ИНФОРМАТИКА

2013 **июль-сентябрь** **№ 3**

ЗАЩИТА ИНФОРМАЦИИ

УДК 004.421.6: 519.237

Ю.С. Харин, В.Ю. Палуха

ИНФОРМАТИВНЫЕ ПРИЗНАКИ ДЛЯ СТАТИСТИЧЕСКОГО РАСПОЗНАВАНИЯ КРИПТОГРАФИЧЕСКИХ ГЕНЕРАТОРОВ

Разрабатывается общий подход к построению информативных признаков для решения задачи статистического распознавания криптографических генераторов. Описываются признаки, построенные в соответствии с этим подходом. Приводятся результаты компьютерных экспериментов применения построенных информативных признаков для статистического распознавания генераторов.

Введение

Современная криптография невозможна без случайных и псевдослучайных последовательностей $x_1, x_2, \ldots \in V = \{0,1\}$ [1]. Случайные последовательности генерируются при помощи физических генераторов, а псевдослучайные – при помощи программных генераторов. Практическую значимость имеют генераторы последовательностей, близких по своим свойствам к равномерно распределенной случайной последовательности (РРСП). РРСП – это случайная последовательность x_1, x_2, …, x_t, x_{t+1}, …, определенная на вероятностном пространстве $(\Omega, \mathcal{F}, P)$ и удовлетворяющая двум требованиям [1]:

C_1: Для любого $n \in \mathbb{N}$ и произвольных значений индексов $1 \le t_1 < \ldots < t_n$ случайные величины $x_{t_1}, \ldots, x_{t_n}$ независимы в совокупности.

C_2: Для любого номера $t \in \mathbb{N}$ случайная величина x_t является бернуллиевой и имеет равномерное распределение вероятностей $P\{x_t = i\} = \frac{1}{2}, \quad i \in V = \{0,1\}$.

Гипотезу о том, что выходная последовательность генератора $\{x_t\}$ является равномерно распределенной, будем обозначать H_* = {$\{x_t\}$ есть РРСП}, а альтернативу – $H = \overline{H_*}$.

При проведении испытаний средств криптографической защиты информации [1, 2] с целью оценки их надежности возникают следующие задачи: определить, какой тип генератора случайной или псевдослучайной последовательности использовался; обнаружить, не ухудшились ли криптографические свойства генератора с течением времени. Математической сущностью этих практических задач является задача статистического распознавания генераторов случайных и псевдослучайных последовательностей, т. е. задача отнесения (классификации) наблюдаемой выходной последовательности генератора $x_1, x_2, \ldots, x_T \in V = \{0,1\}$ некоторой конечной длительности T к одному из L $(2 \le L < +\infty)$ классов Ω_1, …, Ω_L.

Генераторы могут быть разделены на следующие классы $\{\Omega_i\}$: физические и программные генераторы; программные генераторы различных типов; программные генераторы одного типа, но с различными параметрами; программные генераторы с одинаковыми параметрами, но с различной инициализирующей информацией. Будем придерживаться классификации криптографических генераторов [2], представленной на рис. 1.

Приведем описания прореживающего и самосжимающего генераторов, которые в статье будут использоваться в компьютерных экспериментах.

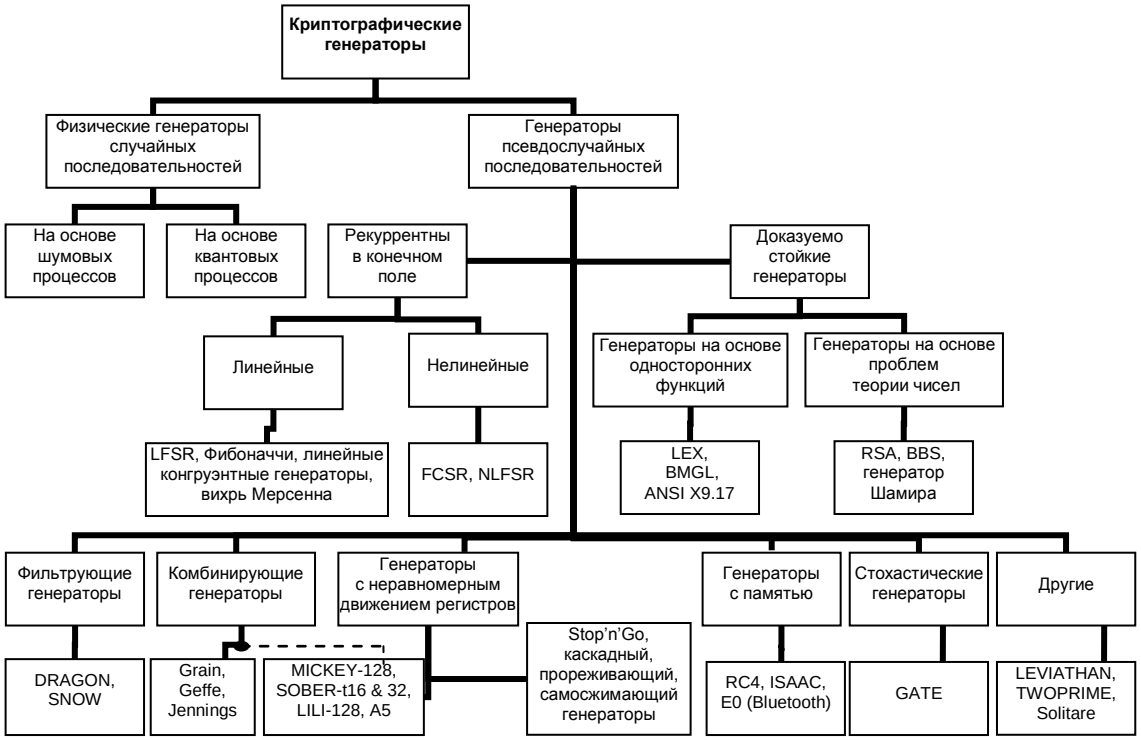

Рис. 1. Схема классификации криптографических генераторов

Пусть LFSR-генератор (регистр сдвига с линейной обратной связью) с порождающим многочленом степени L_1 генерирует «элементарную» двоичную последовательность $\{a_t\}$, а LFSR-генератор с порождающим многочленом степени L_2 генерирует двоичную «управляющую» последовательность $\{s_t\}$. С помощью этих двух последовательностей $\{a_t\}$, $\{s_t\}$ строится выходная последовательность $\{x_t\}$, которая включает те биты a_t, для которых соответствующее значение $s_t = 1$; если $s_t = 0$, то значение a_t игнорируется (отбрасывается). Такой генератор называется прореживающим [3]. Самосжимающий генератор [4] является модификацией прореживающего. Вместо генерации от двух разных регистров управляющей последовательности и последовательности, подлежащей сжатию, они обе берутся из одного и того же регистра. Регистр с порождающим многочленом степени d сдвигается на два такта. Если первый бит в паре равен 1, то выход генератора – второй бит этой пары. Если первый бит равен 0, то оба бита игнорируются и делаются еще два такта.

Как известно, при решении задачи статистического распознавания образов [5] выделяются два этапа: наиболее трудный этап построения пространства M информативных признаков $\rho_1, \ldots, \rho_M$, несущих информацию о разделимости классов $\{\Omega_i\}$, и этап построения решающего правила (разделяющих поверхностей) в пространстве найденных информативных признаков $\rho_1, \ldots, \rho_M$. Данная статья посвящена задаче построения пространства информативных признаков для статистического распознавания криптографических генераторов.

1. Подход к построению информативных признаков генераторов

Будем предполагать, что выходная последовательность генератора $x_t \in V$ является случайной последовательностью на некотором вероятностном пространстве $(\Omega, \mathcal{F}, \mathrm{P})$. Разобьем последовательность $x_1, x_2, \ldots, x_T$ на $l = [T / n]$ фрагментов (n-грамм) длины n $X^{(1)}, X^{(2)}, \ldots, X^{(l)}$, $X^{(k)} = (x_{(k-1)n+1}, \ldots, x_{kn}) \in V^n$ (если $nl < T$, $x_{nl+1}, \ldots, x_T$ не рассматриваем). Пусть наблюдается некоторая статистика $a_k(n) = f(X^{(k)})$, $k = 1, 2, \ldots, l$, при различных длинах фрагмента $n \in [n_-, n_+]$, $1 \le n_- < n_+$; $a_*(n) = E_{H_*}\{a_k(n)\}$ – математическое ожидание этой статистики при истинной гипотезе H_*; $a(n)$ – наблюдаемое выборочное среднее значение статистики:

$a(n)=\frac{1}{l}\sum_{k=1}^{l}a_k(n)$ (рис. 2). В качестве признака предлагается использовать уклонение $a(n)$ от математического ожидания в l_p-метрике $(p\in\mathbb{N})$:

$$\rho^{(p)}=\frac{1}{n_+-n_-+1}\sqrt[p]{\sum_{n=n_-}^{n_+}|a(n)-a_*(n)|^p}. \qquad (1)$$

Чем $\rho^{(p)}$ больше, тем больше свойства генератора отличаются от РРСП.

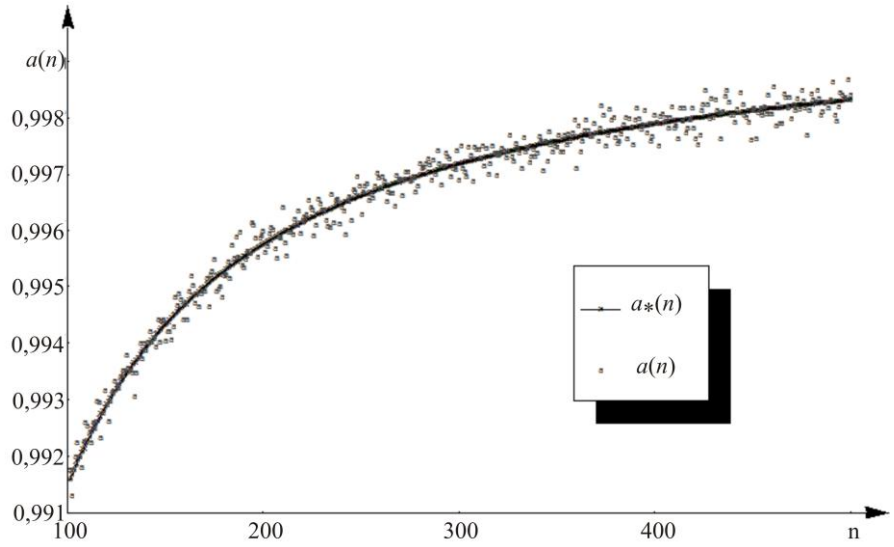

Рис. 2. Статистика $a(n)$ и ее математическое ожидание $a_*(n)$

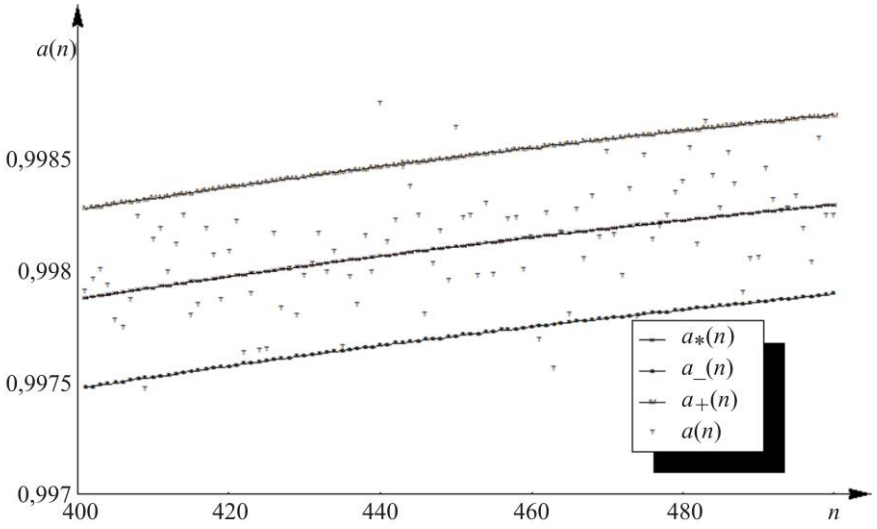

Рис. 3. Статистика $a(n)$, ее математическое ожидание $a_*(n)$ и границы допустимой окрестности

Построим некоторую допустимую окрестность $(a_-(n), a_+(n))$ ожидаемого значения $a_*(n)$: $a_-(n) < a_*(n) < a_+(n)$ – и будем учитывать выбросы за пределы этой окрестности. В качестве такой окрестности $(a_-(n), a_+(n))$ целесообразно использовать доверительный интервал заданного доверительного уровня (рис. 3). Тогда в качестве признака, учитывающего величины уклонений $a(n)$ от допустимого интервала $(a_-(n), a_+(n))$, можно взять величину

$$r = \frac{1}{n_+ - n_- + 1} \sum_{n=n_-}^{n_+} \min\{|a(n) - a_-(n)|, |a(n) - a_+(n)|\}(1 - I_{[a_-(n), a_+(n)]}(a(n))), \quad (2)$$

где $I_{[a,b]}(z) = \begin{cases} 1, z \in [a,b]; \\ 0, z \notin [a,b]. \end{cases}$

2. Нелинейные признаки на основе энтропии

Пусть $p_{i_1,\ldots,i_n} = P\{x_{t+1} = i_1, \ldots, x_{t+n} = i_n\}$ – распределение вероятностей n-граммы $(x_{t+1}, \ldots, x_{t+n}) \in V_n$, которое предполагается не зависящим от $t \in \mathbb{N}$. Многомерная (n-мерная) энтропия Шеннона для фрагмента длины n находится из выражения [1]

$$h(n) = -\sum_{i_1,\ldots,i_n} p_{i_1,\ldots,i_n} \ln p_{i_1,\ldots,i_n} . \quad (3)$$

Теорема 1. *Если справедлива гипотеза* H_*, *то*

$$h_*(n) = h(n)\big|_{H_*} = n \ln 2 = n h_*(1) . \quad (4)$$

Доказательство. При верной гипотезе имеем H_* $p_{i_1,\ldots,i_n} = 2^{-n}$, $(i_1, \ldots, i_n) \in V_n$. Поэтому согласно (3) $h_*(n) = -\sum_{i_1,\ldots,i_n} 2^{-n} \ln 2^{-n} = n \ln 2$. В силу (3) при n = 1 с учетом $p_0 = p_1 = \frac{1}{2}$ получаем $h_*(1) = -(p_0 \ln p_0 + p_1 \ln p_1) = \ln 2$, что доказывает (4). ■

Обозначим: $i = \sum_{j=1}^{n} 2^{j-1} i_j$ – представление числа $i \in \{0, 1, \ldots, 2^n - 1\}$ в двоичной системе счисления, $p_i(n) = P\{\sum_{j=1}^{n} 2^{j-1} x_j = i\} = p_{i_1,\ldots,i_n}$, $i = 0, \ldots, 2^n - 1$. Пусть наблюдается l фрагментов $X^{(1)}$, …, $X^{(l)}$. Построим статистические оценки распределения вероятностей $\{p_i(n)\}$, $i = 0, \ldots, 2^n - 1$:

$$\hat{p}_i(n) = \frac{1}{l} \sum_{k=1}^{l} \delta_{\bar{X}^{(k)}, i}, \quad \bar{X}^{(k)} = \sum_{j=1}^{n} 2^{j-1} x_j^{(k)}, \quad \delta_{\bar{X}^{(k)}, i} = \begin{cases} 1, \bar{X}^{(k)} = i; \\ 0, \bar{X}^{(k)} \neq i. \end{cases}$$

Используя подстановочный принцип, построим статистическую оценку энтропии (3):

$$\hat{h}(n) = -\sum_{i=0}^{2^n - 1} \hat{p}_i(n) \ln \hat{p}_i(n). \quad (5)$$

Теорема 2. *Если справедлива гипотеза* H_*, *то при* $l \to \infty$ *распределение вероятностей статистики* $\hat{h}(n)$ *сходится к нормальному распределению вероятностей* $\mathcal{N}(a_*(n), \sigma_*^2(n))$ *с математическим ожиданием* $a_*(n)$ *и дисперсией* $\sigma_*^2(n)$ *:*

$$a_*(n) = h_*(n), \quad \sigma_*^2(n) = \sum_{i,j=0}^{2^n - 1} (\ln p_i(n) + 1) \cdot \sigma_{ij}(n) \cdot (\ln p_j(n) + 1), \quad \sigma_{ij}(n) = \begin{cases} p_i(n) \cdot (1 - p_i(n)), i = j; \\ -p_i(n) \cdot p_j(n), i \neq j. \end{cases} \quad (6)$$

Доказательство. Свойство асимптотической нормальности и соотношение (6) вытекают из третьей теоремы непрерывности и центральной предельной теоремы [6]. ■

В качестве статистики $a(n)$ из разд. 1 можно взять оценку (5), математическое ожидание которой при истинной гипотезе H_* определяется (4). На основании статистик $\{\hat{h}(n): n_- \le n \le n_+\}$ построим информативные признаки согласно (1):

$$\rho^{(p)} = \frac{1}{n_+ - n_- + 1} \sqrt[p]{\sum_{n=n_-}^{n_+} |\hat{h}(n) - h_*(n)|^p} \,. \tag{7}$$

Для построения информативного признака согласно (2) целесообразно использовать найденное выражение дисперсии (6).

Метод оценивания многомерной энтропии согласно (5) требует вычислительных затрат порядка $O(2^n)$, для снижения которых в [7] предложено вычислять многомерную энтропию, используя расстояние Хэмминга. Опираясь на эту идею, предложим следующий подход к построению информативных признаков на основе энтропии.

Используя k-й фрагмент $X^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)}) \in V^n$ выходной последовательности и вес Хэмминга $W(x_1, ..., x_n) = \sum_{t=1}^{n} x_t$, определим семейство целочисленных статистик:

$$w_1^{(k)} = w_1(X^{(k)}) = W(X^{(k)}) = \sum_{i=1}^{n} x_i^{(k)}; \; w_2^{(k)} = w_2(X^{(k)}) = W((x_1^{(k)}x_2^{(k)}, x_1^{(k)}x_3^{(k)}, ..., x_{n-1}^{(k)}x_n^{(k)})) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} x_i x_j; ...;$$

$$w_m^{(k)} = w_m(X^{(k)}) = W((x_1^{(k)}x_2^{(k)}...x_m^{(k)}, ..., x_{n-m+1}^{(k)}...x_{n-1}^{(k)}x_n^{(k)})) = \sum_{1 \le i_1 < i_2 < ... < i_m \le n} x_{i_1}^{(k)} x_{i_2}^{(k)} ... x_{i_m}^{(k)} \in \{0\} \cup \{C_i^m : m \le i \le n\}.$$

При $m = 1$ получаем частный случай, предложенный в [7]. Множество значений, которые могут принимать величины $w_j^{(k)}$, обозначим как $W_j = \{0\} \cup \{C_i^j : j \le i \le n\}$.

Теорема 3. *Если справедлива гипотеза H_*, то*

$$E_{H_*}\{w_j^{(k)}\} = \frac{C_n^j}{2^j}, \quad j = 1, 2, \ldots, m.$$

Доказательство. Математическое ожидание $w_j^{(k)}$ не зависит от k, поэтому для удобства записи индекс k в дальнейшем будем опускать. Исходя из определения и свойств математического ожидания, получим

$$E_{H_*}\{w_j\} = \sum_{1 \le i_1 < ... < i_j \le n} E_{H_*}\{x_{i_1} \ldots x_{i_j}\} = \sum_{1 \le i_1 < ... < i_j \le n} P\{x_{i_1} = \ldots = x_{i_j} = 1\} = \sum_{1 \le i_1 < ... < i_j \le n} 2^{-j} = \frac{C_n^j}{2^j}, \quad j = 1, \ldots, m. \; ■$$

Пусть $\tilde{p}_{i_1,...,i_m}(n) = P\{w_1 = i_1, ..., w_m = i_m\}$. Тогда величину

$$\tilde{h}^{(m)}(n) = -\sum_{i_1,\ldots,i_m} \tilde{p}_{i_1,\ldots,i_m}(n) \ln \tilde{p}_{i_1,\ldots,i_m}(n) \tag{8}$$

будем называть n-мерной псевдоэнтропией m-го типа.

Пусть наблюдается l фрагментов $X^{(1)}, \ldots, X^{(l)}$. Построим статистические оценки распределения вероятностей $p_i^{(j)}(n) = P\{w_j = i\}, \quad i \in W_j, \; j = 1, \ldots, m$:

$$\hat{p}_i^{(j)}(n) = \frac{1}{l}\sum_{k=1}^{l} \delta_{w_j^{(k)}, i}.$$

Используя подстановочный принцип, построим статистическую оценку псевдоэнтропии (8):

$$\widehat{\tilde{h}^{(m)}}(n) = -\sum_{i_1 \in W_1, \ldots, i_m \in W_m} \hat{p}_{i_1}^{(1)}(n) \ldots \hat{p}_{i_m}^{(m)}(n) \ln(\hat{p}_{i_1}^{(1)}(n) \ldots \hat{p}_{i_m}^{(m)}(n)).$$

Теорема 4. *Если справедлива гипотеза H_*, то статистическая оценка n-мерной псевдоэнтропии 1-го типа находится из выражения*

$$\tilde{h}_*(n) = \tilde{h}^{(1)}(n)\Big|_{H_*} = -\sum_{i=0}^{n} \frac{C_n^i}{2^n} \ln \frac{C_n^i}{2^n}.$$

Доказательство. Так как при верной гипотезе H_* вероятность $P\{x_t = 1\} = P\{x_t = 0\} = \frac{1}{2}$, то $p_i^{(1)}(n) = P\{w_1(X) = i\} = \frac{C_n^i}{2^n}$. Следовательно, $\tilde{h}^{(1)}(n) = -\sum_{i=0}^{n} p_i^{(1)}(n) \ln p_i^{(1)}(n) = -\sum_{i=0}^{n} \frac{C_n^i}{2^n} \ln \frac{C_n^i}{2^n}$. ■

Теорема 5. *Если справедлива гипотеза H_*, то при $l \to \infty$ распределение вероятностей статистики $\widehat{\tilde{h}^{(1)}}(n)$ сходится к нормальному распределению вероятностей $\mathcal{N}(a_*(n), \sigma_*^2(n))$:*

$$a_*(n) = \tilde{h}_*(n), \ \sigma_*^2(n) = \sum_{i,j=0}^{n} (\ln p_i(n) + 1) \cdot \sigma_{ij}(n) \cdot (\ln p_j(n) + 1), \ \sigma_{ij}(n) = \begin{cases} p_i(n) \cdot (1 - p_i(n)), i = j; \\ -p_i(n) \cdot p_j(n), i \neq j. \end{cases}$$

Доказательство. Теорема доказывается аналогично теореме 2. ■

Информативные признаки согласно (2) и теореме 4 имеют вид

$$\rho^{(p)} = \frac{1}{n_+ - n_- + 1} \sqrt[p]{\sum_{n=n_-}^{n_+} |\widehat{\tilde{h}^{(1)}}(n) - \tilde{h}_*(n)|^p}.$$

Для построения информативного признака согласно (2) предлагается использовать выражение дисперсии из теоремы 5.

3. Нелинейные признаки на основе рангов матриц

Пусть на $V_n = \{0, 1\}^n$ определены $N \le n$ линейно независимых над V_n двоичных функций от n двоичных переменных $\psi_1(x_1, \ldots, x_n) \in V$, …, $\psi_N(x_1, \ldots, x_n) \in V$. Разбиваем наблюдаемый ряд $x_1, x_2, \ldots$ на фрагменты $X^{(1)}, X^{(2)}, \ldots \in V^{nN}$ длины nN: $X^{(k)} = (x_{(k-1)nN+1}, \ldots, x_{knN})$. Используя k-й фрагмент $X^{(k)} = (x_1^{(k)}, \ldots, x_{nN}^{(k)}) \in V^{nN}$ выходной последовательности, построим $(N \times N)$-матрицу

$$A^{(k)} = (a_{ij}^{(k)}) = \begin{pmatrix} \psi_1(x_1^{(k)}, \ldots, x_n^{(k)}) & \ldots \psi_N(x_1^{(k)}, \ldots, x_n^{(k)}) \\ \psi_1(x_{n+1}^{(k)}, \ldots, x_{2n}^{(k)}) & \ldots \psi_N(x_{n+1}^{(k)}, \ldots, x_{2n}^{(k)}) \\ \ldots & \ldots \quad \ldots \\ \psi_1(x_{(N-1)n+1}^{(k)}, \ldots, x_{Nn}^{(k)}) \ldots & \psi_N(x_{(N-1)n+1}^{(k)}, \ldots, x_{Nn}^{(k)}) \end{pmatrix} \in V^{N \times N} \quad (9)$$

или поэлементно $a_{ij}^{(k)} = \psi_j(x_{(i-1)n+1}^{(k)}, \ldots, x_{in}^{(k)}) \in V, \quad i, j \in \{1, 2, \ldots, N\}$.

Ранг матрицы (9) отражает наличие функциональной зависимости в последовательности. Если $\mathrm{rank}(A^{(k)}) = r < N$, то в матрице $A^{(k)}$ имеется $N - r$ линейно зависимых над V_n строк. В рас-

сматриваемом случае это означает, что найдена зависимость элементов последовательности от предыдущих:

$$\psi_j(x^{(k)}_{(i-1)n+1}, ..., x^{(k)}_{in}) = \bigoplus_{m=1}^{r} \alpha_m \psi_j(x^{(k)}_{(m-1)n+1}, ..., x^{(k)}_{mn}), \quad i \in \{r+1, ..., N\},\ j \in \{1, ..., N\},\ \alpha_m \in V = \{0, 1\}.$$

За счет выбора функций $\psi_1, \ldots, \psi_N$ можно добиться выявления этой зависимости. Поэтому в качестве признака для фрагмента $X^{(k)}$ будем использовать ранг матрицы (9):

$$r^{(k)} = rank(A^{(k)}) \in \{0, 1, ..., \min\{n, N\}\}.$$

Рассмотрим более подробно частный случай: $N = n$, $\psi_j(x_1, \ldots, x_n) = x_j$. Тогда $a^{(k)}_{ij} = x^{(k)}_{(i-1)n+j}$. Пусть наблюдается l фрагментов, т. е. $T = l \cdot N^2$. Определим статистику

$$v(n) = \frac{1}{nl} \sum_{k=1}^{l} r^{(k)}, \tag{10}$$

имеющую смысл среднего относительного ранга.

Теорема 6. *При верной гипотезе H_* математическое ожидание и дисперсия статистики* (*10*) *имеют вид*

$$E_{H_*}\{v(n)\} = v_*(n) = \frac{1}{n} \sum_{j=0}^{n} q_{nj}\, j; \tag{11}$$

$$D_{H_*}\{v(n)\} = \frac{1}{l}\left(\frac{1}{n^2} \sum_{j=0}^{n} q_{nj}\, j(j - \sum_{k=0}^{n} q_{nk} k) \right), \tag{12}$$

где $q_{nj} = P_{H_0}\{r^{(k)} = j\} = 2^{j(2n-j)-n^2} \prod_{i=0}^{j-1} \frac{(1-2^{i-n})^2}{1-2^{i-j}}, \quad j \in \{0, 1, ..., n\}$.

Доказательство. При верной гипотезе H_* статистика $r^{(k)}$ имеет известное распределение вероятностей [1]:

$$q_{nj} = P_{H_0}\{r^{(k)} = j\} = 2^{j(2n-j)-n^2} \prod_{i=0}^{j-1} \frac{(1-2^{i-n})^2}{1-2^{i-j}},\ j \in \{0, 1, ..., n\}.$$

Тогда $E_{H_*}\{v(n)\} = E_{H_*}\left\{ \frac{1}{nl} \sum_{k=1}^{l} r^{(k)} \right\} = \frac{1}{n}\left(\frac{1}{l} \sum_{k=1}^{l} E_{H_*}\{r^{(k)}\} \right) = \frac{1}{n} E_{H_*}\{r^{(k)}\} = \frac{1}{n} \sum_{j=0}^{n} q_{nj}\, j.$

Математическое ожидание $r^{(k)}$ не зависит от k, поэтому для удобства записи индекс k будем далее опускать при вычислении вероятностных характеристик $r^{(k)}$. Используя свойства дисперсии и учитывая независимость $r^{(k)}$, получим $D_{H_*}\{v(n)\} = D_{H_*}\left\{ \frac{1}{l} \sum_{k=1}^{l} \frac{r^{(k)}}{n} \right\} = \frac{1}{l} D_{H_*}\left\{ \frac{r}{n} \right\}$.

Из определения дисперсии и (11) имеем $D_{H_*}\{v(n)\} = \frac{1}{l}\left(E_{H_*}\left\{ \frac{r^2}{n^2} \right\} - E^2_{H_*}\left\{ \frac{r}{n} \right\} \right) =$

$= \frac{1}{l}\left(\frac{1}{n^2} \sum_{j=0}^{n} q_{nj}\, j^2 - E^2_{H_*}\{v(n)\} \right) = \frac{1}{l}\left(\frac{1}{n^2} \sum_{j=0}^{n} q_{nj}\, j(j - \sum_{k=0}^{n} q_{nk} k) \right)$. ■

На основании статистик $\{v(n): n_- \le n \le n_+\}$ построим информативные признаки согласно (1):

$$\rho^{(p)} = \frac{1}{n_+ - n_- + 1} \sqrt[p]{\sum_{n=n_-}^{n_+} | v(n) - v_*(n) |^p}. \tag{13}$$

Для построения информативного признака согласно (2) целесообразно использовать найденное выражение дисперсии (12).

4. Признаки, основанные на определителях матриц

Определитель матрицы позволяет учитывать зависимости между элементами матрицы. Для наблюдаемой двоичной последовательности $x_1, x_2, ..., x_T \in V = \{0, 1\}$ вычислим следующие статистики, основанные на детерминантах $(n, T \in \mathbb{N}, \quad T >> n^2)$:

$$\begin{aligned} &\alpha(1) = \frac{1}{T}\sum_{t=1}^{T} x_t, \quad \alpha(2) = \frac{1}{T-3}\sum_{t=1}^{T-3} D_t^{(2)}, D_t^{(2)} = \begin{vmatrix} x_t & x_{t+1} \\ x_{t+2} & x_{t+3} \end{vmatrix}, ..., \\ &\alpha(n) = \frac{1}{T-n^2+1}\sum_{t=1}^{T-n^2+1} D_t^{(n)}, D_t^{(n)} = \begin{vmatrix} x_t & x_{t+1} & \dots & x_{t+n-1} \\ x_{t+n} & x_{t+n+1} & \dots & x_{t+2n-1} \\ \dots & \dots & \dots & \dots \\ x_{t+(n-1)n} & x_{t+(n-1)n+1} & \dots & x_{t+n^2-1} \end{vmatrix}. \end{aligned} \tag{14}$$

Вычислим также выборочное среднеквадратичное отклонение для $\{D_t^{(n)}\}$:

$$s_n = \sqrt{\frac{1}{(T-n^2)(T-n^2+1)}\sum_{t=1}^{T-n^2+1} (D_t^{(n)} - \alpha(n))^2}. \tag{15}$$

Исследуем вероятностные характеристики статистики (14) для построения признаков согласно (1), (2). Обозначим: $\alpha_*(n) = E_{H_*}\{D_t^{(n)}\}, n \in \mathbb{N}$.

Теорема 7. *При верной гипотезе H_* математическое ожидание случайного детерминанта из (14) определяется следующими соотношениями:*

$$\alpha_*(1) = \frac{1}{2}, \quad \alpha_*(n) = 0, n = 2,\ 3,\ \dots, \quad t = 1,\ 2,\ \dots \tag{16}$$

Доказательство. $E_{H_*}\{D_t^{(1)}\} = E\{x_t\} = \frac{1}{2}$. Пусть теперь $i_1, \dots, i_n$ – перестановка чисел от 1 до n, $n > 1$; $N(i_1,\dots,i_n) \in \{0,1\}$ – четность перестановки, т. е. четность числа ее инверсий [8]. Воспользуемся тем фактом, что число четных перестановок равно числу нечетных перестановок, получим

$$E_{H_*}\{D_t^{(n)}\} = E_{H_*}\left\{\sum_{i_1,\dots,i_n} (-1)^{N(i_1,\dots,i_n)} x_{t-1+(i_1-1)n+1} \dots x_{t-1+(i_n-1)n+n}\right\} = \sum_{i_1,\dots,i_n} (-1)^{N(i_1,\dots,i_n)} E_{H_*}\left\{x_{t+(i_1-1)n} \dots x_{t-1+(i_n-1)n+n}\right\} =$$

$$= \sum_{i_1,\dots,i_n} (-1)^{N(i_1,\dots,i_n)} P\left\{x_{t+(i_1-1)n} = \dots = x_{t-1+(i_n-1)n+n} = 1\right\} = \sum_{i_1,\dots,i_n} (-1)^{N(i_1,\dots,i_n)} 2^{-n} = 2^{-n} \sum_{i_1,\dots,i_n} (-1)^{N(i_1,\dots,i_n)} = 0. \blacksquare$$

На основании статистик $\{\alpha(n): n_- \le n \le n_+\}$ построим информативные признаки согласно (1):

$$\rho^{(p)} = \frac{1}{n_+ - n_- + 1} \sqrt[p]{\sum_{n=n_-}^{n_+} |\alpha(n) - \alpha_*(n)|^p}.$$

Для построения информативного признака согласно (2) целесообразно использовать выборочное среднеквадратичное отклонение (15).

5. Нелинейные признаки на основе дискретного преобразования Фурье

Разбиваем наблюдаемый ряд $x_1, x_2, \ldots$ на фрагменты $X^{(1)}, X^{(2)}, \ldots \in V^{2^n}$ длины 2^n. Далее фрагмент $X^{(k)} = (x_1^{(k)}, \ldots, x_{2^n}^{(k)}) \in V^{2^n}$ подвергаем дискретному преобразованию Фурье (ДПФ) [9]:

$$X^{(k)} \leftrightarrow Y^{(k)} = \begin{pmatrix} y_1^{(k)} \\ \vdots \\ y_{2^n}^{(k)} \end{pmatrix} \in \mathbb{R}^{2^n}, k = 1,\ 2,\ \ldots. \tag{17}$$

Преобразуем $Y^{(k)}$ в вектор-столбец $U^{(k)} = (u_i^{(k)}) \in \mathbb{R}^n$ при помощи нелинейного преобразования

$$u_1^{(k)} = y_1^{(k)}, u_i^{(k)} = (y_i^{(k)})^2,\ i = 2, 3, ..., 2^n. \tag{18}$$

Пусть наблюдается l фрагментов $U^{(1)}, \ldots, U^{(l)}$. Вычислим выборочную ковариационную матрицу

$$\hat{\Sigma} = (\hat{\sigma}_{ij}) = \frac{1}{l-1} \sum_{k=1}^{l} (U_k - \bar{U})(U_k - \bar{U})', \bar{U} = (\bar{u}_i) = \frac{1}{l} \sum_{k=1}^{l} U_k, \tag{19}$$

или поэлементно

$$\hat{\sigma}_{ij} = \frac{1}{l-1} \sum_{k=1}^{l} (u_{ki} - \bar{u}_i)(u_{kj} - \bar{u}_j),\ i, j = 1,\ ..., 2^n.$$

Оценим собственные значения $0 \le \hat{\lambda}_1 \le ... \le \hat{\lambda}_{2^n}$ выборочной ковариационной матрицы (19) и соответствующие им собственные векторы:

$$\hat{v}_1 = (\hat{v}_{11}, ..., \hat{v}_{1\,2^n})', ..., \hat{v}_{2^n} = (\hat{v}_{2^n 1}, ..., \hat{v}_{2^n 2^n})'. \tag{20}$$

Используя (18) и (20), определим статистики исходя из свойств собственных значений и векторов ковариационной матрицы для наблюдаемого ряда:

$$T^{(1)}(U) = \hat{v}_{11}(u_1 - \bar{u}_1) + ... + \hat{v}_{1\,2^n}(u_{2^n} - \bar{u}_{2^n}), \tag{21}$$

где используется собственный вектор $\hat{v}_1$, соответствующий наименьшему собственному значению $\hat{\lambda}_1$. Построим по наблюдаемому ряду выборку из значений статистик (21):

$$t_k^{(1)} = T^{(1)}(U(X^{(k)})), k = 1, ..., l. \tag{22}$$

Для выборки (22) вычислим выборочное среднее и выборочную дисперсию:

$$\overline{t}^{(1)} = \frac{1}{l}\sum_{k=1}^{l} t_k^{(1)},\ \ s^{2(1)} = \frac{1}{l-1}\sum_{k=1}^{l}(t_k^{(1)} - \overline{t}^{(1)})^2. \tag{23}$$

Статистики (23) можно использовать для построения признаков согласно (1) и (2), а именно $\overline{t}^{(1)}$ – в качестве статистики $a(n)$, а $s^{2(1)}$ – для построения допустимой окрестности. При этом полагаем $a_*(n) = 0, n \in \mathbb{N}$.

6. Решающее правило

Пусть методами из разд. 1 – 5 данной статьи построено пространство из M информативных признаков $(\rho_1, \ldots, \rho_M) \in \mathbb{R}^M$, где количество признаков M и конкретный набор признаков из множества признаков, построенных ранее, определяются исходя из перечня криптографических генераторов, подлежащих распознаванию.

Опишем применение метода дискриминантного анализа [6, 10] для решения задачи распознавания в случае двух классов, когда признаки имеют нормальное распределение. Обоснованием предположения нормальности распределения вероятностей признаков является общий вид признаков (2.1), допускающий применение центральной предельной теоремы при $n_+ >> 1$. Пусть по входной последовательности построен вектор признаков ρ и имеется $m_i > M$ обучающих реализаций $\rho_1^{(i)}, \ldots, \rho_{m_i}^{(i)} \in \mathbb{R}^M$, принадлежащих классу Ω_i, $i \in \{1, 2\}$. Построим оценки математического ожидания вектора признаков и ковариационной матрицы для каждого из классов. Оценка для математического ожидания в классе Ω_i есть выборочное среднее:

$$\hat{\mu}_i = \frac{1}{m_i}\sum_{m=1}^{m_i}\rho_m^{(i)}, \tag{24}$$

а оценка для ковариационной матрицы – выборочная ковариационная матрица:

$$\hat{\Sigma}_i = \frac{1}{m_i - 1}\sum_{m=1}^{m_i}(\rho_m^{(i)} - \hat{\mu}_i)(\rho_m^{(i)} - \hat{\mu}_i)'. \tag{25}$$

Байесовское решающее правило имеет вид [10]

$$d = d_0(\rho) = \arg\min_{i\in\{1,2\}}(\ln|\hat{\Sigma}_i| + (\rho - \hat{\mu}_i)'\hat{\Sigma}_i^{-1}(\rho - \hat{\mu}_i) - 2\ln\hat{\pi}_i), \quad \rho \in \mathbb{R}^M, \tag{26}$$

где оценки $\{\hat{\mu}_i\}$ и $\{\hat{\Sigma}_i\}$ вычисляются по формулам (24) и (25) соответственно, априорные вероятности классов полагаем равными, т. е. $\hat{\pi}_i = P\{\rho \in \Omega_i\} = \frac{1}{2}$, $i \in \{1, 2\}$.

В частности, при равных ковариационных матрицах $\Sigma_1 = \Sigma_2 = \Sigma$ получаем линейное решающее правило

$$d = d_0(\rho) = \arg\min_{i\in\{1,2\}}((\hat{\mu}_i - 2\rho)'\hat{\Sigma}^{-1}\hat{\mu}_i), \quad \rho \in \mathbb{R}^M. \tag{27}$$

7. Результаты компьютерных экспериментов

Проведены две серии компьютерных экспериментов.

Серия 1. Рассмотрим задачу распознавания двух описанных выше криптографических генераторов: Ω_1 – самосжимающий генератор псевдослучайных двоичных последовательностей с порождающим примитивным многочленом порождающего регистра сдвига $f_1(x) = x^{32} + x^{16} + x^7 + x^2 + 1$; Ω_2 – самосжимающий генератор с порождающим примитивным многочленом $f_2(x) = x^{63} + x^{54} + x^{44} + x^{20} + 1$.

В качестве признака ρ используется статистика (6). Таким образом, имеем одномерное пространство признаков ($M = 1$). Для каждого из классов Ω_1 и Ω_2 сгенерированы 40 последова-

тельностей одинаковой длины $T = 10^7$, которые отличаются начальным заполнением порождающего регистра сдвига. По две последовательности от каждого класса ($m_1 = m_2 = 2$) используются для обучения, на остальных 38 оценивается вероятность ошибки распознавания.

Заданы параметры $n_- = 1$, $n_+ = 20$, $p = 1$. Получены следующие значения оценок параметров (24), (25):

$$\hat{\mu}_1 = 0{,}1050088479,\ \hat{\Sigma}_1 = (0{,}3208798039\cdot 10^{-8});$$
$$\hat{\mu}_2 = 0{,}1048425232,\ \hat{\Sigma}_2 = (0{,}3079484568\cdot 10^{-8}).$$

Признаки (7) имеют асимптотически нормальное распределение при $l \to \infty$. Это следует из теоремы о функциональном преобразовании нормально распределенных величин [10] и теоремы 2. Разделимость классов Ω_1, Ω_2 по информативному признаку (7) проиллюстрирована на рис. 4, где темно-серый цвет соответствует Ω_1, светло-серый – Ω_2, серый – пересечению классов. Для распознавания применяется решающее правило (26), которое при $M = 1$ принимает следующий вид:

$$d(\rho) = \begin{cases} 1, & \rho \in [0;\, 0{,}09683808) \cup (0{,}104925219;\, +\infty); \\ 2, & \rho \in [0{,}09683808;\quad 0{,}104925219]. \end{cases}$$

Результаты распознавания представлены в табл. 1, оценка безусловной вероятности ошибки равна 0,22.

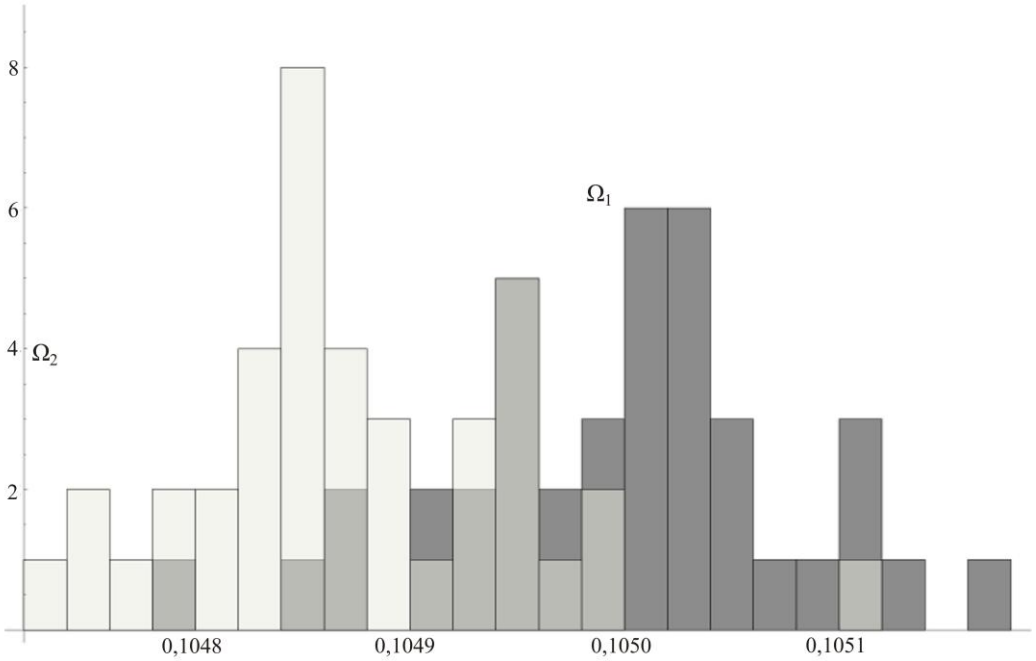

Рис. 4. Гистограммы значений признаков классов Ω_1 и Ω_2

Таблица 1

Результаты распознавания

Класс	Число решений в пользу Ω_1	Число решений в пользу Ω_2	Условная вероятность ошибки
Ω_1	32	6	0,16
Ω_2	11	27	0,29

Серия 2. Рассмотрим также задачу распознавания двух описанных ранее криптографических генераторов. Пусть Ω_1 – прореживающий генератор псевдослучайных двоичных последовательностей с порождающим примитивным многочленом 63-й степени порождающего реги-

стра сдвига и порождающим многочленом $f(x) = x^{13} + x^8 + x^5 + x^3 + 1$ управляющего регистра сдвига; Ω_2 – самосжимающий генератор псевдослучайных двоичных последовательностей с порождающим примитивным многочленом 63-й степени порождающего регистра сдвига.

В качестве признака ρ используется значение величины (13). Таким образом, имеем одномерное пространство признаков ($M = 1$). Для каждого из классов Ω_1 и Ω_2 сгенерированы 30 последовательностей по следующему принципу. В качестве характеристических многочленов порождающих регистров сдвига используются три примитивных многочлена: $g_1(x) = x^{63} + x + 1$, $g_2(x) = x^{63} + x^{54} + x^{44} + x^{20} + 1$, $g_3(x) = x^{63} + x^{60} + x^{22} + x^{18} + x^8 + x^5 + 1$. Для каждого многочлена в каждом из классов сгенерированы 10 последовательностей одинаковой длины $T = 10^7$, которые отличаются начальным заполнением порождающего регистра сдвига. Для обучения используется по одной последовательности от каждого многочлена, т. е. для каждого из классов имеем три обучающие реализации ($m_1 = m_2 = 3$). Оставшиеся 27 реализаций каждого из классов используются для оценивания вероятности ошибки распознавания.

Заданы параметры $n_- = 1$, $n_+ = 2236$, $p = 1$. Получены следующие значения оценок параметров (24), (25):

$$\hat{\mu}_1 = 0{,}001575833409,\ \hat{\Sigma}_1 = (0{,}1120998871\cdot 10^{-8});$$

$$\hat{\mu}_2 = 0{,}0001741820352,\ \hat{\Sigma}_2 = (0{,}706244773\cdot 10^{-12}).$$

Можно применить метод дискриминантного анализа, описанный в разд. 6. Однако вид гистограммы значений построенных признаков (рис. 5) указывает на то, что можно применить более простое линейное решающее правило:

$$d(\rho) = \begin{cases} 1, & \rho > \dfrac{\hat{\mu}_1 + \hat{\mu}_2}{2} = 0{,}008750077221; \\ 2, & \rho \le \dfrac{\hat{\mu}_1 + \hat{\mu}_2}{2} = 0{,}008750077221. \end{cases}$$

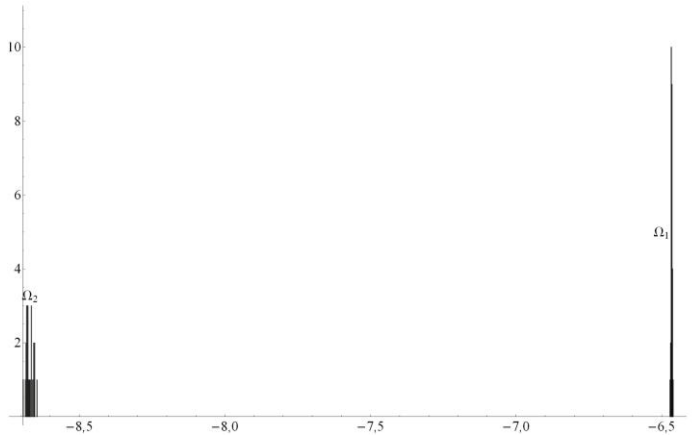

Рис. 5. Гистограмма значений признаков классов Ω_1 и Ω_2 в логарифмической шкале

Результаты распознавания представлены в табл. 2, оценка безусловной вероятности ошибки равна 0.

Таблица 2

Результаты распознавания

Класс	Число решений в пользу Ω_1	Число решений в пользу Ω_2	Условная вероятность ошибки
Ω_1	27	0	0
Ω_2	0	27	0

Заключение

Для решения актуальной задачи распознавания криптографических генераторов разработан общий подход к построению информативных признаков. Этот подход учитывает динамику изменения вероятностных характеристик генераторов при изменении длины применяемого фрагмента. Построенные согласно данному подходу признаки универсальны, так как не используют специфических свойств генераторов, и поэтому их можно применять для распознавания различных типов генераторов. Проиллюстрировано применение одного из признаков для распознавания прореживающего и самосжимающего генераторов, другого – для распознавания самосжимающих генераторов с различными параметрами.

Список литературы

1. Математические и компьютерные основы криптологии / Ю.С. Харин [и др.]. – Минск : Новое знание, 2003. – 382 с.

2. Харин, Ю.С. Проблемы математики и информатики в области защиты информации / Ю.С. Харин // Весці Нацыянальнай акадэміі навук Беларусі. Сер. фіз.-мат. навук. – 2007. – № 4. – С. 84–95.

3. Coppersmith, D. The shrinking generator / D. Coppersmith, Y. Krawchuk, Y. Mansour // Advanced in Cryptology: Proc. of Crypto 93, LNCS 773. – Berlin : Springer-Verlag, 1994. – P. 22–39.

4. Meier, W. Analysis of pseudo random sequences generated by cellular automata / W. Meier, O. Staffelbach // Advances in Cryptology / Eurocrypt '91. – Berlin : Springer-Verlag, 1992. – P. 186–199.

5. Kharin, Y. Robustness in Statistical Pattern Recognition / Y. Kharin. – Dordrecht : Kluwer Academic Publishers, 1996. – 302 p.

6. Боровков, А.А. Математическая статистика / А.А. Боровков. – Новосибирск : Наука, 1997. – 772 с.

7. Куликов, С.В. Оценка энтропии биометрических образов через переход к дискретному представлению с асимметричным распределением меры Хемминга / С.В. Куликов // Труды науч.-техн. конф. кластера пензенских предприятий, обеспечивающих безопасность информационных технологий. – Пенза, 2012. – Т. 8. – С. 18–21.

8. Милованов, М.В. Алгебра и аналитическая геометрия / М.В. Милованов, Р.И. Тышкевич, А.С. Феденко. – Минск : Вышэйшая школа, 1984. – 302 с.

9. Лукин, А.С. Введение в цифровую обработку сигналов (математические основы) / А.С. Лукин. – М. : МГУ, 2007. – 54 с.

10. Харин, Ю.С. Математическая и прикладная статистика / Ю.С. Харин, Е.Е. Жук. – Минск : БГУ, 2004. – 272 с.

Поступила 17.04.2013

Учреждение БГУ «НИИ прикладных проблем математики и информатики», Минск, пр. Независимости, 4
e-mail: kharin@bsu.by,
fpm.paluha@bsu.by

Yu.S. Kharin, V.Yu. Palukha

INFORMATIVE DESCRIPTORS FOR STATISTICAL RECOGNITION OF CRYPTOGRAPHIC GENERATORS

An approach to developing informative descriptors for solving the problem of statistical recognition of cryptographic generators is proposed. The descriptors developed by this approach are given. Theoretical results are illustrated by the computer experiments.