > \alpha^*, \\ H_{1j}, \; \textit{если} \;\; P_j \le \alpha^*, \end{cases}$$

где $P-$значение для каждого уровня разрешения вычисляется как $P_j = 1 - F^{L_j}_{\chi^2_1}\left(I^{(\psi)}_{j,\max}\right)$, а для набора из M критериев в соответствии с [5] $\alpha^* \le \alpha/M$, α - уровень значимости критерия.

Гипотеза H_0 принимается в том случае, если на всех уровнях разрешения $j = \overline{1,M}$ принимаются гипотезы H_{0j}, а альтернативная гипотеза H_1 принимается, когда хотя бы на одном уровне разрешения принимается гипотеза H_{1j}.

Критерий, основанный на скалограмме

В соответствии с [3] определим скалограмму $S^{(\psi)}(j)$:

$$S^{(\psi)}(j) = \sum_{k=0}^{2^{M-j}-1} [d^{(\psi)}_{j,k}]^2, \quad j = \overline{1,M}. \tag{10}$$

На основании следствия из теоремы 1 скалограмма (10) имеет при $T \to \infty$ асимптотически χ^2-распределение с $L_j = 2^{M-j}$ степенями свободы как сумма квадратов независимых стандартных гауссовских величин: $(F_{S^{(\psi)}(j)}(x) - F_{\chi^2_{L_j}}(x)) \xrightarrow{D} 0,\ x \ge 0.$

Критерий для обнаружения разладки последовательности (1), основанный на скалограмме, аналогичен по построению критерию, основанному на максимуме вейвлет-периодограммы, при этом $P-$значение для каждого уровня разрешения определяется как $P_j = 1 - F_{\chi^2_{L_j}}\left(S^{(\psi)}(j)\right)$.

Критерий, основанный на пороговом значении для вейвлет-коэффициентов

Идея этого непараметрического критерия состоит в том, что если рассматриваемая последовательность (1) не имеет разладок, то абсолютные значения вейвлет-коэффициентов на каждом уровне разрешения не превосходят определенного порогового значения. В качестве порогового значения будем использовать универсальное пороговое значение для всех уровней разрешения, подобное предложенному в [6] для сглаживания стационарных временных рядов: $\lambda = \sigma\sqrt{2\log_2 T}$, где $\sigma-$ среднее квадратическое отклонение элементов анализируемой последовательности (1).

Критерий обнаружения разладки, основанный на использовании порогового значения, состоит в следующем: последовательность имеет разладку, если хотя бы на одном уровне разрешения выполняется условие: $\left|d^{(\psi)}_{j,k}\right| > \lambda,\ j = \overline{1, M},\ k = \overline{0, 2^{M-j} - 1}$.

Замечание. Если разладка в последовательности (1) обнаружена, то для критериев, основанных на максимуме вейвлет-периодограммы и пороговом значении, оценку момента разладки можно определить по следующей формуле: $\hat{\tau} = 2^{j^*} \cdot (k^* + 0.5)$, где $(j^*, k^*) = \arg\max_{j,k} |d^{(\psi)}_{j,k}|$.

Анализ эффективности критериев методом статистического моделирования

Число моделируемых бинарных последовательностей для оценки вероятностей ошибок первого и второго рода рассмотренных выше критериев – $K = 100$. Проверка гипотез H_0 и H_1 проводилась на уровнях разрешения $(j > M - 3)$.

Для оценки вероятностей ошибок первого рода моделировались бинарные последовательности длиной $T = 2^8, 2^{10}, 2^{12}$ с вероятностью $p_1 = 0.5$. Уровень значимости для проверки нулевой гипотезы H_0 задавался равным α=0.05. В таблице 1 приведены оценки вероятностей ошибок первого рода $\hat{\alpha}$ критериев в зависимости от длины последовательности T, которые сравнимы с уровнем значимости.

Таблица 1

Оценки вероятностей ошибок первого рода

Критерий\T	2^8	2^{10}	2^{12}
Максимума вейвлет-периодограммы	0.08	0.07	0.07
Скалограммы	0.09	0. 06	0.05
Порогового значения	0.01	0.01	0.01

Для оценки вероятностей ошибок второго рода моделируемые последовательности состояли из двух однородных фрагметов длиной $T_1 = T_2 = T/2$ с разладкой в момент времени соответственно $\tau = 2^7, 2^9, 2^{11}$. Первый фрагмент представлял случайную бинарную последовательность с вероятностями $p_1 = p_2 = 0.5$, а для второго фрагмента вероятность p_1 принимала значения $p_1 =$ {0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45}. В таблицах 2-4 приведены оценки вероятностей ошибок второго рода $\hat{\beta}$ рассмотренных критериев в зависимости от вероятности p_1 и длины последовательности T.

Таблица 2

Оценки вероятностей ошибок второго рода критерия, основанного на максимуме вейвлет-периодограммы

$T \setminus p_1$	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45
2^8	0.00	0.00	0.00	0.01	0.06	0.17	0.54	0.73	0.87
2^{10}	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.22	0.77
2^{12}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13

Таблица 3

Оценки вероятности ошибки второго рода критерия, основанного на скалограмме

$T \setminus p_1$	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45
2^8	0.00	0.00	0.00	0.01	0.09	0.28	0.53	0.83	0.92
2^{10}	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.25	0.90
2^{12}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.34

Таблица 4

Оценки вероятностей ошибок второго рода критерия, основанного на пороговом значении для вейвлет-коэффициентов

$T \setminus p_1$	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45
2^8	0.00	0.00	0.02	0.12	0.37	0.84	0.95	0.98	0.99
2^{10}	0.00	0.00	0.00	0.00	0.00	0.01	0.40	0.89	0.98
2^{12}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.96

Из таблиц 2-4 видно, что оценки вероятностей ошибок второго рода $\hat{\beta}$ уменьшаются с ростом длины последовательности T, а наименьшие оценки вероятностей ошибок второго рода имеет критерий, основанный на максимуме вейвлет-периодограммы.

Результаты проведенных экспериментов показывают работоспособность и эффективность предложенных критериев обнаружения разладки бинарных последовательностей.

Данные исследования были частично поддержаны в рамках ГКПНИ «Инфотех».

1. Харин Ю.С., Абрамович М.С. Об обнаружении спектральной «разладки» двумерного временного ряда // Автометрия. 1999. № 2.

2. Abramovich F., Bailey T., Sapatinas T. Wavelet analysis and statistical applications // JRSSD. 2000. № 48.

3. Chiann C., Morettin P. A. A Wavelet Analysis for Times Series // Journal of Nonparametric Statistics. 1998. № 10.

4. Феллер В. Введение в теорию вероятностей и ее применения: В 2 т. М.: Наука. 1984.

5. Кокс Д., Хинкли Д. Теоретическая статистика. М.: Наука. 1978.

6. Donoho, D. L. and Johnstone, I. M. Minimax Estimation via Wavelet shrinkage // Annals of Statistics. 1998. Vol. 26. № 3.