Математические модели оценок качества обнаружения моментов изменения свойств случайных процессов в алгоритме невязок

Е. Н. Бендерская
Санкт-Петербургский политехнический университет Петра Великого
helen.bend@gmail.com

***Аннотация*. Разработаны математические модели для расчета среднего числа шагов запаздывания в обнаружении изменения свойств случайных процессов в случае принятия решения на основе решающей статистики, построенной по принципу «невязок». Предложенные модели позволяют выбрать параметры алгоритма невязок с ориентацией на наихудший случай, что в системах реального времени важно для обеспечения обнаружения момента изменения свойств случайного процесса с запаздыванием, не превышающем заданного значения.**

Ключевые слова: моменты изменения свойств случайных процессов; задача о разладке; показатели качества; решающая статистика; невязка

I. Введение

Задача обнаружения моментов изменения свойств случайных процессов возникает во многих практических приложениях – начиная с таких традиционных, как медицинская и техническая диагностика [1, 2], и заканчивая относительно новыми – мониторинг активности в компьютерных сетях и анализ поведения программного обеспечения [3, 4]. Несмотря на то, что в настоящее время активно развиваются новые методы обработки как первичных сигналов, так и информации, включая природоподобные вычисления [5, 6], тем не менее интерес к классическим подходам на основе статистических алгоритмов и их актуальность сохраняются [7, 8]. При этом основное внимание уделяется вопросам настройки (определению параметров алгоритмов) и модификации алгоритмов обнаружения с учетом того, что априорная информация о параметрах процессов часто не доступна, а режимы накопления первичной статистики ограничены [9–11].

Данная работа посвящена определению показателей качества работы алгоритма «невязок» для последующей его настройки на основе разработанных математических моделей.

В соответствии с [12], решающую функцию алгоритма «невязок» для обнаружения «разладки» (изменение параметров модели случайного процесса (СП)) можно представить следующим образом:

$$G_k = \frac{\sum_{t=1}^{k}(g_t - 1)}{\sqrt{2k}}, k = \overline{1, N_C}\ , \quad (1)$$

$$g_t = \left\{ \frac{e_t - f(E_t^n, \theta^{(1)})}{\beta^{(1)}} \right\}^2 , \quad (2)$$

где $E_t^n = (e_{t-1}, e_{t-2}, ..., e_{t-n})$ - вектор n предыдущих отсчетов СП e_i, соответствующий моменту времени t; $(\theta^{(i)}\beta^{(i)}) = (\theta^{(i)}{}_1, ..., \theta^{(i)}{}_m, \beta^{(i)})$ – $(1+m)$ – мерные векторы параметров (уравнений) СП до (i=1) и после (i=2) «разладки»; $f(E_t^n, \theta^{(1)})$ – известная функция; N_c – период накопления решающей функции (после выполнения операции сравнения решающая функция обнуляется $G_{N_C} = 0$, так как с течением времени скорость роста решающей функции уменьшается).

Решающие правила можно представить следующим образом [12]. Пусть гипотеза H^0 соответствует случаю наличия «разладки», а гипотезы H^1 – ее отсутствию, тогда:

$$\begin{aligned} H^0 &: |G_t| \le h \\ H^1 &: |G_t| > h \end{aligned} \quad (3)$$

Рассмотрим случай, когда СП представлен моделью авторегрессии первого порядка АР(1) (в этом случае *n=1*) и тогда:

$$f(E_t^n, \theta^{(i)}) = (e_{t-1} - \mu^{(i)}), \quad (4)$$

$$(\theta^{(i)}\beta^{(i)}) = (\mu^{(i)}, a^{(i)}, \beta^{(i)}), \quad (5)$$

$$e_t = \mu^{(i)} + a^{(i)}(e_{t-1} - \mu^{(i)}) + \beta^{(i)} z_t\ , \quad (6)$$

где *i=1* – значения параметров АР до «разладки», *i=2* – после «разладки»; z_t – дискретный белый шум с μ=*0* и σ=*1*; $\beta^{2(i)} = \sigma^{2(i)}(1 - a^{2(i)})$ [12]. В дальнейшем считается, что $a^{(1)} = a^{(2)} = a$, а решающая функция G_k используется для обнаружения «разладок» вида «изменение математического ожидания» и вида «изменение дисперсии», так как в большинстве случаев необходимо обнаруживать именно эти виды «разладок».

Вероятность «ложного» обнаружения для этого алгоритма может быть определена с помощью представления работы алгоритма в виде конечного марковского графа [13].

II. Математические Модели для Среднего Времени Запаздывания в Обнаружении

В [12] показано, что при отсутствии «разладки» $M\,[G_{k+1}]=0$. Пусть «разладка» возникает в момент $t=n_0$, тогда, учитывая (1)–(6), произведем оценку среднего времени, за которое математическое ожидание решающей функции G_k достигнет уровня h, это и будет средним временем запаздывания в обнаружении «разладки»:

$$M[G_{k+1}]=\frac{\sum_{n_0}^{k}\frac{(\mu^{(2)}-\mu^{(1)})^2(1-a)^2+\beta^{(2)2}-1}{\beta^{(1)2}}}{\sqrt{2k}}=$$

$$=(k-n_0+1)\frac{\frac{(\mu^{(2)}-\mu^{(1)})^2(1-a)^2+\beta^{(2)2}}{\beta^{(1)2}}-1}{\sqrt{2k}}=$$

$$=(k-n_0+1)\frac{\frac{(\mu^{(2)}-\mu^{(1)})^2(1-a)^2}{\sigma^{(1)2}(1-a^2)}+\frac{\sigma^{(2)2}-\sigma^{(1)2}}{\sigma^{(1)2}}}{\sqrt{2k}},$$

откуда следует

$$M[G_k]=\left\{u^2+2u+r^2\frac{1-a}{1+a}\right\}\frac{k-n_0+1}{\sqrt{2k}}, \qquad (7)$$

где $u=\frac{\sigma_2-\sigma_1}{\sigma_1}$, $r=\mu_2-\mu_1$, а k определяется исходя из

$$M[G_k]=h, \qquad (8)$$
$$N_{обн.р}=k-n_0+1.$$

Из (7) видно, что время запаздывания в обнаружении для алгоритма «невязок» зависит от момента возникновения «разладки» n_0. При этом возможны следующие случаи:

1) обнаружение произойдет в том же периоде накопления, в котором произошла «разладка», т.е.

$$\exists k:\ M[G_k]=h,\ k=\overline{n_0,N_C};$$

2) обнаружение произойдет в момент времени j на следующем периоде накопления, т.е.,

$$\forall k, M[G_k]<h,\ k=\overline{n_0,N_C}$$
$$\exists\, j:\frac{mj}{\sqrt{2j}}=h,\ j=\overline{1,N_C},$$
$$m=u^2+2u+r^2\frac{1-a}{1+a},\ u=\frac{\sigma_2-\sigma_1}{\sigma_1},\ r=\mu_2-\mu_1, \qquad (9)$$

а момент времени j отсчитывается от начала второго периода накопления.

3) «разладка» не обнаружиться никогда (случай необнаруживаемой «разладки»), если за весь период N_c решающая функция не успевает достигать порога h

$$\forall k, M[G_k]<h,\ k=\overline{n_0,N_C}$$
$$\forall\, j:\frac{mj}{\sqrt{2j}}<h,\ j=\overline{1,N_C},$$

где m, u, r – вычисляются по (9).

«Разладка» будет обнаружена на том же периоде, на котором она возникла, для тех значений n_0 и при таком сочетании значений h, N_c, r, u, для которых выполняется следующее условие:

$$(N_C-n_0+1)m\ge h\sqrt{2N_C} \qquad (10)$$

Это условие является более сильным и позволяет ориентироваться на наихудший случай. Исходя из этого, введем следующее обозначение

$$j^*=\left]2\mathrm{N_c}+1\text{-h}\frac{\sqrt{2\mathrm{N_C}}}{\mathrm{m}}\right[, \qquad (11)$$

где $]X[$ – ближайшее целое, не больше X. Тогда, учитывая (10), можно считать, что обнаружение будет происходить на первом периоде накопления (на том же, на котором возникла «разладка»), если $1\le n_0\le j^*$, а при $j^*+1\le n_0\le N_C$ – на втором периоде накопления. При этом $n_0=(j^*+1)$ – такой момент возникновения «разладки», при котором решающая функция не успевает достичь порога (будет обнулена) за время $t=N_C-n_0$, и поэтому обнаружение произойдет на втором периоде. С этой точки зрения – это самый неблагоприятный момент возникновения «разладки», так как в этом случае среднее время запаздывания в обнаружении будет максимальным при прочих равных условиях (увеличиться на t) и чем ближе к концу периода накопления возникает «разладка» в случае 2, тем меньше время t ожидания окончания текущего периода накопления). В случае 1 (при любом моменте возникновения «разладки» – обнаружение на том же периоде), чем позже возникает «разладка», тем больше время запаздывания в обнаружении из-за уменьшения скорости роста решающей функции с течением времени. Общее условие обнаружения «разладки» выглядит следующим образом:

$$N_C m\ge h\sqrt{2N_C}$$

или $j^*\ge 1$. Среднее время запаздывания в обнаружении определяется из соотношений в зависимости от рассматриваемого случая:

1) $\overline{N}_{обн.р}=j_{обн}-n_0+1$, а $j_{обн}$ находится из уравнения:

$$mj_{обн}-h\sqrt{2j_{обн}}+mn_0=0,$$

откуда следует

$$j_{обн} = \frac{(h\sqrt{2} + \sqrt{2h^2 + 4m^2(N_C - 1)})^2}{4m^2} \qquad (12)$$

2) $\overline{N}_{обн.р} = N_C - n_0 + j_{обн} + 1,$

а $j_{обн}$ находится из уравнения

$$mj_{обн} - h\sqrt{2j_{обн}} = 0,$$

откуда следует

$$j_{обн} = \frac{2h^2}{m^2}. \qquad (13)$$

В виду того, что реально момент возникновения «разладки» неизвестен, целесообразно в качестве оценки среднего числа шагов запаздывания в обнаружении принять его оценку по наихудшему случаю. Для определения наличия моментов возникновения «разладки», при которых могут быть обнаружения на втором периоде накопления решающей функции, предлагается вычислять значение параметра *j**. по соотношению (11). Кроме того, значение *j** будет показывать, является ли «разладка» заданной величины, вообще обнаруживаемой при выбранных параметрах алгоритма. Если указанные моменты существуют, то $\overline{N}_{обн.р}$ вычисляется как для случая 2 при $n_0 = (j^* + 1)$, а если таких моментов нет (т.е. при любом моменте возникновения «разладки» она будет обнаружена на том же периоде накопления) – как в первом случае для $n_0 = N_C$.

Общее аналитическое соотношение выглядит следующим образом:

$$\begin{cases} j^* < 1 \text{ - необнаруживаемая "разладка"} \\ \overline{\mathrm{N}_{\text{обн.р}}} = \dfrac{(h\sqrt{2} + \sqrt{2h^2 + 4m^2(N_C - 1)})^2}{4m^2} + 1 - \mathrm{N_C}, \text{ если } \mathrm{j^* = N_C} \\ \overline{\mathrm{N}_{\text{обн.р}}} = N_C - j^* + \dfrac{2h^2}{m^2}, \text{ если } \mathrm{j^* + 1 \le N_C}. \end{cases} \qquad (14)$$

Случай многократных «разладок» применительно алгоритму «невязок» не рассматривается, так как обнаружение восстановления может быть только с существенным запаздыванием. С учетом наихудшего случая за оценку среднего времени запаздывания в обнаружении восстановления свойств СП можно взять период накопления решающей функции.

III. Заключение

Проверка достоверности предложенных математических моделей по оценке запаздывания в обнаружении «разладки» осуществлялась путем имитационного моделирования на разработанной имитационной модели. Для этого исследована зависимость среднего числа шагов запаздывания в обнаружении «разладки» от момента ее возникновения для различных периодов накопления решающей функции при различных уровнях «ложного» обнаружения и произведено сравнение с зависимостями, полученными по аналитической модели.

Анализ проведенных экспериментов показал, что аналитические модели полностью отражают характер исследуемой зависимости. Кроме того, разработанная модель (14) обеспечивает оценку по наихудшему случаю для «разладки» вида «изменения математического ожидания». Однако, как показывает анализ результатов, если при любом моменте возникновения «разладки» она обнаруживается на первом периоде накопления решающей функции (см. соотношение (12)), то наихудший момент возникновения «разладки» приходится на середину периода накопления решающей функции. Поэтому оценка среднего числа шагов запаздывания в обнаружении может быть смягчена и вычислена по (14) для $n_0 = \frac{N_C}{2}$.

Несмотря на одинаковый характер поведения аналитических и имитационных зависимостей, гарантированной оценки среднего числа шагов запаздывания в обнаружении «разладки» вида «изменения дисперсии» модель (14) не обеспечивает, и, следовательно, в математической модели необходимо учесть дисперсию решающей функции.

Дисперсию решающей функции (1) можно определить из следующего выражения, полученного по аналогии с выражением для математического ожидания (7):

$$D[G_k] = \frac{(n_0 - 1) + (k - n_0 + 1)\{2(u+1)^2 r^2 \frac{1-a}{1+a} + (1+u)^4\}}{k} \qquad (15)$$

Для того, чтобы учесть дисперсию решающей функции при определении среднего числа шагов запаздывания в обнаружении «разладки» вида «изменение дисперсии» и в то же время сильно не загрубить искомую оценку, предлагается рассматривать вместо (8) момент пресечения порога *h* нижней границей интервала

$$\left[\mathrm{M[G_k]} - \alpha_1 \sigma_{\mathrm{G}};\ \mathrm{M[G_k]} \alpha_1 + \sigma_{\mathrm{G}}\right],$$

где α_1 – коэффициент, определяющий границы, в пределах которых лежит большинство реализаций G_k; σ_G – среднеквадратическое отклонение решающей функции G_k.

Тогда, принимая во внимание рассуждения, приведенные выше при выводе математической модели по оценке $\overline{N}_{обн.р}$, при обнаружении «разладки» вида «изменение дисперсии», получим следующее соотношение для среднего числа шагов запаздывания:

$$\begin{cases} j^* < 1 \text{ - необнаруживаемая "разладка"} \\ \overline{\mathrm{N}_{\text{обн.р}}} = \dfrac{((h+\alpha_1\sigma_G)\sqrt{2}+\sqrt{2(h+\alpha_1\sigma_G)^2+4m^2(N_C-1))^2}}{4m^2}+1-\mathrm{N_C}, \text{ если } \mathrm{j^*=N_C} \\ \overline{\mathrm{N}_{\text{обн.р}}} = N_C - j^* + \dfrac{2(h+\alpha_1\sigma_G)^2}{m^2}, \text{ если } \mathrm{j^*+1 \le N_C} \end{cases}$$

В результате имитационных экспериментов по определению α_1, проведенных в широком диапазоне условий и при различных возможных значениях «разладки» вида «изменение дисперсии», было установлено, что значение коэффициента α_1 равное 0.1 обеспечивает гарантированные расчетные оценки среднего числа шагов запаздывания в обнаружении «разладки».

Список литературы

[1] Basseville M., Nikiforov I.V. Detection of abrupt changes: theory and application. N.J.: Prentice Hall Englewood Cliffs. 1993. 528 p.

[2] A. Kalmuk, O. Granichin, O. Granichina and M. Ding. Detection of abrupt changes in autonomous system fault analysis using spatial adaptive estimation of nonparametric regression // American Control Conference (ACC). Boston. MA. 2016. pp. 6839-6844

[3] Tartakovsky A.G., Polunchonko A.S., Sokolov G. Efficient computer network anomaly detection by changepoint detection methods // IEEE Journal of Selected Topics in Signal Processing. 2013. Vol. 7. №1. pp. 4-11

[4] D'Ettorre S., Viktor H.L., Paquet E. Context-Based Abrupt Change Detection and Adaptation for Categorical Data Streams // In: Yamamoto A., Kida T., Uno T., Kuboyama T. (eds) Discovery Science. Lecture Notes in Computer Science. Springer. 2017. Vol 10558. pp. 3–17.

[5] Агеев Е.В., Бендерская Е.Н. Обзор природных вычислений: основные направления и тенденции // Научно-технические ведомости. Информатика. Телекоммуникации. Управление. 2014. №2 (193). С. 9-22.

[6] Агеев Е.В. Бендерская Е.Н. Природоподобные вычисления: от природных явлений к практическим задачам // XX Международная конференция по мягким вычислениям и измерениям (SCM-2017). Сборник докладов в 3-х томах. Санкт-Петербург. 24-26 мая 2017 г. Т.2. с. 531-533.

[7] Artemov A., Burnaev E. Ensembles of detectors for online detection of transient changes // Eighth International Conference on Machine Vision. International Society for Optics and Photonics, 2015. pp. 98751Z–98751Z-5

[8] Rodionov S.N. A comparison of two methods for detecting abrupt changes in the variance of climatic time series // Adv. Stat. Clim. Meteorol. Oceanogr. 2016. №2. pp. 63-78.

[9] Du W., Polunchenko A. S., Sokolov G. On Robustness of the Shiryaev-Roberts Procedure for Quickest Change-Point Detection under Parameter Misspecification in the Post-Change Distribution // arXiv preprint arXiv:1504.04722. 2015. arXiv: arXiv:1504.04722v1

[10] Alippi C., Boracchi G., Roveri M. Hierarchical Change-Detection Tests // IEEE Transactions on Neural Networks and Learning Systems. 2017. Vol. 28. Issue 2. pp. 246-258.

[11] Jin-Peng Qi, Jie Qi, Qing Zhang. A Fast Framework for Abrupt Change Detection Based on Binary Search Trees and Kolmogorov Statistic // Computational Intelligence and Neuroscience. 2016. Article ID 8343187. 16 p.

[12] Бородкин Л.И., Моттль В.В. Алгоритм обнаружения моментов изменения параметров уравнения случайного процесса // Автоматика и телемеханика. 1976. №6. с.23-32.

[13] Бендерская Е.Н., Колесников Д.Н., Пахомова В.И. Функциональная диагностика систем управления. Учебное пособие. СПб.: СПбГТУ. 2000. 144 с.