УДК 519.671:003.26

ЗАЩИТА ИНФОРМАЦИИ И СТОХАСТИКА

Ю.С. Харин[1, 2]

[1]*Научно-исследовательский институт прикладных проблем математики и информатики,*
[2]*Белорусский государственный университет,*
пр. Независимости, 4, 220030, г. Минск, Беларусь, kharin@bsu.by

Рассматривается проблема статистического тестирования дискретно-значных временных рядов на «чистую случайность». Предложены модели отклонений от «чистой случайности», методы и алгоритмы обнаружения этих отклонений. Приводятся результаты компьютерных экспериментов.

Ключевые слова: временной ряд; цепь Маркова; статистический тест.

INFORMATION PROTECTION AND STOCHASTICS

Yu.S. Kharin[a,b]

[a]*Research Institute for Applied Problems of Mathematics and Informatics,*
[b]*Belarusian State University, 4 Nezavisimosti avenue, Minsk 220030, Belarus,*
kharin@bsu.by

Problem of statistical testing for “pure randomness” of discrete-valued time series is considered. Models of deviations from “pure randomness”, methods and algorithms for detection of these deviations are proposed. Results of computer experiments are given.

Keywords: time series; Markov chain; statistical test.

Введение

В современных системах обеспечения информационной и компьютерной безопасности важнейшим способом защиты информации является криптографический способ, позволяющий с гарантированной стойкостью решить главные практические задачи: 1) конфиденциальность; 2) аутентификация источника сообщения; 3) проверка целостности; 4) невозможность отречения от авторства. Криптографический способ базируется на новой науке Криптологии [1], объединяющей Криптографию и Криптоанализ. Криптология и Стохастика тесно связаны: Стохастика представляет математический инструментарий для решения задач Криптологии, Криптология стимулирует Стохастику к разработке новых моделей для исследования сложных последовательностей, циркулирующих в криптосистемах.

Настоящий доклад посвящен представлению и использованию новых стохастических моделей в криптографической защите информации.

1. Проблема «чистой случайности» в защите информации

Многие задачи криптологии (статистическое тестирование криптографических генераторов, статистический криптоанализ, разностный криптоанализ, линейный криптоанализ, криптоатаки по побочным каналам, стеганография) сводятся к задаче различения некоторой зарегистрированной последовательности символов $x_1, x_2, \ldots$ от «чисто случайной» и оценки величины этого различия.

Математической моделью последовательностей, порождаемых генераторами, а также последовательностей, возникающих в различных узлах средств криптографической защиты информации, является дискретный временной ряд (ДВР). ДВР – это случайный процесс $x_t \in A$ на вероятностном пространстве $(\Omega, F, \mathbf{P})$ с дискретным временем $t \in \mathbf{N} = \{1, 2, \ldots\}$ и дискретным множеством состояний (алфавитом) $A = \{0, 1, \ldots, N-1\}$ мощности $|A| = N$, $2 \le N < +\infty$.

В криптологии [1] согласно Шенноновской теории совершенных криптосистем большое внимание уделяется так называемому «чисто случайному» ДВР – равномерно распределенной случайной последовательности (РРСП) $x_1, x_2, \ldots \in A$, обладающей двумя свойствами:

C_1) для любого числа $n \in \mathbf{N}$ и произвольных индексов $1 < t_1 < \ldots < t_n$ случайные элементы $x_{t_1}, \ldots, x_{t_n}$ независимы в совокупности;

C_2) для любого $t \in \mathbf{N}$ случайная величина x_t имеет равномерное на A распределение вероятностей: $\mathbf{P}\{x_t = i\} = N^{-1}$, $i \in A$.

В настоящее время известно более сотни методов генерации последовательностей, по своим свойствам приближающихся к РРСП. Еще больше разработано статистических тестов криптографических генераторов, заключающихся в проверке простой гипотезы $H_0 = \{\{x_t\}$ есть РРСП$\}$ против сложной альтернативы $H_1 = \overline{H}_0 = \{$нарушены свойства $C_1 \cup C_2\}$. Обзор статистических тестов показывает:

1) многие из существующих тестов не ориентированы на проверку главного свойства C_1, а лишь частных случаев свойств C_1, C_2, т.е. частных случаев альтернативы H_1;

2) многие тесты построены «эвристически» и не фиксируют H_1;

3) многие тесты не имеют оценок мощности;

4) при включении нескольких тестов в батарею не удается оптимизировать «составной» тест.

В связи с этим актуальна проблема разработки адекватных стохастических моделей отклонений H_1 от модели РРСП и построения тестов для обнаружения и оценивания таких отклонений.

2. Модели ДВР на основе семейства отклонений от s-мерной равномерности и их энтропийное тестирование

Определим вложенное в H_1 семейство «альтернатив s-мерной неравномерности»: $H_{1(s)} = \{\{x_1, x_2, \ldots\} = \{X_1, X_2, \ldots\}\} \subset H_1$, где $X_1, X_2, \ldots \in A^s$ – независимые одинаково распределенные s-фрагменты (слова) над алфавитом A с некоторым s-мерным дискретным распределением вероятностей $\mathbf{P}_{i_1,\ldots,i_s} = \mathbf{P}\{x_1 = i_1, \ldots, x_s = i_s\}$, $i_1, \ldots, i_s \in A$, отличным от равномерного: $\Delta_s = \sum_{i_1,\ldots,i_s \in A} \left| \mathbf{P}_{i_1,\ldots,i_s} - N^{-s} \right| > 0$, $\sum_{i_1,\ldots,i_s \in A} \mathbf{P}_{i_1,\ldots,i_s} \equiv 1$. Это семейство моделей ДВР обладает двумя свойствами: 1) при $s \to \infty$ семейство этих альтернатив имеет в пределе альтернативу $H_1 = \bar{H}_0$ общего вида; 2) чем меньше Δ_s, тем ближе альтернатива $H_{1(s)}$ к H_0.

Обозначим: $\{x_1, x_2, \ldots, x_T\} = \{X_1, X_2, \ldots, X_M\}$ – наблюдаемая реализация выходной последовательности генератора длиной $T = M \cdot s$, разбитая на M непересекающихся фрагментов длины s; $I\{B\}$ – индикатор события B; статистическая оценка для $\mathbf{P}_{i_1,\ldots,i_s}$

$$\hat{\mathbf{P}}_{i_1,\ldots,i_s} = \frac{1}{M} \sum_{m=1}^{M} I\{X_m = (i_1, \ldots, i_s)\}, \; i_1, \ldots, i_s \in A. \qquad (1)$$

Тест обобщенного отношения правдоподобия для проверки H_0, $H_{1(s)}$ на основе статистик (1) имеет вид:

$$\text{принимается} \begin{cases} H_0, \text{ если } \hat{H}_s - s \ln N > -\dfrac{1}{2M} G^{-1}_{N^s - 1}(1 - \varepsilon), \\ H_{1(s)} \text{ в противном случае,} \end{cases} \qquad (2)$$

$\hat{H}_s = -\sum_{i_1,\ldots,i_s \in A} \hat{P}_{i_1,\ldots,i_s} \ln \hat{P}_{i_1,\ldots,i_s}$ – статистическая оценка s-мерной энтропии Шеннона, $G_K^{-1}(\cdot)$ – обратная функция распределения хи-квадрат с K степенями свободы, $\varepsilon \in (0,1)$ – заданный уровень значимости теста.

Тест (1), (2) мы предлагаем использовать для визуализации процесса принятия решений в виде так называемого «энтропийного профиля (портрета)» – графика зависимости нормированного отклонения оценки s-мерной энтропии от ее математического ожидания при H_0 (см. рис. 1, 2 для $N = 2,\ \varepsilon = 0.05,$ где штриховые линии – границы области решений):

$$\alpha(s) = 2M\left(\hat{H}_s - s\ln N\right) / G_{N^s-1}^{-1}(1-\varepsilon),\quad s \in \{s_{min}, s_{min}+1, \ldots, s_{max}\}\ . \qquad (3)$$

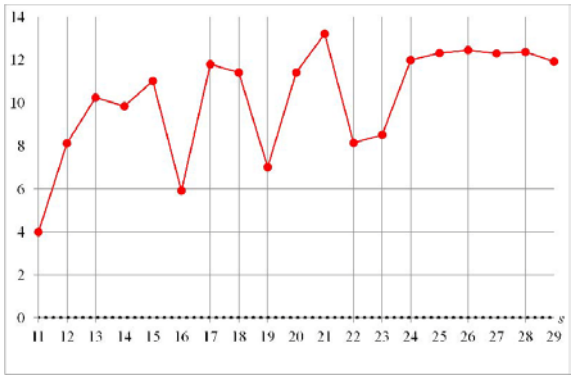

Рисунок 1 – Энтропийный профиль $\ln|\alpha(s)|$ нелинейного регистра сдвига порядка 24 ($T = 2^{32}/s$)

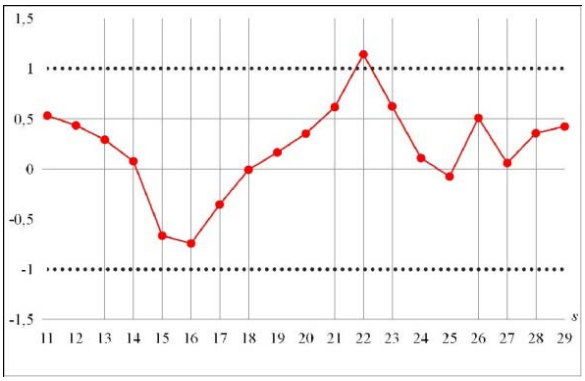

Рисунок 2 – Энтропийный профиль $\alpha(s)$ генератора BelT (СТБ 34.101.27-2011, $T = 2^{29}/s$)

Отметим еще, что вместо энтропии Шеннона в (1) – (3) могут использоваться энтропийные функционалы Реньи и Тсаллиса [2].

3. Модели ДВР на основе марковских зависимостей высокого порядка и их статистическое тестирование

3.1. Тестирование на основе цепи Маркова высокого порядка

Учитывая, что универсальной моделью стохастической зависимости элементов выходной последовательности $\{x_t\}$ криптографического генератора является цепь Маркова достаточно высокого порядка s, определим вложенное в $H_1 = \overline{H}_0$ семейство альтернатив марковской зависимо-

сти: $H_1^{(s)} = \{\{x_t\}$ – однородная цепь Маркова порядка s с $(s+1)$-матрицей переходов $\mathbf{P}\}$, где $\mathbf{P} = \left(p_{i_1,\ldots,i_{s+1}}\right)$, $i_1,\ldots,i_{s+1} \in A$,

$$p_{i_1,\ldots,i_{s+1}} = \mathbf{P}\{x_{t+1} = i_{s+1} \mid x_t = i_s,\ldots,x_{t-s+1} = i_1\}, \Delta_s = \sum_{i_1,\ldots,i_S \in A} \left| p_{i_1,\ldots,i_{s+1}} - N^{-1} \right| > 0. \qquad (4)$$

Тест обобщенного отношения правдоподобия для проверки гипотез $H_0, H_1^{(s)}$ основан на оценке $\hat{h}_s$ условной энтропии $h_s = H\{x_t \mid x_{t-1},\ldots,x_{t-s}\}$:

$$\text{принимается} \begin{cases} H_0, \text{если } \hat{h}_s - \ln N > -G_f^{-1}(1-\varepsilon)/\left(2(T-s)\right),\ f = N^s(N-1), \\ H_1^{(s)} \text{ в противном случае.} \end{cases} \qquad (5)$$

Аналогично (3) с помощью $\hat{h}_s$ строится энтропийный профиль.

Тесты (2), (5), анализирующие стохастические зависимости глубины s в выходной последовательности $\{x_t\}$, требуют экспоненциально растущей с ростом порядка s длины анализируемой последовательности $T = O\left(N^{s+1}\right)$. Для преодоления этой трудности целесообразно использовать «малопараметрические модели цепей Маркова высокого порядка» [1, 3], т.е. модели цепей Маркова s-го порядка, для которых $\left(N^s \times N\right)$-матрица вероятностей переходов зависит от «малого» числа параметров $D << N^s(N-1)$; $\kappa = D/\left(N^s(N-1)\right) << 1$ – коэффициент сжатия, равный относительному числу параметров модели.

3.2. Построение малопараметрических цепей Маркова

Подход I: «сжатие множества значений элементов матрицы» $\mathbf{P}$.

Пусть $Q = \left(q_{j_i,\ldots,j_r,j_{r+1}}\right)$ – некоторая $(r+1)$-мерная матрица, $1 \le r < s$, $\sum_{j_{r+1} \in A} q_{j_i,\ldots,j_r,j_{r+1}} \equiv 1, 0 \le q_{j_i,\ldots,j_r,j_{r+1}} \le 1$; $B(\cdot)\colon A^s \to A^r$ – некоторая дискретная функция. С помощью $B(\cdot)$ $(s+1)$-мерная матрица $\mathbf{P}$ «сжимается» в $(r+1)$-мерную матрицу Q:

$$p_{i_1,\ldots,i_s,i_{s+1}} = q_{B(i_1,\ldots,i_s),i_{s+1}};\ \kappa_{\mathrm{I}} = N^{r-s} \le 1. \qquad (6)$$

Примеры малопараметрических ДВР: $\mathrm{MC}(s,r)$, $\mathrm{MCCO}(s,L)$, VLMC [4].

Подход II. Этот подход заключается в использовании порождающего уравнения для условного распределения вероятностей (4) будущего состояния $x_t \in A$ при условии предыстории $X_{t-s}^{t-1} = \left(x_{t-1},\ldots,x_{t-s}\right)' \in A^s$:

$$p_{i_1,\ldots,i_s,i_{s+1}} = q_{i_{s+1}}\left(\theta\left(i_1,\ldots,i_s;a\right)\right),\ i_1,\ldots,i_{s+1} \in A, \qquad (7)$$

где $\left\{q_j(\theta): j \in A\right\}$ – некоторое стандартное вероятностное распределение на A, зависящее от параметра $\theta = \left(\theta_j\right) \in \Theta \subseteq R^L$; $\theta = \theta\left(i_1,\ldots,i_s;a\right)$ – некоторая функция, известная с точностью до вектора параметров $a = \left(a_k\right) \in R^m$. Коэффициент сжатия: $\kappa_{\text{II}} = m/\left(N^s\left(N-1\right)\right) \le 1$.

Примеры малопараметрических ДВР: модель Джекобса – Льюиса, MTD-модель, $\text{DAR}(s)$, $\text{BCNAR}(s)$, $\text{BiCNAR}(s)$, $\text{PCNAR}(s)$.

4. Малопараметрические модели ДВР на основе подхода I и их статистическое тестирование

4.1. Цепь Маркова $\text{MC}(s,r)$ порядка s с r частичными связями

Эта модель определяется (6) с $B(j_1\ldots,j_s) = \left(j_{m_1^0},\ldots,j_{m_r^0}\right)$ [1, 3]:

$$p_{J_1^{s+1}} = p_{j_1,\ldots,j_s,j_{s+1}} = q_{j_{m_1^0},\ldots,j_{m_r^0},j_{s+1}},\ J_1^{s+1} \in A^{s+1}, \qquad (8)$$

где $J_i^k = \left(j_i, j_{i+1},\ldots,j_k\right) \in A^{k-i+1}$ – последовательность $k-i+1$ индексов $\left(k \ge i\right)$; r – число связей; $M_r^0 = \left(m_1^0,\ldots,m_r^0\right)$ – вектор с r упорядоченными компонентами $1 = m_1^0 < m_2^0 < \ldots < m_r^0 \le s$, называемый шаблоном связей; $Q = \left(q_{J_1^{r+1}}\right)_{J_1^{r+1} \in A^{r+1}}$ – $\left(r+1\right)$-мерная стохастическая матрица.

Статистическую оценку $\hat{Q}$ удобно использовать для визуализации отклонения от гипотезы H_0 (для которой $q_{i_1,\ldots,i_{r+1}} = N^{-1}$). На рис. 3, 4 представлены результаты такой визуализации для генератора со случайной обратной связью и генератора BelT (СТБ 34.101.27-2011 в режиме гаммирования) соответственно; здесь красный цвет – оценка условной вероятности перехода в «0» $\hat{q}_{K_1^r,0}$, зеленый – в «1» $\hat{q}_{K_1^r,1}$); по оси абсцисс откладывается $K_1^r = B\left(J_1^s;\hat{M}_r\right) \in A^r$.

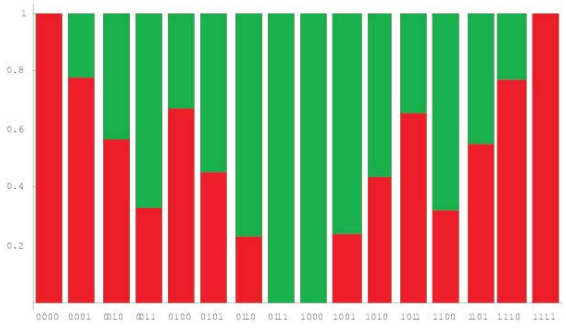

Рисунок 3 – Оценка $\hat{Q}$
$\left(s=64,\ r=4,\ T=10^5\right)$

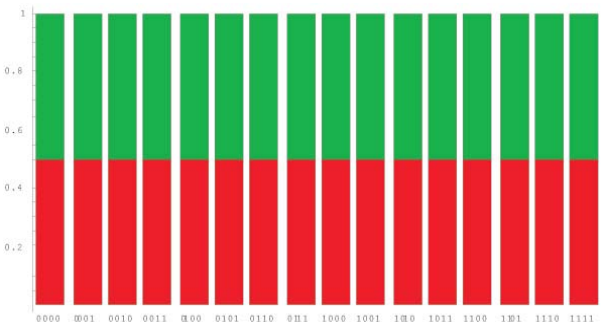

Рисунок 4 – Оценка $\hat{Q}$
$\left(s=32,\ r=4,\ T=8\cdot10^6\right)$

4.2. Модель Джекобса – Льюиса

Эта модель порождается стохастическим разностным уравнением [5]:

$$x_t=\mu_t x_{t-\eta_t}+\left(1-\mu_t\right)\xi_t, \tag{9}$$

где $t>s$, $\{\xi_t,\eta_t,\mu_t\}$ – независимые в совокупности случайные величины с вероятностными распределениями:

$$\mathbf{P}\{\mu_t=1\}=1-\mathbf{P}\{\mu_t=0\}=\rho;\ \mathbf{P}\{\xi_t=k\}=\pi_k\ ,\ k\in A\ ,\ \sum_{k\in A}\pi_k=1\ ;$$

$$\mathbf{P}\{\eta_t=i\}=\lambda_i,\ i\in\{1,2,\ldots,s\},\ \sum_{i=1}^{s}\lambda_i=1,\ \lambda_s\neq0; \tag{10}$$

$$\mathbf{P}\{x_1=k\}=\ldots=\mathbf{P}\{x_s=k\}=\pi_k,\ k\in A.$$

Число параметров этой модели (9), (10) линейно зависит от s, так что коэффициент сжатия $\kappa_{JL}=\left(N+s-1\right)/\left(N^s\left(N-1\right)\right)$. Методы и алгоритмы статистического анализа этой модели представлены в [1].

4.3. MTD-модель Рафтери

MTD (Mixture Transition Distribution)-модель [6] определяется следующим частным случаем уравнения (7): $p_{i_1,\ldots,i_s,i_{s+1}}=\sum_{j=1}^{s}\lambda_j q_{i_j,i_{s+1}},i_1,\ldots,i_{s+1}\in A,$ где $Q=\left(q_{i,k}\right)$ – некоторая стохастическая $\left(N\times N\right)$-матрица, $0\le q_{i,k}\le1$, $\sum_{k\in A}q_{i,k}\equiv1$, $i,k\in A$, $\lambda=\left(\lambda_1,\ldots,\lambda_s\right)'$ – некоторое дискретное распределение вероятностей, $\lambda_1>0$.

Обобщенная MTDg (generalized MTD)-модель определяется параметризацией $(s+1)$-мерной матрицы $\mathbf{P}$: $p_{i_1,\ldots,i_s,i_{s+1}} = \sum_{j=1}^{s} \lambda_j q^{(j)}_{i_j,i_{s+1}}$, $i_1,\ldots,i_{s+1} \in A$, где $Q^{(j)} = \left(q^{(j)}_{i,k}\right)$ – некоторая стохастическая матрица для j-го лага, $\kappa_{\mathrm{MTDg}} = \left(s\left(N(N-1)/2+1\right)-1\right)/\left(N^s(N-1)\right)$. Методы и алгоритмы статистического тестирования даны в [1].

4.4. Биномиальная условно нелинейная авторегрессионная модель BiCNAR(s)

Эта модель порождается биномиальным случаем уравнения (7):

$$p_{i_1,\ldots,i_s,i_{s+1}} = C^{i_{s+1}}_{N-1}\, \theta^{i_{s+1}} (1-\theta)^{N-1-i_{s+1}}, i_{s+1} \in A = \{0,1,\ldots,N-1\},$$

$$\theta = \theta\left(\mathrm{I}^{\mathrm{s}}_1\right) = \mathrm{F}\left(a'\Psi\left(\mathrm{I}^{\mathrm{s}}_1\right)\right),\ \mathrm{I}^{\mathrm{s}}_1 = (i_1,\ldots,i_s)' \in A^s,$$

где $\Psi\left(\mathrm{I}^{\mathrm{s}}_1\right) = \left(\psi_1\left(\mathrm{I}^{\mathrm{s}}_1\right),\ldots,\psi_m\left(\mathrm{I}^{\mathrm{s}}_1\right)\right)'$: $A^s \to R^m$ – вектор-столбец $m \le N^s$ линейно независимых функций, например, полиномов; $\mathrm{F}(\cdot)$: $R^1 \to [0,1]$ – некоторая функция распределения, например, логистическая, нормальная или Коши; $a = (a_1,\ldots,a_m)'$ – вектор-столбец m неизвестных параметров модели. Относительное число параметров модели: $\kappa = m\left(N^s(N-1)\right)^{-1} < 1$.

Методы и алгоритмы статистического анализа BiCNAR(s)-модели, ее частных случаев и обобщений представлены в [7, 8].

Заключение

1. В криптологии актуальна проблема построения и статистического анализа моделей ДВР, адекватно описывающих отклонения от РРСП.
2. Представлены семейства моделей ДВР на основе отклонений от s-мерной равномерности и на основе марковских зависимостей порядка s.
3. Для преодоления «проклятия размерности» представлены подходы к построению малопараметрических цепей Маркова порядка s.
4. Разработаны методы и алгоритмы статистического оценивания параметров и проверка гипотез H_0, H_1 для малопараметрических моделей, построенных на основе предложенных подходов.

5. Теоретические результаты иллюстрируются результатами компьютерных экспериментов по тестированию выходных последовательностей известных криптографических генераторов.

Библиографические ссылки

1. Харин Ю.С., Агиевич С.В., Васильев Д.В., Матвеев Г.В. Криптология. Минск: БГУ, 2014. 512 с.
2. Харин Ю.С., Палуха В.Ю. Энтропийный анализ криптографических генераторов случайных и псевдослучайных последовательностей // Веснік сувязі. 2017. № 146(1). С. 46–49.
3. Харин Ю.С. Цепи Маркова с S-частичными связями и их статистическое оценивание // Доклады НАН Беларуси. 2004. № 48(1). С. 40–44.
4. Buhlmann P., Wyner A.J. Variable length Markov chains // The Annals of Statistics. 1999. № 27(2). P. 480–513.
5. Jacobs P.A., Lewis P.A.W. Discrete time series generated by mixtures I: correlational and runs properties // Journal of the Royal Statistical Society. Ser. B. 1978. № 40(1). P. 94–105.
6. Raftery A. A model for high-order Markov chains // Journal of the Royal Statistical Society. Ser. B. 1985. № 47(3). P. 528–539.
7. Харин Ю.С., Волошко В.А. Биномиальные условно нелинейные авторегрессионные модели дискретных временных рядов и их вероятностные и статистические свойства // Труды Института математики НАН Беларуси. 2019. № 26(1). С. 95–105.
8. Kharin Yu.S., Voloshko V.A., Medved E.A. Statistical estimation of parameters for binary conditionally nonlinear autoregressive time series // Mathematical Methods of Statistics. 2019. № 26(2). P. 103–118.