УДК 519.2

ЭНТРОПИЙНЫЕ ФУНКЦИОНАЛЫ РЕНЬИ И ТСАЛЛИСА В ЗАДАЧАХ ЗАЩИТЫ ИНФОРМАЦИИ

В.Ю. Палуха, Ю.С. Харин

Рассматривается задача оценки качества криптографических генераторов по их выходным последовательностям. Построены статистические оценки функционалов энтропии Реньи и Тсаллиса с использованием частотных статистик, приведены вероятностные свойства построенных оценок энтропии. Разработан критерий качества криптографических генераторов псевдослучайных последовательностей, в основе которого лежит вычисление оценки энтропии наблюдаемой выходной последовательности генератора.

Введение

Основным элементом современных средств криптографической защиты информации являются генераторы псевдослучайных последовательностей. Стойкость криптосистем зависит от того, насколько близка генерируемая последовательность по своим вероятностным свойствам к равномерно распределённой случайной последовательности (РРСП). Одним из подходов к оценке качества генератора является статистическое оценивание энтропии и сравнение полученной оценки с ожидаемым значением для РРСП. Помимо наиболее известной энтропии Шеннона, в последнее время внимание исследователей привлекают также функционалы энтропии Реньи и Тсаллиса [1]. Применению этих функционалов к задаче оценки качества генераторов псевдослучайных последовательностей и посвящена данная статья.

Математическая модель

Пусть на вероятностном пространстве (Ω, F, P) с пространством элементарных событий $\Omega = \{\omega_1, \ldots, \omega_N\}$ определена случайная величина $x = x(\omega) = \omega$ с дискретным распределением вероятностей $p_k = P\{x = \omega_k\}$, $p_k \geq 0$, $\sum_{k=1}^{N} p_k = 1, k = 1, \ldots, N$. Энтропия Реньи определяется формулой [1]

$$H_r(P) = \frac{1}{1-r} \log\left(\sum_{i=1}^{M} p_i^r \right), \tag{1}$$

а энтропия Тсаллиса – формулой [1]

$$S_r(P) = \frac{1}{r-1}\left(1 - \sum_{i=1}^{M} p_i^r \right), \tag{2}$$

где r – действительный параметр, $r > 0, r \neq 1$. Ограничимся рассмотрением натуральных $r > 1$.

Построение статистических оценок энтропии на основе частотных оценок вероятностей

Пусть имеется случайная последовательность $\{x_t : t = 1, \ldots, n\}$ длины n из распределения вероятностей $\{p_k\}$. Обозначим через $I\{A\}$ индикатор события A. Построим частотные оценки распределения вероятностей $\{p_k : k = 1, \ldots, N\}$:

$$\hat{p}_k = \frac{v_k}{n}, \quad v_k = \sum_{t=1}^{n} I\{x_t = \omega_k\}, \quad I\{x_t = \omega_k\} = \begin{cases} 1, x_t = \omega_k; \\ 0, x_t \neq \omega_k. \end{cases} \tag{3}$$

Введём в рассмотрение гипотезу H_* = {{x_t} есть РРСП} = {{x_t} – независимые одинаково распределённые случайные величины, $p_k = 1/N$, k = 1, …, N}.

Следуя [2], будем полагать, что имеет место схема серий. В таком случае вектор $(v_1, \ldots, v_N)^T$, составленный из величин v_k из (3), имеет полиномиальное распределение вероятностей Pol(n, N, p_1, …, p_N), а каждая из компонент распределена по биномиальному закону $Bi(n, p_k)$. Рассмотрим асимптотику:

$$n, N \to \infty, n/N \to \lambda, 0 < \lambda < \infty. \tag{4}$$

В асимптотике (4) для распределения вероятностей величин $\{v_k\}$ справедлива аппроксимация законом Пуассона $\Pi(\lambda_k)$ с параметром $\lambda_k = np_k$ [3]. При истинной гипотезе H_* имеем $p_k = 1/N$, $k = 1, \ldots, N$, поэтому все величины $\{v_k\}$ имеют одинаковый параметр распределения $\lambda = n/N$.

Как видно из формул (1) и (2), энтропии Реньи и Тсаллиса являются функциями от величины

$$P_r(P) = \sum_{k=1}^{N} p_k^r. \tag{5}$$

Рассмотрим задачу статистического оценивания величины $P_r(P)$. Оценка (5) по подстановочному принципу $\hat{P}_r(P) = \sum_{k=1}^{N} \hat{p}_k^r = \sum_{k=1}^{N} \left(\frac{v_k}{n} \right)^r$ является смещённой [4]. Для построения несмещённой оценки определим r-ую нисходящую факториальную степень x:

$$x^{\underline{r}} = x(x-1)\ldots(x-r+1) = \frac{x!}{(x-r)!} = \sum_{i=0}^{r} s(r,i) x^i, \tag{6}$$

где $s(r, i)$ – число Стирлинга первого рода [5]. По определению, при $x < r$ полагают $x^{\underline{r}} ::= 0$. В [4] предложена статистическая оценка для величины (5), которая основана на (6) и при справедливости (4) является асимптотически несмещённой:

$$\hat{P}_r(P) = \sum_{k=1}^{N} \frac{v_k^{\underline{r}}}{n^r}. \tag{7}$$

Положим

$$Z_{n,r} = \sum_{k=1}^{N} v_k^{\underline{r}}. \tag{8}$$

Из [2] следует, что в асимптотике (4) статистика (8) имеет асимптотически нормальное распределение $Z_{n,r} \square \mathrm{N}_1(\mu_{n,r}, \sigma_{n,r}^2)$:

$$\mu_{n,r} = \sum_{k=1}^{N} E\{v_k^{\underline{r}}\}, \tag{9}$$

$$\sigma_{n,r}^2 = \sum_{k=1}^{N} D\{v_k^{\underline{r}}\} - \left(\sum_{k=1}^{N} \mathrm{cov}\{v_k, v_k^{\underline{r}}\} \right)^2 \Big/ n, \tag{10}$$

где $\mathrm{N}_1(\mu_{n,r}, \sigma_{n,r}^2)$ – одномерный нормальный закон распределения вероятностей с математическим ожиданием $\mu_{n,r}$ и дисперсией $\sigma_{n,r}^2$, $E\{\xi\}$ и $D\{\xi\}$ – соответственно математическое ожидание и дисперсия случайной величины ξ, $\mathrm{cov}\{\xi, \eta\}$ – ковариация случайных величин ξ и η.

При истинной гипотезе H_* (9) и (10) принимают вид:

$$\mu_{n,r} = \sum_{k=1}^{N} E\{v_k^{\underline{r}}\} = NE\{v^{\underline{r}}\},$$

$$\sigma_{n,r}^2 = ND\{v^{\underline{r}}\} - N^2 \mathrm{cov}^2\{v, v^{\underline{r}}\}/n = N\left(D\{v^{\underline{r}}\} - \mathrm{cov}^2\{v, v^{\underline{r}}\}/\lambda \right).$$

Из [4] следует, что если случайная величина v распределена по закону Пуассона с параметром λ, т.е. $\mathrm{L}\{v\} = \Pi(\lambda)$, то $E\{v^{\underline{r}}\} = \lambda^r$. Тогда легко показать, что $D\{v^{\underline{r}}\} = \lambda^r \left(E\{(v+r)^{\underline{r}}\} - \lambda^r \right) = \lambda^r E\{(v+r)^{\underline{r}} - v^{\underline{r}}\}$, $\mathrm{cov}\{v, v^{\underline{r}}\} = r\lambda^r$. Кроме того, согласно [6], $E\{v^r\} = \sum_{i=0}^{r} S(r,i)\lambda^i$, где $S(r, i)$ – число Стирлинга второго рода [5]. Из вышесказанного следует справедливость следующей теоремы.

Теорема 1. При истинной гипотезе H_* в асимптотике (4) статистика (8) имеет асимптотически нормальное распределение $Z_{n,r} \square \mathrm{N}_1(\mu_{n,r}, \sigma_{n,r}^2)$:

$$\mu_{n,r} = N\lambda^r = n\lambda^{r-1},$$

$$\sigma_{n,r}^2 = N\lambda^r \left(\sum_{i=1}^{r} s(r,i) \sum_{j=0}^{i-1} C_i^j r^{i-j} \sum_{k=1}^{j} S(j,k)\lambda^k - r^2\lambda^{r-1} + r! \right) =$$

$$= n\lambda^{r-1} \left(\sum_{i=1}^{r} s(r,i) \sum_{j=0}^{i-1} C_i^j r^{i-j} \sum_{k=1}^{j} S(j,k)\lambda^k - r^2\lambda^{r-1} + r! \right).$$

Следствие 1. При $r = 2$ для параметров асимптотического распределения случайной величины $Z_{n,2}$ справедливы выражения:

$$\mu_{n,2} = n\lambda, \ \ \sigma_{n,2}^2 = 2n\lambda.$$

Статистические оценки энтропии Реньи и Тсаллиса, построенные с использованием (6), выражаются через статистику (8):

$$\hat{H}_r = \frac{1}{1-r}\ln\left(\sum_{k=1}^{N}\frac{\nu_k^r}{n^r}\right) = \ln n + \frac{1}{r-1}\left(\ln n - \ln Z_{n,r}\right), \tag{11}$$

$$\hat{S}_r = \frac{1}{r-1}\left(1-\sum_{k=1}^{N}\frac{\nu_k^r}{n^r}\right) = \frac{1}{r-1}\left(1-\frac{Z_{n,r}}{n^r}\right). \tag{12}$$

Статистическая оценка энтропии Тсаллиса и её свойства

Поскольку оценка энтропии Тсаллиса (12) является линейным преобразованием статистики (8), справедлива следующая теорема об асимптотическом распределении вероятностей оценки (12).

Теорема 2. При истинной гипотезе H_* в асимптотике (4) статистика (12) имеет асимптотически нормальное распределение $\hat{S}_r \sim \mathrm{N}_1(\mu_{S,r}, \sigma^2_{S,r})$:

$$\mu_{S,r} = \frac{1}{r-1}\left(1-\frac{1}{N^{r-1}}\right), \tag{13}$$

$$\sigma^2_{S,r} = \frac{\lambda^{r-1}}{(r-1)^2 n^{2r-1}}\left(\sum_{i=1}^{r} s(r,i)\sum_{j=0}^{i-1} C_i^j r^{i-j}\sum_{k=1}^{j} S(j,k)\lambda^k - r^2\lambda^{r-1} + r!\right). \tag{14}$$

Следствие 2. При $r = 2$ для математического ожидания и дисперсии асимптотического распределения оценки (12) справедливы выражения:

$$\mu_{S,2} = 1-\frac{1}{N},\ \sigma^2_{S,2} = \frac{2N}{n^2}.$$

Покажем, что при истинной гипотезе H_* в асимптотике (4) статистика (12) является несмещённой и состоятельной оценкой энтропии Тсаллиса. Поскольку при истинной гипотезе H_* $p_k = 1/N$, $k = 1, \ldots, N$, то значение энтропии Тсаллиса равно

$$S_r(P) = \frac{1}{r-1}\left(1-\sum_{k=1}^{N} p_k^r\right) = \frac{1}{r-1}\left(1-\sum_{k=1}^{N}\frac{1}{N^r}\right) = \frac{1}{r-1}\left(1-\frac{1}{N^{r-1}}\right),$$

что совпадает с (13). В асимптотике (4) дисперсия оценки (14) стремится к 0, откуда с учётом несмещённости согласно [3] следует состоятельность оценки.

Статистическая оценка энтропии Реньи и её свойства

Оценка энтропии Реньи (11) является функцией от статистики (8). Из теоремы 4.2.5 в [7] вытекает справедливость следующей теоремы об асимптотическом распределении вероятностей оценки (11).

Теорема 3. При истинной гипотезе H_* в асимптотике (4) статистика (11) имеет асимптотически нормальное распределение $\hat{H}_r \sim \mathrm{N}_1(\mu_{H,r}, \sigma^2_{H,r})$:

$$\mu_{H,r} = \ln N, \tag{15}$$

$$\sigma^2_{H,r} = \frac{\sigma^2_{n,r}}{(r-1)^2 n^2 \lambda^{2r-2}}, \qquad (16)$$

где $\sigma^2_{n,r}$ – дисперсия статистики (8).

Следствие 3. При $r = 2$ для дисперсии асимптотического распределения оценки (11) справедливо выражение

$$\sigma^2_{H,2} = \frac{2}{n\lambda}.$$

Покажем, что при истинной гипотезе H_* в асимптотике (4) статистика (11) является несмещённой и состоятельной оценкой энтропии Реньи. При истинной гипотезе H_* $p_k = 1/N$, $k = 1, \ldots, N$, поэтому значение энтропии Реньи равно

$$H_r(P) = \frac{1}{1-r}\ln\left(\sum_{k=1}^{N} p_k^r\right) = \frac{1}{1-r}\ln\left(\sum_{k=1}^{N}\frac{1}{N^r}\right) = \frac{1}{1-r}\ln\frac{1}{N^{r-1}} = \frac{r-1}{r-1}\ln N = \ln N,$$

что совпадает с (15). В асимптотике (4) дисперсия оценки (16) стремится к 0, откуда с учётом несмещённости согласно [3] следует состоятельность оценки.

Проверка гипотезы о «чистой случайности» последовательности на основе оценок энтропии Реньи и Тсаллиса

Поскольку построенные точечные оценки функционалов энтропии Реньи и Тсаллиса являются состоятельными, и нам известно их асимптотическое распределение, мы можем построить на их основе интервальные оценки. Зададим уровень значимости $\varepsilon \in (0,1)$. Введём обозначения: $\hat{h}_r(n, N)$ – статистическая оценка энтропии Тсаллиса (12) (или Реньи (11)), μ – асимптотическое математическое ожидание статистической оценки энтропии Тсаллиса (13) (или Реньи (15)), σ^2 – асимптотическая дисперсия статистической оценки энтропии Тсаллиса (14) (или Реньи (16)) при истинной гипотезе H_*. Тогда с вероятностью $1 - \varepsilon$ $\hat{h}_r(n, N) \in (t_-, t_+)$, $t_\pm = \mu \pm \sigma\Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right)$, где $\Phi^{-1}(\cdot)$ – квантиль стандартного нормального закона [3].

На основе интервальной оценки построим решающее правило для проверки гипотез о том, является ли наблюдаемая последовательность генератора «чисто случайной»: H_* и $\overline{H_*}$. Вычислим для наблюдаемой последовательности статистику $\hat{h}_r(n, N)$. Решающее правило, основанное на статистике $\hat{h}_r(n, N)$, имеет вид:

$$\begin{cases} H_*, & \text{если } t_- < \hat{h}_r(n, N) < t_+; \\ \overline{H}_*, & \text{в противном случае,} \end{cases} \qquad t_\pm = \mu \pm \sigma\Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right).$$

В случае принятия решения о справедливости гипотезы H_* можно сделать вывод о том, что генератор пригоден для использования в криптосистемах,

поскольку по своим энтропийным свойствам он неотличим от чисто случайной последовательности на основе выборки объёма n.

Заключение

В данном докладе описан способ построения статистических оценок функционалов энтропии Реньи и Тсаллиса на основе частотных оценок вероятностей. Приведены асимптотические законы распределения и формулы моментов построенных оценок. На основе оценок энтропии построено решающее правило для проверки гипотезы о том, является ли наблюдаемая выходная последовательность криптографического генератора равномерно распределённой случайной последовательностью.

Список литературы

1. Bonachela, J.A. Entropy estimates of small data sets / J.A. Bonachela, H. Hinrichsen, M.A. Muñoz // Journal of Physics A: Mathematical and Theoretical. – 2008. – Vol. 41, № 20. –202001 (9 pp).
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. Jayadev, A. Estimating Renyi Entropy of Discrete Distributions [Electronic resource] / A. Jayadev, [et al.] – Mode of access: http://arxiv.org/pdf/1408.1000v2.pdf. – Date of access: 13.12.2015.
5. Энвин, А.Ю. Дискретная математика: Конспект лекций / А.Ю. Энвин. – Челябинск: Издательство ЮУрГУ, 1998. – 176 с.
6. Riordan, J. (1937). Moment recurrence relations for binomial, Poisson and hypergeometric frequency distributions / J. Riordan // Annals of Mathematical Statistics. – 1937. – Vol. 8, № 2. – P. 103–111.
7. Андерсон, Т. Введение в многомерный статистический анализ / Т. Андерсон. – Пер. с англ. Ю.Ф. Кичатова, Е.С. Кочеткова, Н.С. Райбмана. – Под ред. Б.В. Гнеденко. – Москва: Физматгиз, 1963 г. – 500 с.

Палуха Владимир Юрьевич, аспирант факультета прикладной математики и информатики Белорусского государственного университета, младший научный сотрудник Научно-исследовательского института прикладных проблем математики и информатики, palukha@bsu.by

Харин Юрий Семёнович, член-корреспондент Национальной академии наук Беларуси, директор Научно-исследовательского института прикладных проблем математики и информатики, заведующий кафедрой математического моделирования и анализа данных Белорусского государственного университета, доктор физико-математических наук, профессор, kharin@bsu.by