И. В. Чугунков, Д. А. Степаньян

ПОВЫШЕНИЕ ЭФФЕКТИВНОСТИ ТЕСТИРОВАНИЯ ГЕНЕРАТОРОВ ПСЕВДОСЛУЧАЙНЫХ ЧИСЕЛ

Принцип работы практически всех существующих тестов [1–4], используемых для оценки статистической безопасности генераторов псевдослучайных чисел (ПСЧ), основан на анализе тестовой статистики, при определении которой подсчитывается число исходов, относящихся к заданным категориям (например, в частотном тесте – это число появлений 0 и 1, в проверке равномерности – количество появлений каждого из возможных чисел [5]). В общем случае для хранения числа появлений наборов, состоящих из km-разрядных чисел, в последовательности длиной n чисел потребуется дополнительный объем памяти

$$V_H = 2^{km}\left\lceil \log_2\left(\left\lfloor \frac{n}{k} \right\rfloor + 1\right)\right\rceil$$

в случае непересекающихся наборов и

$$V_П = 2^{km}\left\lceil \log_2(n+1)\right\rceil$$

в случае пересекающихся.

Как можно заметить, с увеличением длины тестируемой последовательности возрастает и объем дополнительной памяти, требуемый для реализации теста. В предельном случае

$$V_H \underset{n\to\infty}{\to} \infty,\ V_П \underset{n\to\infty}{\to} \infty.$$

Таким образом, возникает задача уменьшения затрачиваемой для реализации теста памяти.

В настоящее время большинство разработчиков оценочных тестов (в том числе и авторы наиболее популярного на сегодняшний день Руководства НИСТ США [2]) решают данную проблему путем введения ограничений на размеры анализируемых наборов и длину исследуемой последовательности.

Уменьшение размеров наборов приводит к существенному ослаблению теста, так как, зная логику работы последнего, можно внести соответствующие изменения в алгоритм работы генератора ПСЧ или непосредственно в саму псевдослучайную последовательность (ПСП), благодаря чему свойства последней будут неотличимы от свойств истинно случайной последовательности для типа неслучайности, проверяемого заданным тестом. Данное утверждение косвенно подтверждается результатами исследований наиболее эффективных из существующих генераторов ПСЧ [5]. Все они с легкостью проходят тест проверки серий с наборами из 2 и 3 бит [1–5] и полностью проваливают критерий равномерности с набором из 8 бит [1, 2]. И это только для наборов, состоящих из одного числа. Можно предположить, что увеличение количества чисел в наборах сделает статистику прохождения еще более удручающей.

Уменьшение длины исследуемой последовательности также снижает качество тестирования. Для примера, разработчики Руководства НИСТ рекомендуют использовать для анализа последовательности длиной всего лишь 2^{20} бит или 128 Кбайт, что для существующих объемов передаваемой информации, исчисляемой мегабайтами, гигабайтами и даже терабайтами, является просто недопустимым.

Один из вариантов решения данной проблемы косвенно предложен в Руководстве НИСТ. Суть его заключается в тестировании не всей последовательности целиком, а подпоследовательностей. Однако данный подход не лишен недостатков. Основной из них связан с выбором размера подпоследовательности. Во-первых, не существует детерминированного алгоритма определения требуемого размера подпоследовательности для заданных теста и длины

тестируемой ПСП. Во-вторых, при увеличении размера подпоследовательности опять возникает проблема дополнительной памяти. Уменьшение же размера подпоследовательностей приводит к необходимости оценки корреляции между последними для исключения влияния периодичности.

Выход из сложившейся ситуации видится в модификации механизма работы тестов, суть которой заключается в том, чтобы анализировать не число появлений определенных наборов, а число отсутствующих наборов. В этом случае для хранения информации о наборе достаточно будет 1 бита, что приведет к уменьшению объема памяти до 2^{km} бит как для пересекающихся, так и для непересекающихся наборов, при этом цель теста не будет изменена и он продолжит выявлять те же статистические отклонения, что и оригинальный тест. Таким образом, размер требуемой для реализации теста памяти перестает зависеть от длины исследуемой последовательности. К примеру, для реализации тестов НИСТ при рекомендуемой длине в 128 Кбайт объем дополнительной памяти уменьшается в $\lceil \log_2(2^{20}+1) \rceil = 21$ раз.

Рассмотрим, как меняется механизм вычисления статистики теста. Для заданной последовательности длиной n, состоящей из наборов по km-разрядных чисел, подсчитываем число отсутствующих наборов $\mu_э$. Дж. Марсалья в своей работе [6] показал, что число отсутствующих наборов аппроксимируется с нормальным распределением. Найдем среднее μ и отклонение σ.

Для расчета среднего необходимо рассмотреть разложение в ряд Тейлора производящей функции [7], соответствующей значениям k и m. Очевидно, это не очень удобно, поскольку для различных наборов придется заново определять производящую функцию для каждого типа набора и раскладывать ее в ряд Тейлора. Например, для $k = 2$ расчет осуществляется следующим образом [6]:

- вычисляется ρ_1 — коэффициент при z_n в разложении производящей функции

$$\frac{1}{1-z+p^2z^2};$$

- вычисляется ρ_2 — коэффициент при z_n в разложении производящей функции

$$\frac{1+pz}{1-(1-p)z-(p-p^2)z^2};$$

- вычисляется среднее для числа отсутствующих слов

$$\mu = \left(2^{km}-2^m\right)\cdot p_1 + 2^m \cdot p_2 .$$

С увеличением k увеличивается число производящих функций, а также их сложность. Поэтому для расчета предлагается использовать подход Дж. Марсальи, который в работе [6] показал, что среднее можно вычислить (при условии, что $n > 1000$) по формуле

$$\mu_{\approx} = 2^{km} e^{-\frac{n}{2^{km}}}.$$

Действительно, для случая $n = 2^{21}$, $k = 2$, $m = 10$ имеем:

$$p_1 = 0{,}135335283236469;$$

$$p_2 = 0{,}135599351997986596411;$$

$$\mu = \left(2^{2\times10}-2^{10}\right)\times 0{,}135335283236469 + 2^{10}\times 0{,}135599351997986596411$$

$$\approx 141909{,}60;$$

$$\mu_{\approx} = 2^{2\cdot10} e^{-\frac{2^{21}}{2^{2\cdot10}}} \approx 141909{,}33;$$

$$\mu_{\approx} \approx \mu .$$

Расчет отклонения осуществить гораздо сложнее. Точное значение отклонения равно

$$\sigma = \sqrt{\sum_{i_1=1}^{2^m}\sum_{i_2=1}^{2^m}\ldots\sum_{i_k=1}^{2^m} cov\left(n_{i_1}, n_{i_2}, \ldots, n_{i_k}\right)},$$

где $cov\left(n_{i_1}, n_{i_2}, \ldots, n_{i_k}\right)$ – ковариация $n_{i_1}, n_{i_2}, \ldots, n_{i_k}$ [7],

$$n_i = \begin{cases} 1, \text{если буква, равная } 2^i, \text{ присутствует в слове;} \\ 0, \text{если буква, равная } 2^i, \text{ отсутствует в слове.} \end{cases}$$

Расчет ковариации для двух переменных уже является сложной задачей. Как правило, используют готовые ковариационные матрицы, в том числе и полученные в результате испытаний. Кроме этого, некоторые специализированные программные продукты, например MathCad, позволяют вычислять ковариации для небольшого числа переменных.

В связи с этим предлагается две методики оценки полученного в результате теста числа отсутствующих наборов $\mu_э$. Первая заключается в выборе небольшого значения k, например не более 5, расчете отклонения и вычислении статистики теста

$$P\text{-}value = \frac{erfc\left(\frac{\mu_э - \mu_\approx}{\sqrt{2}\sigma}\right)}{2},$$

где $erfc(\,)$ – дополнительная функция ошибок [7].

Вторая методика связана с использованием функции Лапласа [8]. Пусть для заданной последовательности длины n результаты испытаний можно разделить на s категорий с вероятностью попадания в каждую категорию, равной p_s^T. Пусть $p_s^э$, $i = \overline{1,s}$ – частота попадания в s-ю категорию, полученная в результате испытаний. Тогда вероятность того, что отклонение $p_s^э$ от p_s^T не превышает заданного значения $\xi > 0$, приближенно [8] равна удвоенной функции Лапласа от $\xi \cdot \sqrt{\frac{n}{p_s^T\left(1-p_s^T\right)}}$:

$$P\left(\left|p_s^э - p_s^T\right| \le \xi\right) \approx 2\,\Phi_L\left(\xi \cdot \sqrt{\frac{n}{p_s^T\left(1-p_s^T\right)}}\right).$$

Для уровня значимости теста α

$$P\left(\left|\frac{\nu_i}{n} - \frac{1}{2^m}\right| \le \xi\right) = 1 - \alpha.$$

Следовательно,

$$2\,\Phi_L\left(\xi \cdot \sqrt{\frac{n}{p_s^T\left(1-p_s^T\right)}}\right) = 1 - \alpha,$$

$$\xi = \Phi_L^{-1}\left(\frac{1-\alpha}{2}\right) \cdot \sqrt{\frac{p_s^T\left(1-p_s^T\right)}{n}},$$

где $\Phi_L^{-1}()$ – обратная функция Лапласа.

Таким образом, с вероятностью $1 - \alpha$

$$\left|p_s^э - p_s^T\right| \le \Phi_L^{-1}\left(\frac{1-\alpha}{2}\right) \cdot \sqrt{\frac{p_s^T\left(1-p_s^T\right)}{n}}$$

или

$$p_s^T - \Phi_L^{-1}\left(\frac{1-\alpha}{2}\right)\cdot\sqrt{\frac{p_s^T\left(1-p_s^T\right)}{n}} \le p_s^{\text{Э}} \le p_s^T + \Phi_L^{-1}\left(\frac{1-\alpha}{2}\right)\cdot\sqrt{\frac{p_s^T\left(1-p_s^T\right)}{n}}.$$

С учетом того, что

$$p_s^{\text{Э}} = \frac{\mu_{\text{Э}}}{n} \text{ и } p_s^T = \frac{\mu_{\approx}}{n},$$

получаем

$$\mu_{\approx} - \Phi_L^{-1}\left(\frac{1-\alpha}{2}\right)\cdot\sqrt{\frac{\mu_{\approx}\left(n-\mu_{\approx}\right)}{n}} \le \mu_{\text{Э}} \le \mu_{\approx} + \Phi_L^{-1}\left(\frac{1-\alpha}{2}\right)\cdot\sqrt{\frac{\mu_{\approx}\left(n-\mu_{\approx}\right)}{n}}$$

Подсчет числа отсутствий набора вместо числа появлений может использоваться в ряде существующих тестов (таблица 1). При этом следует отметить, что просто заменить методику подсчета нельзя, так как в большинстве случаев размеры наборов слишком малы, чтобы число отсутствующих наборов могло аппроксимироваться с нормальным законом распределения.

Таблица 1. Тесты, допускающие возможность использования подсчета числа пропущенных слов

№	Название теста
Система «DIEHARD»	
1	Проверка пересекающихся перестановок
2	Проверка рангов матриц
3	Обезьяньи тесты
4	Подсчет единиц
5	Тест игры в кости
Подборка Д. Кнута	
6	Проверка равномерности
7	Проверка серий
8	Покер-тест
9	Тест собирателя купонов
10	Проверка перестановок
Руководство НИСТ	
11	Проверка рангов матриц
12	Проверка непересекающихся шаблонов
13	Проверка пересекающихся шаблонов
14	Проверка серий
Система CRYPT-X	
15	Проверка серий
16	Бинарное ускорение в подпоследовательностях

В связи с этим рекомендуется осуществлять комбинированное тестирование — вначале классическим тестом применительно к небольшим размерам наборов (несколько бит), затем

модификацией теста с подсчетом числа отсутствий для наборов большого размера. Для достижения требуемой аппроксимации желательно, чтобы $n = t2^{km}$, $t = \overline{1,4}$ [8].

Можно выделить следующие положительные эффекты от использования подсчета числа отсутствующих наборов.

- Размер вспомогательной памяти для сбора статистики теста теперь не зависит от длины тестируемой последовательности и определяется только размерами набора. Это позволит выделять данную память статически, что ускорит выполнение теста при программной реализации, или прогнозировать объем выделяемой динамической памяти.
- Учитывая, что значение длины последовательности теперь не влияет на объем затрачиваемой памяти, модифицированные оценочные тесты можно будет применять не к части последовательности фиксированной длины (ограниченной в существующих системах несколькими мегабайтами), а практически ко всему периоду ПСП.
- Механизм работы модифицированных тестов предполагает досрочное завершение процесса исследования в том случае, если исследуемая статистика вышла за границы доверительного интервала, что существенно сократит время тестирования.
- Уменьшение объемов требуемой памяти позволит более свободно варьировать параметрами тестирования, увеличив диапазон используемых значений, что существенно повысит функциональность тестов и качество тестирования.

СПИСОК ЛИТЕРАТУРЫ:

1. *Кнут Д.* Искусство программирования. Том 2. Получисленные алгоритмы: Пер. с англ. 3-е изд. М.: Издательский дом «Вильямс». 2007. — 832 с.: илл.
2. A statistical test suite for random and pseudorandom number generators for cryptographic applications. NIST Special Publications 800-22. Revision 1.a. April, 2010.
3. *Gustafson H., Dawson E., Nielsen L., Caelli W.* A computer package for measuring the strength of encryption algorithms // Computer & Security. 1994. Vol. 13. Issue 8. P. 687–697.
4. *Marsaglia G.* DIEHARD: Battery of tests of randomness. URL: http://stat.fsu.edu/pub/diehard.
5. *Иванов М. А., Чугунков И. В.* Теория, применение и оценка качества генераторов псевдослучайных последовательностей. М.: КУДИЦ-ОБРАЗ, 2003. — 240 с.
6. *Marsaglia G., Zaman A.* Monkey tests for random number generators // Computers and Mathematics with Applications. 1993. Vol. 26. № 9. P. 1–10.
7. *Корн Г., Корн Т.* Справочник по математике для научных работников и инженеров: Пер. с англ. / Под ред. И. Г. Арамановича. М.: Наука, 1973. — 832 с.: илл.
8. *Гмурман В. Е.* Теория вероятностей и математическая статистика. М.: Высшая школа, 2003. — 479 с.