УДК 519.21

Г.Б. Філімоніхін, проф., д-р техн. наук
Кіровоградський національний технічний університет,
М.В. Гончарова, студ.
Національний технічний університет України «КПІ», м.Київ

Дослідження задачі про розладнання послідовності з n символів

Досліджується задача про розладнання послідовності з n символів та її застосування в схемах захисту інформації. Побудовані два алгоритми для визначення моменту розладнання та зроблений порівняльний аналіз їх ефективності
момент розладнання, послідовність з n символів

Вступ. Класична задача розладнання [1] полягає у визначенні випадкового моменту зміни ймовірнісних характеристик випадкового процесу, для якого потрібно побудувати оцінки. Залишається актуальною задача побудови алгоритмів визначення моментів розладнання різних типів випадкових послідовностей, що використовуються в сучасних системах захисту інформації.

1. **Постановка задачі та обґрунтування алгоритмів її розв'язання.** Дослідимо задачу розладнання для потоку з n символів алфавіту. Якщо занумерувати символи від 1 до n, то спостережними подіями в потоці n символів будуть події $A_1, A_2, \ldots, A_n$, що вказують на появу символу з відповідним номером. Джерело повідомлень передає кожен символ алфавіту з певною ймовірністю $p_1, p_2, \ldots, p_n$; у деякий момент часу ймовірності передачі символів змінюються відповідно на $q_1, q_2, \ldots, q_n$. Потрібно визначити момент зміни ймовірностей, тобто момент розладнання.

Події $A_1, A_2, \ldots, A_n$ утворюють повну групу подій, тому

$$\sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} q_i = 1 .$$

1.1. Алгоритм 1. Розглядаємо послідовність символів довжиною *М*. Для визначення моменту розладнання послідовності достатньо знайти момент розладнання хоча б для одного її символу, так як розладнання для всіх символів відбувається одночасно.

Кількість K_i символів під номером *i* на інтервалах довжиною N_i до моменту розладнання повинна зберігати майже сталу величину за властивістю стійкості відносної частоти появи випадкової події в серії експериментів, тобто

$$K_i = \nu_{N_i}(A_i) N_i \approx p_i N_i ,$$

де N_i – кількість експериментів;

$\nu_{N_i}(A_i)$ – частота появи *i*-того символу.

Частота $\nu_{N_i}(A_i)$ змінюватиметься при зміні ймовірності, тобто $\nu_{N_i}(A_i) \approx q_i$, а отже зміниться і кількість K_i *i*-тих символів на інтервалах довжиною N_i:

$$K_i \approx q_i N_i .$$

Для визначення моменту розладнання підраховуємо кількість K_i знаків відповідного *i*-

того символу, $i = \overline{1,n}$, на інтервалах від 0 до N_i, $N_i < M$, потім – на інтервалах від 1 до

N_i+1, від 2 до N_i+2, …, від $M-N_i$ до M.

Проте при застосуванні такого алгоритму виникає проблема з визначенням довжини спостережного інтервалу, так як за умовою задачі ймовірності символів до і після моменту розладнання $p_1, p_2, \ldots, p_n$ і $q_1, q_2, \ldots, q_n$ приймають малі значення.

Критичну область для статистики K_i обираємо у вигляді $U_{\text{кр}} = \{K_i > C\}$.

При заданій помилці першого роду α для визначення порогу C використаємо інтегральну теорему Лапласа:

$$P\left(0 < \frac{K_i - Nq_i}{\sqrt{Nq_i(1-q_i)}} < \frac{C - Nq_i}{\sqrt{Nq_i(1-q_i)}}\right) = \Phi\left(\frac{C - Nq_i}{\sqrt{Nq_i(1-q_i)}}\right) = 1-\alpha, \quad (1)$$

де q_i – імовірність надходження символів під номером i після розладнання;

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_0^x e^{-z^2/2}dz .$$

Із співвідношення (1) отримаємо:

$$C = Nq_i + t_{1-\alpha}\sqrt{Nq_i(1-q_i)}, \quad (2)$$

де $t_{1-\alpha}$ – квантиль нормального розподілу;

$\Phi(t_{1-\alpha}) = 1-\alpha$.

Аналогічно за теоремою Лапласа при заданій помилці другого роду β

$$P\left(0 < \frac{K_i - Np_i}{\sqrt{Np_i(1-p_i)}} < \frac{C - Np_i}{\sqrt{Np_i(1-p_i)}}\right) = \Phi\left(\frac{C - Np_i}{\sqrt{Np_i(1-p_i)}}\right) = \beta .$$

Відповідний даному N і визначеному в (2) C, квантиль $t_{1-\beta}$ ймовірності помилки другого роду β рівний

$$t_{1-\beta} = -t_\beta = \frac{Np_i - C}{\sqrt{Np_i(1-p_i)}} . \quad (3)$$

Враховуючи (2), (3), довжина найменшого інтервалу, на якому підраховуємо кількість символів під номером *i*, рівна

$$N = \left(\frac{t_{1-\beta}\sqrt{p_i(1-p_i)} + t_{1-\alpha}\sqrt{q_i(1-q_i)}}{p_i - q_i}\right)^2 . \quad (4)$$

де $i = \overline{1,n}$.

Щоб точніше визначити момент розладнання, вибираємо символ із найменшою довжиною спостережного інтервалу. Моментом розладнання всього потоку символів вважатимемо момент розладнання знайденого символу.

1.2. Алгоритм 2. Для розв’язання задачі про розладнання для потоку з n символів алфавіту можна також використати алгоритм, який базується на властивості стійкості вибіркових середніх.

На основі результатів експериментів встановлюємо момент розладнання потоку символів, який співпадає з моментом зміни значення вибіркового середнього ймовірностей символів потоку, при цьому довжина спостережного інтервалу для потоку повідомлень рівна N, $N < M$.

Підраховуємо кількість K_i символів під номером *i*, $i = \overline{1,n}$, на інтервалах від 0 до N, потім — на інтервалах від 1 до $N+1$, від 2 до $N+2$, ..., від $N-M$ до M. На цих інтервалах обчислюємо вибіркові середні символів потоку. До моменту розладнання вони будуть рівні

$$n_{\text{в}} = \frac{1}{N}\sum_{i=1}^{n} iK_i \approx \sum_{i=1}^{n} ip_i ,$$

а після –

$$n_{\text{в}} = \frac{1}{N}\sum_{i=1}^{n} iK_i \approx \sum_{i=1}^{n} iq_i \,.$$

Можливі випадки.

Перший випадок. Якщо за умовою задачі середні символів потоку до і після моменту розладнання майже рівні, тобто

$$\sum_{i=1}^{n} iq_i \approx \sum_{i=1}^{n} ip_i \,,$$

то різниця між вибірковими середніми незначна і пояснюється випадковим відбором об’єктів вибірки. Тому в цьому випадку визначити момент розладнання за зміною величини вибіркових середніх неможливо, і доцільно використати алгоритм 1.

Другий випадок. Занумеруємо символи так, щоб послідовність різниць $p_i - q_i, \, i = \overline{1,n}$ була неспадною. При цьому величина

$$\sum_{i=1}^{n} ip_i - \sum_{i=1}^{n} iq_i$$

буде більш значною.

Потрібно перевірити нульову гіпотезу H_0:

$$n_{\text{в}} = \sum_{i=1}^{n} iq_i \,,$$

якщо розладнання відбулося, при альтернативній гіпотезі H_1

$$n_{\text{в}} = \sum_{i=1}^{n} ip_i \,,$$

якщо розладнання не відбулося.

Якщо незалежні вибірки мають великий об’єм (не менше 30 кожна), то вибіркові середні мають розподіл, близький до нормального.

У якості критерію перевірки нульової гіпотези приймемо випадкову величину

$$U = \frac{n_{\text{в}} - \sum_{i=1}^{n} iq_i}{\sigma(n_{\text{в}})} = \frac{n_{\text{в}} - \sum_{i=1}^{n} iq_i}{\frac{\sigma_q}{\sqrt{N}}} \,,$$

$$\text{де } \sigma_q^2 = \sum_{i=1}^{n} i^2 q_i - \left(\sum_{i=1}^{n} iq_i \right)^2 ,$$

яка нормально розподілена, до того ж при справедливості нульової гіпотези $MU = 0$, $\sigma(U) = 1$.

Критичну область вибираємо у вигляді $U_{\text{кр}} = \{U > C\}$.

При заданій помилці першого роду α для визначення порогу C використаємо інтегральну теорему Лапласа:

$$P\left(0 < \frac{n_{\text{в}} - \sum_{i=1}^{n} iq_i}{\frac{\sigma_q}{\sqrt{N}}} < \frac{C - \sum_{i=1}^{n} iq_i}{\frac{\sigma_q}{\sqrt{N}}} \right) = \Phi\left(\frac{C - \sum_{i=1}^{n} iq_i}{\frac{\sigma_q}{\sqrt{N}}} \right) = 1 - \alpha \,. \tag{5}$$

Із (5) отримаємо:

$$C = \sum_{i=1}^{n} iq_i + \frac{t_{1-\alpha}\sigma_q}{\sqrt{N}} \,. \tag{6}$$

Аналогічно при заданій помилці другого роду β за теоремою Лапласа

$$P\left(0<\frac{n_{\text{в}}-\sum_{i=1}^{n} ip_i}{\frac{\sigma_p}{\sqrt{N}}}<\frac{C-\sum_{i=1}^{n} ip_i}{\frac{\sigma_p}{\sqrt{N}}}\right)=\Phi\left(\frac{C-\sum_{i=1}^{n} ip_i}{\frac{\sigma_p}{\sqrt{N}}}\right)=\beta, \qquad (7)$$

де $\sigma_p^2=\sum_{i=1}^{n} i^2 p_i-\left(\sum_{i=1}^{n} ip_i\right)^2$.

Із рівності (7) знаходимо квантиль $t_{1-\beta}$ імовірності помилки другого роду β:

$$t_{1-\beta}=t_{-\beta}=\frac{\sum_{i=1}^{n} ip_i-C}{\sigma_p}\sqrt{N}. \qquad (8)$$

Із формул (6) і (8) знаходимо довжину найменшого спостережного інтервалу:

$$N=\left(\frac{t_{1-\beta}\sigma_p+t_{1-\alpha}\sigma_q}{\sum_{i=1}^{n} ip_i-\sum_{i=1}^{n} iq_i}\right)^2. \qquad (9)$$

2. Розв'язання задачі. Приклади. Для розв'язання задачі застосовуємо програму, написану на мові С++.

Приклад 1. Нехай $n=10$; $p_i=0{,}1; q_1=0{,}14; q_2=0{,}16; q_3=0{,}06; q_4=0{,}101; q_5=0{,}051$; $q_6=0{,}105; q_7=0{,}1; q_8=0{,}08; q_9=0{,}15; q_{10}=0{,}053$; $\alpha=0{,}01; \beta=0{,}001$.

За алгоритмом 1 за формулою (4) обчислюємо довжину найменшого інтервалу N для кожного символу. Найменше значення буде $N=863$ для символу $i=5$.

За формулою (2) знаходимо $C=60$.

Отже, для значень $K_i<60$ можна вважати з імовірністю висновку 0,99, що розладнання відбулося.

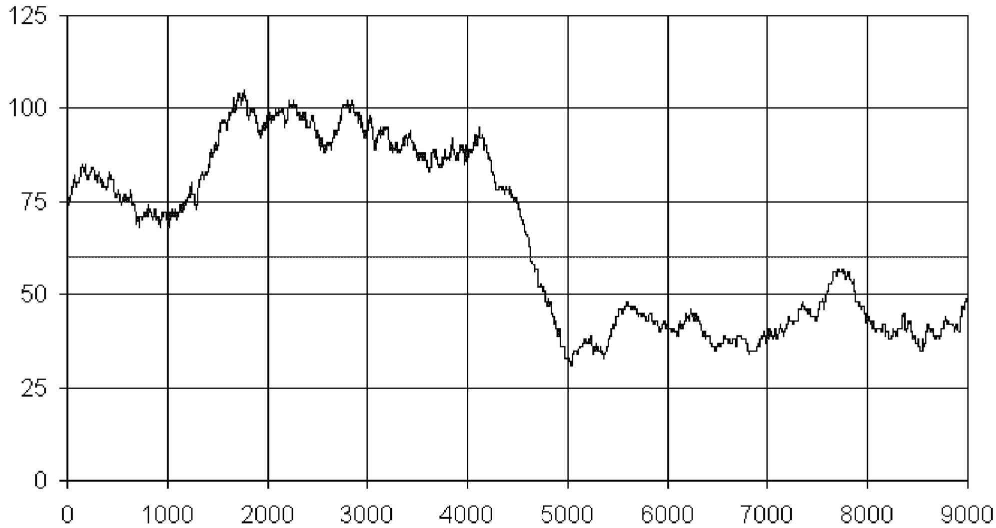

Рисунок 1- Графік, який показує кількість символів K_5 у кожний момент часу. Лінією позначено поріг C

Із графіка (рис. 1) видно, що розладнання відбулося приблизно в позиції 4630.

Приклад 2. Розв'яжемо ту саму задачу за алгоритмом 2.

За формулою (9) отримаємо $N=17600$. При такому об'ємі вибірки неможливо визначити момент розладнання, тому перенумеруємо символи алфавіту: $p_i=0{,}1$; $q_1=0{,}051$;

$q_2 = 0{,}053; q_3 = 0{,}06$; $q_4 = 0{,}08; q_5 = 0{,}1$; $q_6 = 0{,}101; q_7 = 0{,}105$; $q_8 = 0{,}14; q_9 = 0{,}15; q_{10} = 0{,}16$. При цьому зміняться вибіркові середні.

За формулами (9) і (6) знаходимо значення N і C: $N = 1192$, $C \approx 6{,}179$. Отже, для значень $n_в > 6{,}179$ можна вважати, що розладнання відбулося.

Із графіка (рис. 2) видно, що розладнання відбулося приблизно в позиції 4300.

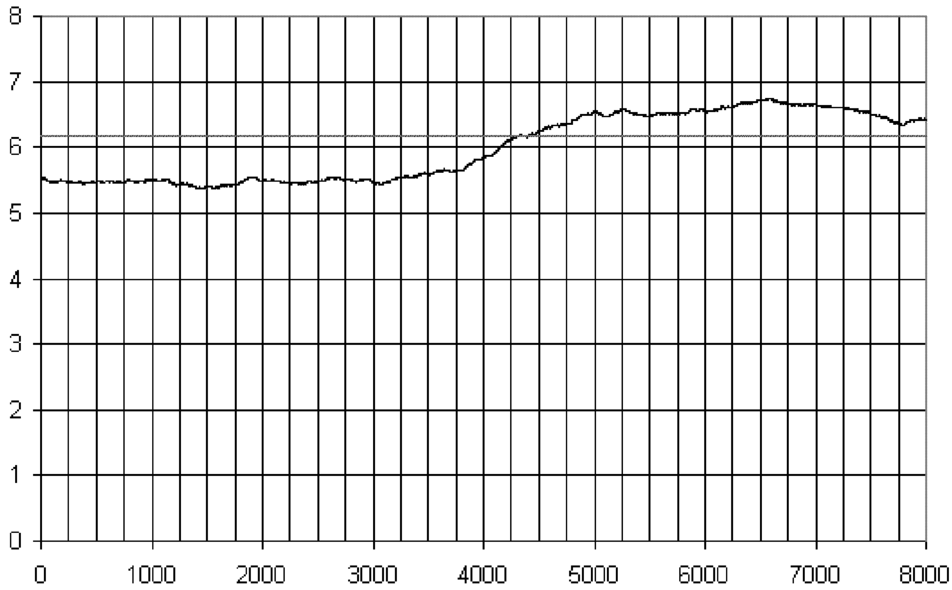

Рисунок 2- Графік випадкової величини $n_в$. Лінією позначена межа критичної області C

Зробимо порівняльний аналіз алгоритмів.

У кожному з алгоритмів був прийнятий рівень значущості $\alpha = 0{,}01$, потужність критерію $1 - \beta = 0{,}999$.

Порівнюючи результати розв'язання задачі за двома алгоритмами, отримано: момент розладнання за алгоритмом 1 відбувається в позиції 4630, за алгоритмом 2 — 4300. Мінімальна довжина N для алгоритмів рівна відповідно 863, 1192.

Таким чином, алгоритм 1 є більш точним, так як довжина спостережного інтервалу для нього є меншою, що дає змогу точніше визначити момент розладнання.

Висновки. Для знаходження моменту розладнання послідовності з n символів запропоновані алгоритми 1 і 2, в основі яких лежить властивість стійкості відносної частоти появи випадкової події в серії експериментів та властивість стійкості вибіркових середніх відповідно.

Практична цінність результатів дослідження визначається можливістю їх використання в схемах захисту інформації, а саме: для перевірки рівноймовірності випадкових послідовностей, при шифруванні, для виявлення зміни мови повідомлення, :для тестування якості випадкових послідовностей, перевірки однорідності послідовностей.

Список літератури

1. Дарховский Б.С., Бродский Б.Е. Непараметрический метод скорейшего обнаружения изменений среднего случайной последовательности // Теория вероятн. и ее примен. — 1987. — 32, №4. — С.703–711.
2. Николаев А.Ф. Об одной постановке задачи о множественной разладке // Теория вероятн. и ее примен. — 1998. — 43, №2. — С.370–374.
3. Кличене Н., Телькснис Л. Методы обнаружения моментов изменения свойств случайных последовательностей // Автоматика и телемеханика. — 1983. — №10. — С.5–56.

Г. Филимонихин, М. Гончарова

Исследование задачи о разладке последовательности из *n* символов

Исследуется задача о разладке последовательности из n символов и ее применение в схемах защиты информации. Построены два алгоритм для определения момента разладки и сделан сравнительный анализ их эффективности.

G. Filimonihin, M. Goncharova

Investigation of problem of the change moment of sequence from *n* symbols

Disorder problem of sequence from n symbols and its application in schemes of privacy are investigated. Two algorithms for determination of the discord moment are compounded. The comparative analysis of its effective is made.

Одержано 15.02.10