МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ
MATHEMATICAL MODELING

УДК 519.816, 519.226, 519.254, 519.244.3
https://doi.org/10.37661/1816-0301-2021-18-3-36-47

Оригинальная статья
Original Paper

Последовательное статистическое принятие решений в задачах анализа потоков данных

А. Ю. Харин

Белорусский государственный университет,
пр. Независимости, 4, Минск, 220030, Беларусь
E-mail: KharinAY@bsu.by

Аннотация. В задачах анализа потоков данных актуальны проблемы статистического принятия решений о параметрах наблюдаемых потоков. Для их решения в работе предлагается использовать последовательные статистические решающие правила. Такие правила построены в статье для трех моделей потоков наблюдений: последовательности независимых однородных наблюдений; последовательности наблюдений, образующих временной ряд с трендом; последовательности зависимых наблюдений, образующих однородную цепь Маркова. Для каждого случая рассмотрена также ситуация, когда модель описывает наблюдаемые стохастические данные с искажениями. В качестве допустимых искажений используются «выбросы» («засорения»), которые адекватно описывают наиболее часто встречающиеся на практике ситуации. Предложены семейства последовательных решающих правил, в рамках которых строятся робастные решающие правила, позволяющие снизить влияние искажений на характеристики эффективности. Для иллюстрации преимуществ построенных решающих правил приводятся результаты компьютерных экспериментов.

Ключевые слова: последовательное решающее правило, статистический тест, временной ряд с трендом, однородная цепь Маркова, искажения

Благодарности. Результаты исследования получены при выполнении НИР «Библиотека процедур для компьютерного мониторинга данных и принятия решений на основе методов последовательного анализа» в рамках ГПНИ «Цифровые и космические технологии, безопасность человека, общества и государства», а также при поддержке стипендии Президента Республики Беларусь талантливым молодым ученым на 2021 г.

Для цитирования. Харин, А. Ю. Последовательное статистическое принятие решений в задачах анализа потоков данных / А. Ю. Харин // Информатика. – 2021. − Т. 18, № 3. – С. 36–47. https://doi.org/ 10.37661/1816-0301-2021-18-3-36-47

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Поступила в редакцию | Received 02.08.2021
Подписана в печать | Accepted 01.09.2021
Опубликована | Published 29.09.2021

Sequential statistical decision making in problems of data flows analysis

Alexey Y. Kharin

Belarusian State University,
av. Nezavisimosti, 4, Minsk, 220030, Belarus
E-mail: KharinAY@bsu.by

Abstract. In the problems of data flows analysis, the problems of statistical decision making on parameters of observed data flows are important. For their solution it is proposed to use sequential statistical decision rules. The rules are constructed for three models of observation flows: sequence of independent homogeneous observations; sequence of observations forming a time series with a trend; sequence of dependent observations forming a homogeneous Markov chain. For each case the situation is considered, where the model describes the observed stochastic data with a distortion. "Outliers" ("contamination") are used as the admissible distortions that adequately describe the majority of situations appear in practice. For such situations the families of sequential decision rules are proposed, and robust decision rules are constructed that allow to reduce influence of distortion to the efficiency characteristics. The results of computer experiments are given to illustrate the constructed decision rules.

Keywords: sequential decision rule, statistical test, time series with trend, homogeneous Markov chain, distortion

Acknowledgements. The results of the paper are obtained within the Project "Procedures library for computer monitoring of data and decision making on the basis of sequential analysis methods" in frames of the State research program "Digital and space technologies, human and state safety", and with the support of the grant of the President of the Republic of Belarus of talented young researchers for the year 2021.

For citation. Kharin A. Y. Sequential statistical decision making in problems of data flows analysis. *Informatics,* 2021, vol. 18, no. 3, pp. 36–47 (In Russ.). https://doi.org/10.37661/ 1816-0301-2021-18-3-36-47

Conflict of interest. The author declare of no conflict of interest.

Введение. В современных прикладных задачах часто возникает необходимость анализа потоков стохастических данных с целью принятия решения об одном из двух режимов функционирования системы, порождающей такой поток [1] (например, «корректная работа» и «типовой сбой»). Режим функционирования характеризуется значением параметра (вектора параметров) распределения вероятностей, моделирующего наблюдения в потоке.

Для принятия решения в пользу одного из режимов в прикладных задачах эффективным подходом представляется последовательный статистический анализ [2]. В нем делается предположение о том, что число наблюдений, необходимых для принятия решения с заданной точностью (малыми значениями вероятностей ошибок), априори не фиксируется, а определяется в зависимости от самих стохастических наблюдений и является случайной величиной. Происходит «подстраивание» числа необходимых наблюдений по мере их поступления к сложности задачи принятия решения для конкретной ситуации, сформированной реальными наблюдениями. Другими словами, после каждого наблюдения имеются две возможности: принимать окончательное решение в пользу одного из двух возможных режимов либо принимать решение о том, что при полученных наблюдениях требуемая точность не может быть обеспечена и необходимо следующее наблюдение. Такая «гибкость» позволяет минимизировать среднее число наблюдений, необходимых для обеспечения заданной точности решений, однако создает значительные сложности теоретического анализа характеристик эффективности последовательных статистических решающих правил (условных вероятностей ошибок и математических ожиданий случайного числа необходимых наблюдений) [3].

Несмотря на трудности теоретического анализа, последовательные статистические решающие правила интенсивно используются в медицине, мониторинге режимов функционирования сложных технических объектов, контроле качества производимой продукции и во многих дру-

гих областях [1]. Эти правила представляют собой естественную схему вовлечения получаемой информации в процесс принятия решений по мере поступления наблюдений и позволяют существенно экономить число необходимых наблюдений, стоимость каждого из которых может быть значительной [2, 3].

Подход, разработанный в книге [4], дает возможность вычислять характеристики эффективности последовательных тестов не только в случае полного соответствия модели тем потокам данных, для описания которых она построена, но и при наличии искажений [5] – отклонений вероятностных законов, характеризующих наблюдения, от гипотетических модельных предположений.

В настоящей работе рассмотрены три модели потока стохастических данных: последовательность независимых одинаково распределенных случайных наблюдений; последовательность неоднородных наблюдений, образующих временной ряд с трендом, и последовательность зависимых наблюдений, образующих однородную цепь Маркова. Для каждой модели представлено последовательное решающее правило. Предложены семейства последовательных решающих правил, позволяющие строить робастные [6] решающие правила, устойчивые к влиянию искажений на характеристики эффективности.

Поток данных, представляющих независимые однородные наблюдения. Рассмотрим вначале простейшую модель потока данных. Пусть наблюдается поток $x_1, x_2, \ldots \in U \subseteq \mathbf{R}^N$ независимых случайных векторов с распределением вероятностей P_θ, имеющим без ограничения общности плотность распределения вероятностей $p_\theta(x)$ относительно некоторой меры $\mu(x)$, $x \in U \subseteq \mathbf{R}^N$, где $\theta \in \Theta = \{0,1\}$ – значение параметра, характеризующего поток; наблюдателю это значение неизвестно. (В случае дискретных распределений вероятностей здесь вместо привычной плотности распределения вероятностей относительно меры Лебега имеются в виду вероятности соответствующих событий.)

Сформулированы две гипотезы H_0 и H_1 о значении параметра, которые соответствуют двум режимам работы системы, порождающей поток случайных наблюдений:

$$H_0 : \theta = 0,\ H_1 : \theta = 1. \tag{1}$$

Последовательное решающее правило задается двумя компонентами: моментом остановки и терминальным решением. Первая компонента представляет собой функцию от случайных наблюдений и указывает номер наблюдения (случайный вследствие этой зависимости), после которого принимается решение о завершении процесса наблюдения. До момента остановки каждый раз принимается решение о продолжении этого процесса и необходимости получения следующего наблюдения, поскольку требуемая точность в контексте малых значений вероятностей возможных ошибок не может быть обеспечена совокупностью полученных наблюдений. В момент остановки принимается терминальное решение в пользу одной из гипотез (1) в соответствии с правилом, определяемым второй компонентой, на основании всех наблюдений, полученных к моменту остановки.

Рассмотрим семейство последовательных решающих правил $\delta_\lambda = (\tau_\lambda, d_\lambda)$, представленных в виде упорядоченных пар и основанных на функции $\lambda(\cdot) : U \to \mathbf{R}$. Здесь

$$\tau_\lambda = \inf\left\{n : \Lambda_n \notin (C_-, C_+)\right\} - \tag{2}$$

зависящий от наблюдений $x_1, \ldots, x_n$ случайный момент остановки процесса наблюдения (первый момент выхода из интервала между порогами). Терминальное решение $d_\lambda = i$, $i \in \{0,1\}$, означает выбор гипотезы H_i в качестве результата в соответствии с правилом

$$d_\lambda = \mathbf{1}_{[C_+, +\infty)}(\Lambda_n). \tag{3}$$

В задании компонент (2), (3)

$$\Lambda_n = \Lambda_n(x_1, \ldots, x_n) = \sum_{t=1}^{n} \lambda(x_t)\ ,\ n \in \mathbf{N} = \{1, 2, 3, \ldots\}, - \tag{4}$$

критериальная статистика; C_-, $C_+ \in \mathbf{R}$ – параметры (называемые порогами) последовательного решающего правила (2)–(4), $C_- < C_+$; $\mathbf{1}_A(\cdot)$ – индикаторная функция принадлежности аргумента множеству *A*.

Последовательное статистическое решающее правило, основанное на статистике отношения правдоподобия, является элементом рассмотренного семейства последовательных решающих правил (2)–(4) при следующей функции:

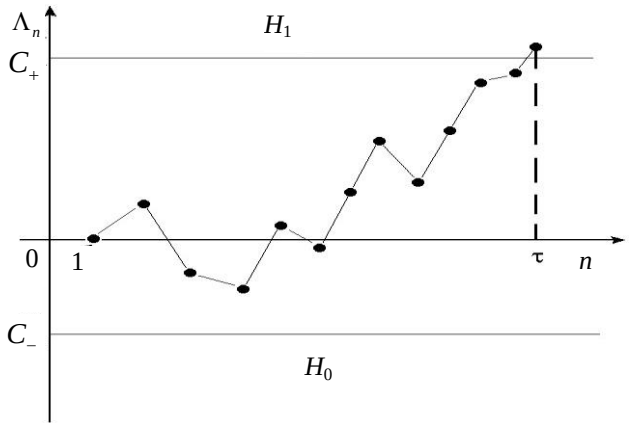
$$\lambda(u) = \lambda_W(u) = \log \tfrac{p_1(u)}{p_0(u)},\ u \in U\,. \tag{5}$$

На практике, как правило, значения порогов C_-, C_+ вычисляют по формулам [2]

$$C_- = \log \tfrac{\beta_0}{1-\alpha_0}\,,\ C_+ = \log \tfrac{1-\beta_0}{\alpha_0}\,, \tag{6}$$

где $\alpha_0, \beta_0 \in (0,1)$ – допустимые значения вероятностей ошибок первого рода (отвергается гипотеза H_0 ($d_\lambda = 1$) при условии, что она верна) и второго рода (отвергается гипотеза H_1 ($d_\lambda = 0$) при условии, что она верна) соответственно. Эти величины задаются пользователем решающего правила с учетом его представлений о допустимости соответствующих малых значений указанных характеристик. Логарифм берется по произвольному основанию. Важно лишь, чтобы при вычислении функции (5) и порогов (6) оно было одним и тем же.

Процесс изменения критериальной статистики в последовательном статистическом решающем правиле схематично изображен на рисунке.



Динамика критериальной статистики в процессе принятия решения в пользу H_1 последовательным правилом

Test statistic dynamics in the decision making process in favor of H_1 with the sequential rule

Фактические значения вероятностей ошибок первого и второго рода для теста, основанного на функции $\lambda(\cdot)$, обозначим соответственно

$$\alpha = \alpha(\delta_\lambda) = \mathrm{E}_0\left\{\mathrm{P}_0\left\{d_\lambda = 1 \mid \tau_\lambda\right\}\right\},\ \beta = \beta(\delta_\lambda) = \mathrm{E}_1\left\{\mathrm{P}_1\left\{d_\lambda = 0 \mid \tau_\lambda\right\}\right\}, \quad (7)$$

где $\mathrm{E}_\theta\{\cdot\}$ – математическое ожидание (среднее значение случайного аргумента) по распределению P_θ. Обозначим условное математическое ожидание случайного числа необходимых наблюдений, когда справедлива гипотеза $H_k,\ k \in \Theta$:

$$t_k = t_k(\delta_\lambda) = \mathrm{E}_k\left\{\tau_\lambda\right\}. \quad (8)$$

В контексте данной работы рассматривается множество из четырех характеристик эффективности последовательного статистического решающего правила $\delta_\lambda = (\tau_\lambda, \delta_\lambda)$, определенных соотношениями (7), (8).

В работе [2] доказано, что для последовательного решающего правила (2)–(6) проверки гипотез (1) о параметре наблюдаемого потока независимых однородных наблюдений выполнены следующие неравенства относительно вероятностей ошибочных решений (7):

$$\alpha(\delta_{\lambda_w}) \le \frac{\alpha_0}{1-\beta_0}, \beta(\delta_{\lambda_w}) \le \frac{\beta_0}{1-\alpha_0}.$$

При этом сумма фактических вероятностей ошибочных решений ограничена суммой заданных значений:

$$\alpha(\delta_{\lambda_w}) + \beta(\delta_{\lambda_w}) \le \alpha_0 + \beta_0.$$

Известно также, что для произвольных положительных значений α, β, в сумме меньших единицы, существуют значения порогов $C_-, C_+ \in \mathbf{R}$, для которых $\alpha(\delta_{\lambda_w}) = \alpha$, $\beta(\delta_{\lambda_w}) = \beta$, и при этом выполняется принцип оптимальности А. Вальда: для любого другого статистического решающего правила δ с фактическими вероятностями ошибочных решений, не превосходящими соответственно α и β, условные средние числа необходимых наблюдений не могут быть меньше, чем у δ_{λ_w}:

$$t_0(\delta_{\lambda_w}) \le t_0(\delta),\ t_1(\delta_{\lambda_w}) \le t_1(\delta).$$

Для условных математических ожиданий числа необходимых наблюдений известны приближенные выражения [14] (в предположении отличия от нуля знаменателей дробей в правых частях, $p_i(x) > 0$, $i = 0,1$, а также $0 < \alpha_0, \beta_0 < 1$)

$$t_0(\delta_{\lambda_w}) \approx \frac{(1-\alpha_0)\log\frac{\beta_0}{1-\alpha_0} + \alpha_0 \log\frac{1-\beta_0}{\alpha_0}}{\mathrm{E}_0\left\{\log\frac{p_1(x)}{p_0(x)}\right\}},\ t_1(\delta_{\lambda_w}) \approx \frac{\beta_0 \log\frac{\beta_0}{1-\alpha_0} + (1-\beta_0)\log\frac{1-\beta_0}{\alpha_0}}{\mathrm{E}_1\left\{\log\frac{p_1(x)}{p_0(x)}\right\}}.$$

Доказано [15], что последовательное решающее правило (2)–(6) с вероятностью единица завершается за конечное число наблюдений при выполнении условий, которые на практике интерпретируются как «различимость» соответствующих распределений вероятностей и, как правило, выполняются.

Фактические значения $t_0(\delta_{\lambda_W})$, $t_1(\delta_{\lambda_W})$ могут существенно отличаться от указанных выше приближений, а задаваемые значения α_0, β_0 – от фактических значений $\alpha(\delta_{\lambda_W}), \beta(\delta_{\lambda_W})$. Поэтому задача оценивания характеристик эффективности (7), (8) для последовательных решающих правил является важной [3]. Для решения этой задачи разработан подход [4], основанный на аппроксимации значений критериальной статистики последовательного решающего правила специальными цепями Маркова. Применим этот подход для простейшей модели независимых многомерных дискретных наблюдений [7], которая подвержена искажениям [8].

Пусть рассмотренная выше вероятностная модель потока наблюдений подвержена искажениям и наблюдения в потоке получены из смеси распределений вероятностей

$$\bar{P}_k(x) = (1-\varepsilon_k)P_k(x) + \varepsilon_k \tilde{P}_k(x),\ x \in U,\ k \in \{0,1\}, \tag{9}$$

где $\varepsilon_k \in [0, \varepsilon_+]$ – вероятность появления «выброса», или «засорения». Как правило, ее значение неизвестно при анализе поступающего потока данных, максимальный уровень искажения ε_+ известен в конкретной решаемой задаче; $\tilde{P}_k(u)$ – произвольное «засоряющее» распределение вероятностей, $u \in U$, $\tilde{P}_k(\cdot) \neq P_k(\cdot)$. С целью снижения влияния искажений (9) на характеристики эффективности построим семейство последовательных решающих правил $\delta_g = (\tau_g, d_g)$:

$$\tau_g = \inf\left\{n:\ \sum_{t=1}^{n} g(\lambda_W(x_t)) \notin (C_-, C_+)\right\},\ d_g = \mathbf{1}_{[C_+,+\infty)}\left(\sum_{t=1}^{n} g(\lambda_W(x_t))\right), \tag{10}$$

где

$$g(z) = g_- \mathbf{1}_{(-\infty, g_-)}(z) + z\mathbf{1}_{[g_-, g_+]}(z) + g_+ \mathbf{1}_{(g_+, +\infty)}(z),\ z \in \mathbf{R},$$

а $g_-, g_+ \in \mathbf{R}$ – дополнительные параметры последовательных решающих правил в рассматриваемом семействе, $g_- < g_+$.

Указанный выше подход при заданных значениях C_-, C_+, g_-, g_+, «засоряющих» распределениях вероятностей $\tilde{P}_k$, $k \in \{0,1\}$, и вероятностях появления «искажений» позволяет вычислять характеристики эффективности $\tilde{\alpha}(\delta_g)$, $\tilde{\beta}(\delta_g)$, $\tilde{t}_0(\delta_g)$, $\tilde{t}_1(\delta_g)$, определяемые аналогично выражениям (7), (8) для последовательных решающих правил (10) при наличии искажений (9).

Сформулируем критерий для построения робастного (устойчивого к искажениям (9)) последовательного решающего правила в рамках семейства (10):

$$\begin{cases} \sup\limits_{\{\tilde{P}_k\},\{\varepsilon_k\}} \left(w_0 \tilde{\alpha}(\delta_g; \tilde{P}_0, \varepsilon_0) + w_1 \tilde{\beta}(\delta_g; \tilde{P}_1, \varepsilon_1)\right) \to \min\limits_{g_-, g_+}, \\ \tilde{t}_0(\delta_g) + \tilde{t}_1(\delta_g) \le C \cdot \left(\tilde{t}_0(\delta_{\lambda_W}) + \tilde{t}_1(\delta_{\lambda_W})\right), \end{cases} \tag{11}$$

где w_0, w_1 – задаваемые величины потерь, вызванных ошибками принятия решений первого и второго рода соответственно; C – коэффициент максимально допустимого увеличения среднего числа наблюдений. Алгоритмическая сложность численного решения задачи (11) составляет $O(K^2)$, где K – число состояний аппроксимирующей цепи Маркова, используемой в рас-

сматриваемом подходе для вычисления характеристик эффективности (7), (8) последовательного решающего правила.

Проиллюстрируем подход результатами компьютерных экспериментов для следующего частного случая:

$$x_t = (x_{t1}, \ldots, x_{t5})',\ x_{ti} \in A = \{1, 2, 3, 4\},\ U = \underbrace{A \times \cdots \times A}_{5},\ |U| = 1024,\ t = 1, 2, \ldots$$

Нулевая гипотеза состоит в предположении дискретного равномерного распределения вероятностей на множестве *U*. Альтернативное предположение заключается в том, что при независимости компонент в первых трех из них имеется преобладание единиц над остальными тремя возможными значениями, четвертая и пятая компоненты распределены дискретно равномерно:

$$\mathrm{P}_0(u) = \tfrac{1}{1024},\ u \in U;$$

$$\mathrm{P}_1\{x_{ti} = 1\} = 0,4,\ \mathrm{P}_1\{x_{ti} = j\} = 0,2,\ j \in \{2, 3, 4\},\ i \in \{1, 2, 3\},$$
$$\mathrm{P}_1\{x_{ti} = j\} = \tfrac{1}{4},\ j \in A,\ i \in \{4, 5\}.$$

Значения порогов C_-, C_+ выбраны в соответствии с выражениями (6), $C = 6$, $w_0 = w_1 = 1$. Последовательное решающее правило, основанное на функции (5), сравнивалось с построенным робастным решающим правилом при $g_- = -0,231,\ g_+ = 0,241$ в рамках искаженной модели, заданной следующим образом:

$$\varepsilon_- = \varepsilon_+ = 0,1,$$

$$\tilde{\mathrm{P}}_0\{x_{ti} = 1\} = 1,\ \tilde{\mathrm{P}}_0\{x_{ti} \neq 1\} = 0,\ i \in \{1, 2, 3\},\ \tilde{\mathrm{P}}_0\{x_{ti} = j\} = \tfrac{1}{4},\ j \in A,\ i \in \{4, 5\};$$

$$\tilde{\mathrm{P}}_1\{x_{ti} = 3\} = 1,\ \tilde{\mathrm{P}}_0\{x_{ti} \neq 3\} = 0,\ i \in \{1, 2, 3\},\ \tilde{\mathrm{P}}_1\{x_{ti} = j\} = \tfrac{1}{4},\ j \in A,\ i \in \{4, 5\}.$$

Результаты экспериментов (таблица) демонстрируют существенный выигрыш в малости вероятностей ошибок первого и второго рода для построенного робастного решающего правила (11) в сравнении с традиционным последовательным решающим правилом, полученный ценой приемлемого увеличения среднего числа необходимых наблюдений.

Результаты вычислительных экспериментов

Results of computer experiments

α_0	β_0	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\alpha}(\delta_g)$	$\tilde{\beta}(\delta_g)$	$\tilde{t}_0$	$\tilde{t}_1$	$\tilde{t}_0(\delta_g)$	$\tilde{t}_1(\delta_g)$
0,01	0,01	0,243	0,124	0,004	0,008	60,1	51,3	224,2	261,6
0,01	0,05	0,227	0,216	0,004	0,041	39,2	40,7	145,0	240,3
0,05	0,05	0,304	0,197	0,023	0,043	28,2	26,6	137,4	156,5
0,05	0,10	0,289	0,291	0,025	0,084	21,7	22,6	106,3	144,1
0,10	0,10	0,352	0,248	0,058	0,085	16,9	16,4	95,5	108,9

Поток данных, образующий временной ряд с трендом. Рассмотрим теперь более сложную модель – поток неоднородных стохастических наблюдений, образующих временной ряд с трендом. Пусть наблюдения $x_1, x_2, \ldots \in \mathbf{R}$ имеют вид

$$x_t = \theta^T \psi(t) + \xi_t,\ t \geq 1, \tag{12}$$

где $\psi(t) = (\psi_1(t), \psi_2(t), \ldots, \psi_m(t))^T$, $t \geq 1$, – вектор заданных базисных функций тренда ($(\cdot)^T$ означает транспонирование); $\theta = (\theta_1, \theta_2, \ldots, \theta_m)^T \in \mathbf{R}^m$ – вектор коэффициентов (его истинное значение в процессе наблюдения неизвестно); $\{\xi_t, t \geq 1\}$ – последовательность независимых одинаково распределенных случайных величин (ошибок, погрешностей наблюдения). Обозначим плотности распределения вероятностей случайных наблюдений (12) через $\{p_n(x, \theta),\ x \in \mathbf{R}, n \geq 1\}$.

Рассмотрим две гипотезы относительно вектора параметров, определяющего тренд наблюдаемого потока данных:

$$H_0 : \theta = \theta^0,\ H_1 : \theta = \theta^1, \tag{13}$$

где $\theta^0, \theta^1 \in \mathbf{R}^m$ – заданные векторы, $\theta^0 \neq \theta^1$, m – число компонент параметра.

Обозначим статистику накопленного логарифмического отношения правдоподобия, вычисленную по n наблюдениям аналогично выражению (4):

$$\Lambda_n = \Lambda_n(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \lambda_i,$$

причем здесь логарифмическое отношение правдоподобия по наблюдению x_i вычисляется следующим образом: $\lambda_i = \ln\left(p_i(x_i, \theta^1) / p_i(x_i, \theta^0) \right)$; $p_i(x, \theta) > 0$ – плотность распределения вероятностей случайного наблюдения номер i при условии, что истинное значение вектора параметров равно θ.

После $n = 1, 2, 3, \ldots$ наблюдений решение принимается согласно правилу

$$d = \mathbf{1}_{[C_+, +\infty)}(\Lambda_n) + 2 \cdot \mathbf{1}_{(C_-, C_+)}(\Lambda_n), \tag{14}$$

где пороги C_-, C_+ являются параметрами и вычисляются в соответствии с формулами (6); $d = 2$ соответствует решению о том, что полученной в n наблюдениях информации недостаточно для обеспечения заданных малых уровней вероятностей ошибок первого и второго рода и требуется получить и обработать следующее наблюдение из потока; $d = 0$ и $d = 1$ означают остановку процесса наблюдения и принятие решения в пользу соответствующей гипотезы из (13).

При выполнении часто встречающегося предположения о распределении вероятностей погрешностей наблюдения $\xi_t \sim N_1(0, \sigma^2)$, $t = 1, 2, 3, \ldots$, где $\sigma > 0$ – заданное среднеквадратическое отклонение, получаем

$$x_t \sim N(\theta^T \psi(t);\ \sigma^2),\ t \geq 1,\ p_t(x, \theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2} (x - \theta^T \psi(t))^2 \right\}.$$

Как следствие, логарифмическое отношение правдоподобия линейно по наблюдениям:

$$\lambda_t = \lambda_t(x_t) = -\frac{1}{2\sigma^2} \{ 2x_t (\theta^0 - \theta^1)^T \psi(t) + (\theta^1)^T \psi(t) \psi^T(t) \theta^1 - (\theta^0)^T \psi(t) \psi^T(t) \theta^0 \}.$$

На практике часто гипотетическая модель потока наблюдений (12) оказывается подверженной «выбросам» [9]:

$$\overline{x}_t = \theta^T \psi(t) + \overline{\xi}_t,\ t \ge 1, \tag{15}$$

где ошибки наблюдения образованы смесью $\overline{\xi}_t = (1-\tau_t)\xi_t + \tau_t\tilde{\xi}_t,\ t \ge 1$, а $\{\tilde{\xi}_t, t \ge 1\}$ – последовательность независимых случайных величин, при этом $\{\tau_t, t \ge 1\}$ – независимые одинаково распределенные случайные величины с распределением Бернулли: $P(\tau_t = 0) = 1-\delta,\ P(\tau_t = 1) = \delta,$ $\tau_t, \xi_t, \tilde{\xi}_t$ независимы, а $\delta \in [0, 1/2)$ – уровень «засорения».

Для получения робастного решающего правила [10] при искажениях (15) построим семейство последовательных решающих правил, основанных на функции $f_{g_-}^{g_+}(\lambda_n)$ вместо λ_n в равенстве (14). Эту функцию по аналогии с потоком однородных наблюдений можно задать в виде

$$f_{g_-}^{g_+}(x) = g_- \cdot \mathbf{1}_{(-\infty, g_-]}(x) + x \cdot \mathbf{1}_{(g_-, g_+)}(x) + g_+ \cdot \mathbf{1}_{[g_+, +\infty)}(x)$$

или, для обеспечения непрерывности функции распределения приращений критериальной статистики, в виде

$$f_{g_-}^{g_+}(x) = \begin{cases} \dfrac{\varepsilon g_-}{x} + g_- - \varepsilon,\ x \le g_-; \\ x,\ g_- < x < g_+; \\ -\dfrac{\varepsilon g_+}{x} + g_+ + \varepsilon,\ x \ge g_+. \end{cases}$$

Оптимальные значения параметров g_-, g_+ определяются по критерию, аналогичному (11).

Поток зависимых наблюдений, образующих однородную цепь Маркова. Пусть теперь модель потока наблюдаемых данных допускает зависимость наблюдений и $x_1, x_2, \ldots$ образуют однородную цепь Маркова [11], принимая значения на множестве $V = \{0, 1, \ldots, M-1\}$. Обозначим вектор вероятностей начальных состояний через $\pi = (\pi_i)$, $i \in V$, а матрицу вероятностей одношаговых переходов – через $P = (p_{ij}),\ i, j \in V$:

$$P\{x_1 = i\} = \pi_i,\ P\{x_n = j \mid x_{n-1} = i\} = p_{ij},\ i, j \in V.$$

Рассмотрим гипотезы, описывающие два типовых режима потока наблюдений в контексте параметров:

$$H_0 : \pi = \pi^{(0)},\ P = P^{(0)};\ H_1 : \pi = \pi^{(1)},\ P = P^{(1)}, \tag{16}$$

где $\pi^{(0)} = (\pi_i^{(0)}),\ \pi^{(1)} = (\pi_i^{(1)})$ – заданные значения вектора ненулевых вероятностей начальных состояний, $P^{(0)} = (p_{ij}^{(0)}) \ne P^{(1)} = (p_{ij}^{(1)})$ – заданные значения матриц ненулевых вероятностей одношаговых переходов для соответствующих гипотез.

При построении последовательного решающего правила проверки гипотез (16) обозначим для модели потока наблюдений статистики

$$\lambda_1 = \log \frac{\pi^{(1)}_{x_1}}{\pi^{(0)}_{x_1}},\ \lambda_k = \log \frac{p^{(1)}_{x_{k-1},x_k}}{p^{(0)}_{x_{k-1},x_k}},\ k>1;\ \Lambda_n = \sum_{k=1}^{n} \lambda_k,\ n \in \mathbf{N}. \tag{17}$$

Последовательное решающее правило проверки гипотез (16) строится в соответствии с выражениями (14) и (17).

Исследуем теперь часто возникающую на практике ситуацию, когда гипотетическая модель, описывающая поток наблюдений, искажена, т. е. фактические значения параметров потока образованы смесью гипотетического и «засоряющего» значений (16) и порождают поток данных, отклоняющийся от постулируемого гипотетической моделью:

$$\overline{\pi}^{(k)} = (1-\varepsilon)\pi^{(k)} + \varepsilon\tilde{\pi}^{(k)},\ \overline{P}^{(k)} = (1-\varepsilon)P^{(k)} + \varepsilon\tilde{P}^{(k)},\ k = 0,1, \tag{18}$$

где $\tilde{\pi}^{(k)}$ и $\tilde{P}(k)$ – вектор вероятностей начальных состояний и матрица вероятностей одношаговых переходов для «засоряющей» цепи Маркова, $P^{(k)} \neq \tilde{P}^{(k)}$, $k = 0,1$, а $\varepsilon \in [0,\ 1/2)$ – вероятность «засорения» (уровень искажения модели).

Построим семейство последовательных решающих правил вида (14), в котором вместо критериальной статистики Λ_n (17) будем использовать равенство

$$\Lambda^g_n(x_1,\ldots,x_n) = g(\lambda_1, M) + \sum_{t=2}^{n} g(\lambda_t, x_{t-1}),\ n \in \mathbf{N}, \tag{19}$$

где $g : \mathbf{R} \times (V \cup \{M\}) \to \mathbf{R}$ – функция, зависящая от $2(M+1)$ дополнительных параметров $g_-(u),\ g_+(u) \in \mathbf{R},\ g_-(u) < g_+(u),\ u \in V \cup \{M\}$:

$$g(y,u) = g_-(u) \cdot \mathbf{1}_{(-\infty,\ g_-(u)]}(y) + y \cdot \mathbf{1}_{(g_-(u),\ g_+(u))}(y) + g_+(u) \cdot \mathbf{1}_{[g_+(u),\ +\infty)}(y),$$

$$y \in \mathbf{R},\ u \in V \cup \{M\}.$$

В рамках семейства (19) робастное последовательное решающее правило при наличии искажений (18) строится по критерию минимакса аналогично (11).

Заключение. В работе рассмотрены три модели потока стохастических наблюдений: независимые однородные наблюдения; неоднородные наблюдения, образующие временной ряд с трендом, и зависимые наблюдения, образующие однородную цепь Маркова. Для них построены последовательные решающие правила для различения двух заданных типовых ситуаций, описываемых в терминах параметров моделей. Отдельно рассмотрен случай, когда гипотетическая модель наблюдений подвержена искажениям, и разработан подход для построения робастных (устойчивых к искажениям) последовательных решающих правил.

Отметим, что предложенный подход позволяет вычислять характеристики эффективности и применим для построения робастных последовательных решающих правил в ситуации, когда типовая ситуация не описывается единственно возможным значением вектора параметров, а соответствует некоторому множеству значений [12]. Кроме того, для других моделей искажений (ошибки спецификации гипотетических значений параметров, уклонения фактических распределений вероятностей в рамках малых окрестностей в соответствующих метрических пространствах, искажения моделей зависимости и т. д.) разработанный в [4] подход для вычисления характеристик эффективности последовательных решающих правил также

применим и позволяет строить робастные последовательные решающие правила. Для модели временных рядов с трендом также удается развить подход вычисления характеристик эффективности последовательных решающих правил на случай, когда каждый из типовых вариантов описывается единственным значением параметра, однако типовых вариантов больше чем два [13].

Представленные в работе последовательные решающие правила применимы также для решения задачи длительного мониторинга потоков данных с целью выявления смены режима функционирования наблюдаемой системы. В этом случае последовательное решающее правило применяется либо с перезапуском после принятия одного из двух решений, либо со сдвигом начала отсчета. При таких схемах принятия решений для обеспечения гарантий всей многоэтапной процедуры требуется дополнительное исследование вероятностей ошибочных решений.

Список использованных источников

1. Mukhopadhyay, N. Sequential Methods and Their Applications / N. Mukhopadhyay, B. de Silva. – N. Y. : Marcel Dekker, 2009. – 504 p.

2. Wald, A. Sequential Analysis / A. Wald. – N. Y. : John Wiley and Sons, 1947. – 212 p.

3. Lai, T. Sequential analysis: Some classical problems and new challenges / T. Lai // Statistica Sinica. – 2001. – Vol. 11. – P. 303–408.

4. Харин, А. Ю. Робастность байесовских и последовательных статистических решающих правил / А. Ю. Харин. – Минск : БГУ, 2013. – 207 с.

5. Huber, P. Robust Statistics / P. Huber, E. Ronchetti. – N. Y. : Wiley, 2009, 380 p.

6. Maevskii, V. V. Robust regressive forecasting under functional distortions in a model / V. V. Maevskii, Y. S. Kharin // Automation and Remote Control. – 2002. – Vol. 63, iss. 11. – P. 1803–1820.

7. Weiss, C. H. Discrete-valued Time Series / C. H. Weiss. – Oxford : John Wiley and Sons, 2018. – 284 p.

8. Kharin, A. Y. Robust sequential test for hypotheses about discrete distributions in the presence of "outliers" / A. Y. Kharin, D. V. Kishylau // J. of Mathematical Sciences. – 2015. – Vol. 205, iss. 1. – P. 68–73.

9. Kharin, A. Performance and robustness analysis of sequential hypotheses testing for time series with trend / A. Kharin, T. T. Tu // Austrian J. of Statistics. – 2017. – Vol. 46, no. 3–4. – P. 23–36.

10. Kharin, A. Y. On error probabilities calculation for the truncated sequential probability ratio test / A. Y. Kharin, T. T. Tu // Журнал Бел. гос. ун-та. Математика. Информатика. – 2018. – № 1. – P. 68–76.

11. Kemeny, J. G. Finite Markov Chains / J. G. Kemeny, J. L. Snell. – N. Y. : Springer, 1960. – 238 p.

12. Kharin, A. Y. An approach to asymptotic robustness analysis of sequential tests for composite parametric hypotheses / A. Y. Kharin // J. of Mathematical Sciences. – 2017. – Vol. 227, iss. 2. – P. 196–203.

13. Tu, T. T. Sequential probability ratio test for many simple hypotheses on parameters of time series with trend / T. T. Tu, A. Y. Kharin // Журнал Бел. гос. ун-та. Математика. Информатика. – 2019. – № 1. – P. 35–45.

14. Siegmund, D. Sequential Analysis. Tests and Confidence Intervals / D. Siegmund. – N. Y. : Springer-Verlag, 1985. – 272 p.

15. Ghosh, B. Sequential Tests of Statistical Hypotheses / B. Ghosh. – Reading : Addison – Wesley, 1970. – 454 p.

References

1. Mukhopadhyay N., Silva B. de. *Sequential Methods and Their Applications*. New York, Marcel Dekker, 2009, 504 p.

2. Wald A. *Sequential Analysis*. New York, John Wiley and Sons, 1947, 212 p.

3. Lai T. Sequential analysis: Some classical problems and new challenges. *Statistica Sinica*, 2001, vol. 11, pp. 303–408.

4. Kharin A. Y. Robastnost bajesovskih i posledovatelnyh statisticheskih reshayuschih pravil. *Robustness of Bayesian and Sequential Statistical Decision Rules*. Minsk, Belorusskij gosudarstvennyj universitet, 2013, 207 p. (In Russ.).

5. Huber P., Ronchetti E. *Robust Statistics*. New York, Wiley, 2009, 380 p.

6. Maevskii V. V., Kharin Y. S. Robust regressive forecasting under functional distortions in a model. *Automation and Remote Control*, 2002, vol. 63, iss. 11, pp. 1803–1820.

7. Weiss C. H. *Discrete-valued Time Series*. Oxford, John Wiley and Sons, 2018, 284 p.

8. Kharin A. Y., Kishylau D. V. Robust sequential test for hypotheses about discrete distributions in the presence of "outliers". *Journal of Mathematical Sciences*, 2015, vol. 205, iss. 1, pp. 68–73.

9. Kharin A., Tu T. T. Performance and robustness analysis of sequential hypotheses testing for time series with trend. *Austrian Journal of Statistics*, 2017, vol. 46, no. 3–4, pp. 23–36.

10. Kharin A. Y., Tu T. T. *On error probabilities calculation for the truncated sequential probability ratio test*. Zhurnal Belorusskogo gosudarstvennogo universiteta. Matematika. Informatika [*Journal of the Belarusian State University. Mathematics and Informatics*], 2018, no. 1, pp. 68–76.

11. Kemeny J. G., Snell J. L. *Finite Markov Chains*. New York, Springer, 1960, 238 p.

12. Kharin A. Y. An approach to asymptotic robustness analysis of sequential tests for composite parametric hypotheses. *Journal of Mathematical Sciences*, 2017, vol. 227, iss. 2, pp. 196–203.

13. Tu T. T., Kharin A. Y. *Sequential probability ratio test for many simple hypotheses on parameters of time series with trend*. Zhurnal Belorusskogo gosudarstvennogo universiteta. Matematika. Informatika [*Journal of the Belarusian State University. Mathematics and Informatics*], 2019, no. 1, pp. 35–45.

14. Siegmund D. *Sequential Analysis. Tests and Confidence Intervals*. New York, Springer-Verlag, 1985, 272 p.

15. Ghosh B. *Sequential Tests of Statistical Hypotheses*. Reading, Addison – Wesley, 1970, 454 p.

Информация об авторе

Харин Алексей Юрьевич, доктор физико-математических наук, доцент, заведующий кафедрой теории вероятностей и математической статистики Белорусского государственного университета, главный научный сотрудник НИЛ статистического анализа и моделирования Учреждения БГУ «НИИ прикладных проблем математики и информатики».
ORCHID ID: 0000-0002-5790-1956
E-mail: KharinAY@bsu.by

Information about the author

Alexey Y. Kharin, Dr. Sci. (Phys.-Math.), Associate Professor, Head of the Department of Probability Theory and Mathematical Statistics, Leading Researcher of the Research Laboratory of Statistical Analysis and Modeling of the Research Institute for Applied Problems of Mathematics and Informatics, Belarusian State University.
ORCHID ID: 0000-0002-5790-1956
E-mail: KharinAY@bsu.by