УДК 519.233.3:519.722:004.421.5

Энтропийный анализ криптографических генераторов случайных и псевдослучайных последовательностей

Представлен метод оценки качества криптографических генераторов на основе энтропии Шеннона. По выходной последовательности вычисляется статистическая оценка энтропии, для которой известно асимптотическое распределение вероятностей, и применяется решающее правило, основанное на вычисленной оценке. Приводится теоретическое обоснование метода и описание класса последовательностей, которые решающее правило может отличить от чисто случайной. Проведены компьютерные эксперименты, демонстрирующие применение решающего правила.

Ю.С. ХАРИН,
член-корреспондент НАН Беларуси, доктор физ.-мат. наук, профессор, директор

В.Ю. ПАЛУХА,
младший научный сотрудник

НИИ прикладных проблем математики и информатики

Ключевые слова:
криптографические генераторы случайных и псевдослучайных последовательностей, энтропия Шеннона, статистическая проверка гипотез.

Введение. Важным структурным элементом средств криптографической защиты информации (криптосистем) являются генераторы случайных и псевдослучайных последовательностей [1]. Стойкость криптосистем зависит от того, насколько близка генерируемая последовательность по своим свойствам к равномерно распределенной случайной последовательности (РРСП), так называемой чисто случайной. Для проверки качества криптографических генераторов используются статистические тесты, суть которых заключается в следующем. Наблюдается выходная последовательность криптографического генератора и вводится гипотеза H_* о том, что последовательность является РРСП. Вычисляется некоторая статистика, распределение вероятностей которой при истинной гипотезе H_* известно. На основании значения статистики гипотеза H_* принимается либо отклоняется. В данной статье рассматривается применение статистической оценки энтропии Шеннона в качестве тестовой статистики для анализа выходных последовательностей криптографических генераторов.

Математическая модель. Пусть на вероятностном пространстве (Ω, F, P) с множеством состояний $\Omega = \{\omega_1, \ldots, \omega_N\}$ определена случайная величина $x = x(\omega) = \omega$ с дискретным распределением

вероятностей $P = \{p_k\}$, $p_k = P\{x = \omega_k\}$, $p_k \geq 0$, $\sum_{k=1}^{N} p_k = 1$, $k = 1, \ldots, N$. Энтропия Шеннона определяется формулой [1]:

$$H(P) = -\sum_{k=1}^{N} p_k \ln p_k. \quad (1)$$

Пусть имеется случайная последовательность $\{x_t: t = 1, \ldots, N\}$ объема n из распределения вероятностей $\{p_k\}$. Построим частотные оценки распределения вероятностей $\{p_k: k = 1, \ldots, N\}$:

$$\hat{p}_k = \frac{\nu_k}{n},$$

$$\nu_k = \sum_{t=1}^{n} I\{x_t = \omega_k\} \in \mathbb{N}_0 = \mathbb{N} \cup \{0\}, \quad (2)$$

$$I\{x_t = \omega_k\} = \begin{cases} 1, x_t = \omega_k; \\ 0, x_t \neq \omega_k. \end{cases}$$

Как было сказано во введении, рассмотрим гипотезу $H_\star = \{\{x_t\}$ является РРСП$\} = \{\{x_t\}$ – независимые одинаково распределенные случайные величины, $p_k = 1 / N$, $k = 1, \ldots, N\}$ и альтернативу $\overline{H_\star}$.

Следуя [2], будем полагать, что имеет место схема серий. В таком случае вектор $(\nu_1, \ldots, \nu_N)^T$, составленный из частот ν_k из (2), имеет полиномиальное распределение вероятностей $\mathrm{Pol}(n, N, p_1, \ldots, p_N)$, а каждая из компонент распределена по биномиальному закону $\mathrm{Bi}(n, p_k)$. Рассмотрим асимптотику, в которой длительность наблюдения n и число значений N растут синхронно:

$$n, N \to \infty, n / N \to \lambda, 0 < \lambda < \infty. \quad (3)$$

В асимптотике (3) для распределения вероятностей статистик $\{\nu_k\}$ справедлива аппроксимация законом Пуассона $\Pi(\lambda_k)$ с параметром $\lambda_k = np_k$ [3]. При истинной гипотезе $H_\star$ все элементарные вероятности равны: $p_k = 1 / N$, $k = 1, \ldots, N$, поэтому все частоты $\{\nu_k\}$ имеют одинаковый параметр распределения Пуассона $\lambda = n / N$.

Статистическая оценка энтропии Шеннона $\hat{H}(n, N)$, построенная по подстановочному принципу с использованием частотных оценок вероятностей, записывается следующим образом:

$$\hat{H} = \hat{H}(n, N) = -\sum_{k=1}^{N} \hat{p}_k \ln \hat{p}_k = -\sum_{k=1}^{N} \frac{\nu_k}{n} \ln \frac{\nu_k}{n}. \quad (4)$$

Справедлива доказанная в [4] теорема об асимптотическом распределении вероятностей статистики (4).

Теорема. В асимптотике (3) статистика (4) при истинной гипотезе $H_\star$ имеет асимптотически нормальное распределение $\mathcal{L}\left\{\frac{\hat{H} - \mu_H}{\sigma_H}\right\} \to \mathcal{N}_1(0,1)$:

$$\mu_H = \ln n - e^{-\lambda} \sum_{k=1}^{+\infty} \frac{\ln(k+1)\lambda^k}{k!}, \quad (5)$$

$$\sigma_H^2 = \frac{e^{-\lambda}}{n} \sum_{k=1}^{+\infty} \frac{(k+1)\lambda^k}{k!} \ln^2(k+1) -$$
$$- \frac{e^{-2\lambda}}{N} \left(\sum_{k=1}^{+\infty} \frac{\ln(k+1)\lambda^k}{k!} \right)^2 -$$
$$- \frac{e^{-2\lambda}}{n} \left(\sum_{k=1}^{+\infty} \ln(k+1) \frac{\lambda^k}{k!} (k+1-\lambda) \right)^2.$$

Знание асимптотического распределения точечной оценки (4) позволяет построить интервальную оценку энтропии Шеннона. С вероятностью $1 - \varepsilon$ оценка энтропии $\hat{H}(P) \in (H_-, H_+)$ такова:

$$H_\pm = \mu_H \pm \sigma_H \Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right), \quad (6)$$

где $\Phi(\cdot)$ – функция распределения стандартного нормального закона [3]. Решающее правило, основанное на интервальной оценке (6), имеет вид:

$$\begin{cases} H_\star, & \text{если } t_- < \hat{H}(n, N) < t_+; \\ \overline{H_\star}, & \text{в противном случае,} \end{cases} \quad t_\pm = \mu_H \pm \sigma_H \Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right). \quad (7)$$

Энтропийный анализ выходных последовательностей. Построенная точечная оценка (4) является смещенной, что продемонстрировано в [5]. Как известно, истинное значение энтропии Шеннона при справедливости гипотезы $H_\star$ достигает своего максимального значения $\ln N$. В отличие от него значение статистической оценки энтропии Шеннона (4) наблюдаемой последовательности может быть как меньше, так и больше математического ожидания (5), если гипотеза $H_\star$ верна, поскольку математическое ожидание (5) в асимптотике (3) не превышает максимальное возможное истинное значение энтропии: $\mu_H \leq \ln N$. В [5] показано, что с уменьшением λ растет разность между μ_H и $\ln N$, поэтому решающее правило (7) является двусторонним. Основанием для отклонения гипотезы $H_\star$ может являться как малое, так и большое значение статистической оценки энтропии Шеннона.

Приведем примеры обеих ситуаций. Поведение оценки (4) зависит от значений частот ν_k. Предположим, что они сильно различаются. Например, половина частот равна нулю – такая ситуация наблюдается при подсчете частот $(s + 1)$-грамм

у выходной последовательности регистра сдвига порядка s. Из неравенства Йенсена [1] следует, что $\hat{H} < \ln N$, а в случае большой разбежки в значениях частот будет выполняться и $\hat{H} \le t_-$, т. е. гипотеза H_* будет отклонена в силу выхода значения оценки энтропии за нижнюю границу доверительного интервала.

Пусть теперь значения частот ν_k приблизительно равны. В асимптотике (3) число наблюдений, приходящихся на каждый элемент алфавита, конечно и относительно невелико, и для чисто случайной последовательности значения частот ν_k могут отличаться между собой в некоторых допустимых пределах. При приблизительно равных значениях частот ν_k значение статистики (4) будет стремиться к максимуму и превысит математическое ожидание (5), которое, как уже было сказано, при малых λ меньше $\ln N$. Однако равные значения частот ν_k могут свидетельствовать о наличии зависимости, некоторого правила при генерации элементов последовательности и, возможно, даже периода. Примером такого генератора является регистр сдвига, в выходной последовательности которого поочередно встречаются все возможные значения s-грамм. Тест (7) позволяет идентифицировать такой генератор, поскольку в этом случае будет выполнено условие $\hat{H} \ge t_+$, т. е. гипотеза H_* будет отклонена в силу выхода значения оценки энтропии за верхнюю границу доверительного интервала.

Теоретически описанная выше ситуация может встретиться и у чисто случайной последовательности, что повлечет отклонение гипотезы H_*. Однако вероятность такого события (и, соответственно, ошибки второго рода) крайне мала. Эта вероятность зависит от длины последовательности, мощности алфавита и уровня значимости, и для установления точного вида зависимости требуются дополнительные исследования.

Чтобы провести качественный энтропийный анализ и уменьшить вероятность ошибок первого и второго рода, для двоичных последовательностей рекомендуется применять следующий подход. Выходная последовательность «разрезается» на непересекающиеся, идущие подряд фрагменты длины s (s-граммы): $X^{(t)} = (X_j^{(t)}) = (y_{(t-1)s+1}, \ldots, y_{ts}) \in \{0,1\}^s$, $t = 1, \ldots, n = [T / s]$. Из полученных s-грамм формируется новая последовательность $\{x_t\}$ из алфавита мощности $N = 2^s$ по правилу $x_t = \sum_{j=1}^{s} 2^{j-1} X_j^{(t)} + 1$. «Разрезание» последовательности на фрагменты различных длин позволяет исследовать сформированные последовательности различных алфавитов. Анализ результатов применения разработанного статистического теста при различных мощностях алфавита позволяет принять взвешенное решение относительно принятия либо отклонения гипотезы H_*. Возможны следующие варианты. Гипотеза H_* принимается, когда решающее правило дало положительный результат:

хотя бы при одном рассмотренном значении s;

при всех рассмотренных значениях s;

при нескольких рассмотренных значениях s, количество которых превышает некоторое заданное наперед пороговое число.

Выбор принципа принятия решения зависит от конкретной ситуации и остается за аналитиком.

Результаты компьютерных экспериментов. Для демонстрации разработанного статистического теста были проведены компьютерные эксперименты на последовательностях физического генератора и регистра сдвига. На рисунках 1–3 приведены значения отклонений оценки энтропии Шеннона (4) от математического ожидания (5), деленные на верхние границы доверительных интервалов: $\dfrac{\hat{H} - \mu_H}{\sigma_H \Phi^{-1}\left(1 - \varepsilon/2\right)}$, на уровне значимости $\varepsilon = 0{,}05$ в зависимости от s. На рисунке 1 представлен результат анализа последовательности длины $T = 2^{31}$ линейного регистра сдвига с характеристическим многочленом 25-й степени, на рисунке 2 – результат анализа последовательности длины $T = 2^{28}$ этого же регистра, на рисунке 3 – результат анализа последовательности длины $T = 500 \times 2^{23}$ физического генератора [6].

Как видно на рисунке 1, гипотеза H_* отклоняется не только при $s > 25$, когда оценка энтропии (4) меньше ожидаемого значения (5) из-за того, что превышен порядок регистра, но и при $s \le 25$, поскольку в этом случае выходная последовательность является «слишком правильной» и значение оценки энтропии (4) превышает математическое ожидание (5). На рисунке 2 показано, что для отклонения гипотезы H_* достаточно меньшей длины последовательности. Она все равно является «слишком правильной», и значение оценки энтропии (4) превышает математическое ожидание (5). Из рисунка 3 следует, что

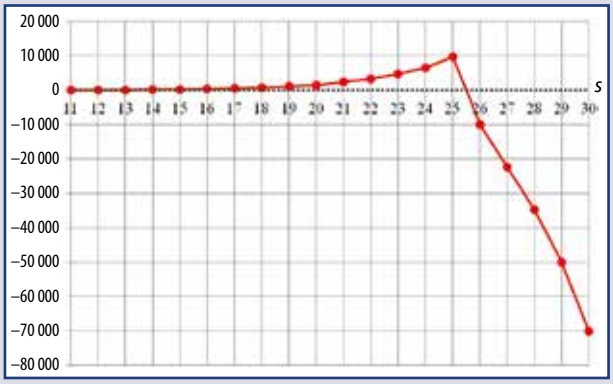

Рисунок 1 – Нормированное отклонение оценки энтропии регистра сдвига, $T = 2^{31}$

Рисунок 2 – Нормированное отклонение оценки энтропии регистра сдвига, $T = 2^{28}$

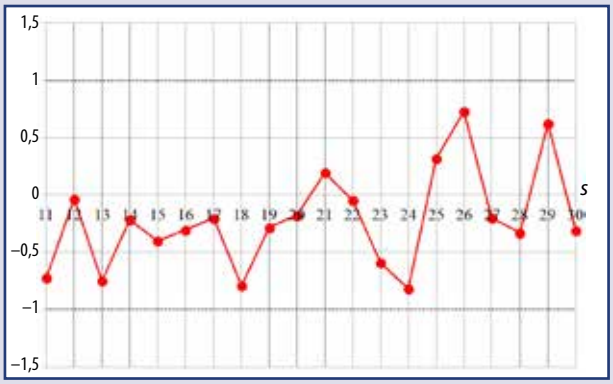

Рисунок 3 – Нормированное отклонение оценки энтропии физического генератора

у выходной последовательности генератора [6] значение оценки (4) не выходит за границы доверительного интервала и гипотеза H_* принимается.

Заключение. В настоящей работе рассмотрена разработка решающего правила (статистического теста) для проверки качества криптографических генераторов на основе статистической оценки энтропии Шеннона выходной последовательности генератора. Приводится пример класса последовательностей, которые разработанное решающее правило может отличить от равномерно распределенной случайной последовательности. Показано применение теста на примере выходных последовательностей физического генератора и регистра сдвига.

ЛИТЕРАТУРА

1. **Харин, Ю.С.** Криптология / Ю.С. Харин [и др.]. – Минск: БГУ, 2013. – 512 с.
2. **Holst, L.** Asymptotic normality and efficiency for certain goodness-of-fit tests / L. Holst // Biometrika. – 1972. – № 59. – P. 137–145.
3. **Харин, Ю.С.** Теория вероятностей, математическая и прикладная статистика / Ю.С. Харин, Н.М. Зуев, Е.Е. Жук. – Минск: БГУ, 2011. – 463 с.
4. **Палуха, В.Ю.** Статистические тесты на основе оценок энтропии для проверки гипотез о равномерном распределении случайной последовательности / В.Ю. Палуха // Весці НАН Беларусі. Серыя фізіка-матэматычных навук. – 2017. – № 1. – С. 79–88.
5. **Палуха, В.Ю.** Энтропийные характеристики двоичных последовательностей в криптографии / В.Ю. Палуха, Ю.С. Харин // Комплексная защита информации. Материалы XX научно-практической конференции. Минск, 19–21 мая 2015 г. – Минск: РИВШ, 2015. – С. 99–102.
6. Speedtest-500MB.bin [Electronic resource] // Humboldt Berlin University, Faculty of Mathematics and Natural Sciences, Department of Physics. – Mode of access: http://qrng.physik.hu-berlin.de/files/speedtest-500MB.bin. – Date of access: 07.05.2016.

A method for testing the cryptographic generators quality based on Shannon entropy is presented. The statistical estimator of entropy, for which the asymptotic probability distribution is known, is calculated for the output sequence and the decision rule based on the calculated estimator is constructed. The theoretical substantiation of the method and the description of the class of sequences that can be distinguished from the pure random sequence is given. Results of computer experiments illustrating the application of the constructed decision rule were presented.

Получено 27.09.17.