УДК 004.4: 519.23

ПРОГРАММНЫЙ КОМПЛЕКС ДЛЯ ЭНТРОПИЙНОГО АНАЛИЗА ДИСКРЕТНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

В.Ю. Палуха, Ю.С. Харин, М.В. Мальцев, А.И. Сергеев, А.А. Орлов

НИИ прикладных проблем математики и информатики БГУ
Пр-т. Независимости, 4, 220030, Минск, Беларусь
palukha@bsu.by, kharin@bsu.by, giftis95@mail.ru,
alex.orlov.official@gmail.com, maltsew@bsu.by

Рассматривается задача анализа дискретных последовательностей на основе оценок функционалов энтропии Шенннона, Реньи и Тсаллиса. Представлен разработанный программный комплекс.

***Ключевые слова*:** функционалы информационной энтропии; энтропия Шеннона; энтропия Реньи; энтропия Тсаллиса; статистические оценки.

SOFTWARE COMPLEX FOR ENTROPY ANALYSIS OF DISCRETE SEQUENCES

U.Yu. Palukha, Yu.S. Kharin, M.U. Maltsau, A.I. Siarheeu, A.A. Arlou

RI for Applied Problems of Mathematics and Informatics, BSU
4 Niezalieznasci Avenue, Minsk 220030, Belarus
Corresponding author: palukha@bsu.by

The problem of analyzing discrete sequences based on the estimates of the Shannnon, Renyi, and Tsallis entropy functionals is considered. The developed software package is presented.

Keywords: information entropy; Shannon entropy; Renyi entropy; Tsallis entropy; statistical estimators.

Введение

Стойкость систем криптографической защиты информации зависит от того, насколько близка используемая ими случайная или псевдослучайная последовательность по своим свойствам к равномерно распределённой случайной последовательности (РРСП) [1] (которая также называется «чисто случайной»), что устанавливается с помощью статистических тестов. В них проверяется гипотеза $H_* = \{\{x_t\}$ является РРСП$\}\}$. В данной работе в качестве тестовых статистик выступают статистические оценки

энтропии Шеннона, Реньи и Тсаллиса. Авторами разработан программный комплекс, который позволяет вычислять оценки указанных функционалов энтропии дискретной последовательности и на их основе принимать или отклонять гипотезу о «чистой случайности» анализируемой последовательности.

Энтропийный анализ

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 $p = \{p_k\}$, $p_k = P\{x = \omega_k\}$, $p_k \geq 0$, $\sum_{k=1}^{N} p_k = 1, k = 1, \ldots, N$. В таблице приведены формулы наиболее распространённых функционалов энтропии.

Таблица – Функционалы энтропии

Энтропия Шеннона	$H(p) = -\sum_{i=1}^{N} p_i \ln p_i$
Энтропия Реньи	$H_r(p) = \frac{1}{1-r} \ln\left(\sum_{i=1}^{N} p_i^r \right), r \in \square, r > 1.$
Энтропия Тсаллиса	$S_r(p) = \frac{1}{r-1}\left(1 - \sum_{i=1}^{N} p_i^r \right), r \in \square, r > 1.$

Пусть наблюдается реализация случайной последовательности $\{x_t : t = 1, \ldots, n\}$ длины n из распределения вероятностей $\{p_k\}$, по которой будет оцениваться энтропия. Частотные оценки вероятностей имеют вид

$$\hat{p}_k = \frac{v_k}{n}, \quad v_k = \sum_{t=1}^{n} I\{x_t = \omega_k\} \in \square_0 = \square \cup \{0\}, \quad I\{x_t = \omega_k\} = \begin{cases} 1, x_t = \omega_k; \\ 0, x_t \neq \omega_k. \end{cases} \quad (1)$$

Рассмотрим асимптотику

$$n, N \to \infty, n/N \to \lambda, 0 < \lambda < \infty. \quad (2)$$

которая отличается от классической ($n \to \infty$, $N < \infty$) тем, что длина последовательности n и мощность алфавита N растут синхронно.

Оценка энтропии Шеннона на основе статистик (1) имеет вид:

$$\hat{H} = \hat{H}(n, N) = -\sum_{k=1}^{N} \hat{p}_k \ln \hat{p}_k = -\sum_{k=1}^{N} \frac{v_k}{n} \ln \frac{v_k}{n} = \ln n - \frac{1}{n} \sum_{k=1}^{N} v_k \ln v_k. \quad (3)$$

Теорема 1 [2]**.** В асимптотике (2) статистика (3) при гипотезе H_* имеет асимптотически нормальное распределение с параметрами

$$\mu_H = \ln n - e^{-\lambda}\sum_{k=1}^{+\infty}\frac{\ln(k+1)\lambda^k}{k!}, \tag{4}$$

$$\sigma_H^2 = \frac{e^{-\lambda}}{n}\sum_{k=1}^{+\infty}\frac{(k+1)\lambda^k}{k!}\ln^2(k+1) - \frac{e^{-2\lambda}}{N}\left(\sum_{k=1}^{+\infty}\frac{\ln(k+1)\lambda^k}{k!}\right)^2 - \\ -\frac{e^{-2\lambda}}{n}\left(\sum_{k=1}^{+\infty}\ln(k+1)\frac{\lambda^k}{k!}(k+1-\lambda)\right)^2. \tag{5}$$

Из теоремы 1 видно, что в асимптотике (2) оценка (3) является смещённой. Для функционалов энтропии Реньи и Тсаллиса можно построить несмещённую оценку в асимптотике (2), в т.ч. и при $\lambda < 1$.

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; при $x < r$ полагают $x^{\underline{r}} ::= 0$.

Статистические оценки энтропии Реньи и Тсаллиса, построенные с использованием нисходящей факториальной степени, имеют вид

$$\hat{H}_r(n, N) = \frac{1}{1-r}\ln\left(\sum_{k=1}^{N}\frac{\nu_k^{\underline{r}}}{n^r}\right) = \ln n + \frac{1}{r-1}\left(\ln n - \ln\sum_{k=1}^{N}\nu_k^{\underline{r}}\right), \tag{7}$$

$$\hat{S}_r(n, N) = \frac{1}{r-1}\left(1 - \sum_{k=1}^{N}\frac{\nu_k^{\underline{r}}}{n^r}\right) = \frac{1}{r-1}\left(1 - \frac{1}{n^r}\sum_{k=1}^{N}\nu_k^r\right). \tag{8}$$

Теорема 2 [2]**.** В асимптотике (2) статистика (8) является состоятельной асимптотически несмещённой оценкой энтропии Тсаллиса и при истинной гипотезе H_* имеет асимптотически нормальное распределение с параметрами:

$$\mu_{S,r} = \frac{1}{r-1}\left(1 - \frac{1}{N^{r-1}}\right), \tag{9}$$

$$\sigma_{S,r}^2 = \frac{\lambda^{r-1}}{(r-1)^2 n^{2r-1}}\left(\sum_{i=1}^{r} s(r,i)\sum_{j=1}^{i-1} C_i^j r^{i-j}\sum_{k=1}^{j} S(j,k)\lambda^k - r^2\lambda^{r-1} + r!\right), \tag{10}$$

где $S(r, i)$ – число Стирлинга второго рода.

Следствие 1. При $r = 2$ для математического ожидания и дисперсии асимптотического распределения оценки (8) справедливы выражения:

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

Теорема 3 [2]**.** В асимптотике (2) статистика (7) является состоятельной оценкой энтропии Реньи и при истинной гипотезе H_* имеет асимптотически нормальное распределение с параметрами:

$$\mu_{H,r} = \ln N, \tag{11}$$

$$\sigma^2_{H,r} = \frac{\sum_{i=2}^{r} s(r,i) \sum_{j=1}^{i-1} C_i^j r^{i-j} \sum_{k=1}^{j} S(j,k) \lambda^k - r^2 \lambda^{r-1} + r!}{(r-1)^2 n \lambda^{r-1}}. \tag{12}$$

Следствие 2. При $r = 2$ для дисперсии асимптотического распределения вероятностей оценки (7) справедливо выражение:

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

Пусть $\alpha \in (0,1)$ – заданный уровень значимости. Введём обозначения: $\hat{h}$ – статистическая оценка энтропии Шеннона (3), Реньи (7) или Тсаллиса (8), μ_h – асимптотическое математическое ожидание статистической оценки энтропии Шеннона (4), Реньи (11) или Тсаллиса (9), σ_h^2 – асимптотическая дисперсия статистической оценки энтропии Шеннона (5), Реньи (12) или Тсаллиса (10) при истинной гипотезе H_*. Решающее правило имеет вид [2]:

$$\text{принимается} \begin{cases} H_*, \text{если} \quad t_- < \hat{h} <; t_+ \\ \overline{H_*}, \text{в противном случае,} \end{cases} \qquad t_\pm = \mu_h \pm \sigma_h \Phi^{-1}\left(1 - \frac{\alpha}{2}\right), \tag{13}$$

где $\Phi(\cdot)$ – функция распределения стандартного нормального закона.

Вычислим нормированную статистику

$$\tilde{h} = \frac{\hat{h} - \mu_h}{\sigma_h}.$$

Она в асимптотике (2) имеет стандартное нормальное распределение: $\tilde{h} \sim \mathcal{N}(0,1)$. Следовательно, двустороннее *p*-значение для неё равно

$$p - value = 2\left(1 - \Phi\left(\left|\tilde{h}\right|\right)\right). \tag{14}$$

Пусть генератор порождает двоичную выходную последовательность $\{y_\tau\}$, $\tau = 1, \ldots, T$. «Нарежем» её на непересекающиеся подряд идущие 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 сформируем новую последовательность $\{x_t\}$ из алфавита мощности $N = 2^s$ по правилу $x_t = \sum_{j=1}^{s} 2^{j-1} X_j^{(t)} + 1$.

На основе критерия (13) мы можем вычислить последовательность нормированных отклонений оценки энтропии от математического ожидания в зависимости от s, которые назовём **энтропийными профилями**:

$$\chi(s) = \frac{\hat{h}(s) - \mu_h(s)}{\sigma_h(s)\Phi^{-1}(1-\alpha/2)} = \frac{\tilde{h}(s)}{\Phi^{-1}(1-\alpha/2)}, \; s = s_-, \ldots, s_+. \qquad (15)$$

Программный комплекс

В НИИ ППМИ разработан программный комплекс (ПК), который позволяет вычислять оценки энтропии Шеннона (3), Реньи (7) и Тсаллиса (8) при $r = 2$, их асимптотические параметры распределений вероятностей при гипотезе H_* с помощью алгоритмов [3], p-значения (14) и энтропийные профили (15) для двоичных файлов. Помимо вывода самих значений, программа выводит графики зависимостей этих величин от длины фрагмента s.

В начале работы необходимо выбрать файл с последовательностью, диапазон $s_-, \ldots, s_+$ и функционалы энтропии. Вычисляемые значения добавляются на экран в режиме реального времени. Имеется возможность изменять уровень значимости $\alpha \in (0,1)$ без пересчёта оценок энтропии и переключаться на различные режимы отображения: непосредственно оценки энтропии $\hat{h}$, нормированные значения (15), p-значения (14).

Для тестирования ПК подготовлена библиотека последовательностей псевдослучайных и физических генераторов. На рисунке 1 представлен результат работы ПК с последовательностью физического генератора [4], на рисунке 2 – с последовательностью, полученной при помощи регистра сдвига с линейной обратной связью (РСЛОС) с примитивным характеристическим многочленом над полем GF(2) $x^{32} + x^{22} + x^2 + x + 1$ [1] на уровне значимости $\alpha = 0.05$. Как видно из рисунков, для физического генератора гипотеза H_* принимается, для РСЛОС начиная с $s = 16$ отклоняется.

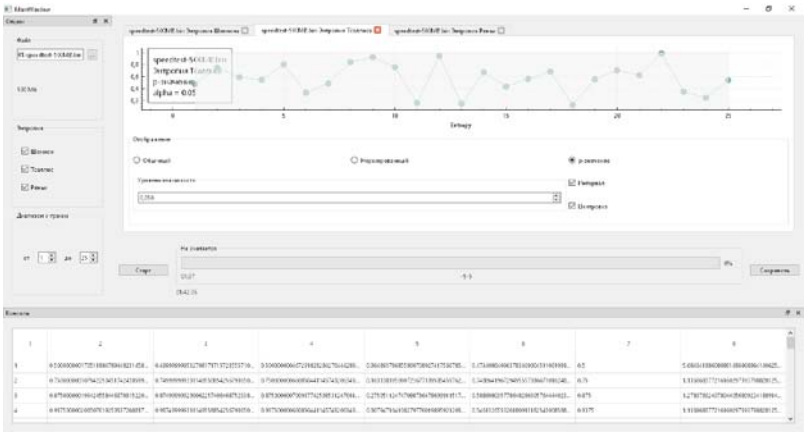

Рисунок 1 – Энтропийный профиль Тсаллиса физического генератора

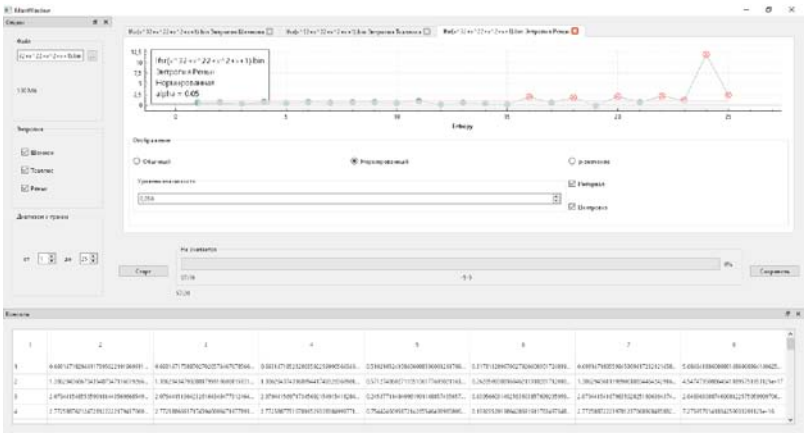

Рисунок 2 – Энтропийный профиль Реньи РСЛОС

Библиографические ссылки

1. Харин Ю.С., Агиевич С.В., Васильев Д.В., Матвеев Г.В. Криптология. Минск: БГУ, 2013. 512 с.
2. Палуха В. Ю. Статистические тесты на основе оценок энтропии для проверки гипотез о равномерном распределении случайной последовательности // Весці НАН Беларусі. Серыя фізіка-матэматычных навук. 2017. № 1. С.: 79–88.
3. Палуха В.Ю., Харин Ю.С. Вычисление статистических оценок функционалов энтропии двоичных последовательностей // Международный конгресс по информатике: информационные системы и технологии [Электронный ресурс]: Материалы международного научного конгресса. Республика Беларусь, Минск, 24–27 октября 2016 года. Минск: БГУ, 2016. С. 472–476.
4. Физический генератор. URL: http://qrng.physik.hu-berlin.de/files/speedtest-500MB.bin.