УДК 004.421.5: 519.237

В. Ю. ПАЛУХА, Ю. С. ХАРИН

ОБ ИНФОРМАТИВНЫХ ПРИЗНАКАХ ДЛЯ СТАТИСТИЧЕСКОГО РАСПОЗНАВАНИЯ КРИПТОГРАФИЧЕСКИХ ГЕНЕРАТОРОВ

Для решения задачи статистического распознавания криптографических генераторов предлагается принцип построения информативных признаков. Описаны признаки, построенные в соответствии с этим принципом. Приводятся результаты компьютерных экспериментов применения признаков для распознавания генераторов.

1. Введение и постановка задачи

Современная криптография не может обойтись без случайных и псевдослучайных последовательностей [1]. Случайные последовательности генерируются при помощи физических датчиков, их использование является затратным. Псевдослучайные последовательности генерируются программно. Большинство симметричных криптографических систем использует ключи, которые генерируются псевдослучайным образом; асимметричные системы используют случайные параметры [1]. Знание способа генерации ключа значительно облегчает работу криптоаналитика. Он может как восстановить значения использованных ключей, так и предсказать значения следующих ключей, которые в будущем будут сгенерированы. Генератор, который используется продолжительное время, необходимо регулярно проверять на соответствие заявленным требованиям. Вследствие этого возникает задача проверки принадлежности полученной псевдослучайной последовательности некоторому эталонному генератору. Поэтому одной из существенных задач криптоанализа является определение того способа, которым были сгенерированы случайные числа. Для решения этой задачи методами математической статистики необходимо построение пространства признаков, которые несут информацию о математической модели последовательности.

Поставим перед собой следующие задачи: построение информативных признаков, описание их вероятностных свойств и применение этих признаков для распознавания криптографических генераторов.

Итак, пусть наблюдается двоичная последовательность $x_1, x_2, \ldots \in V = \{0, 1\}$, порождённая одним из L генераторов $\Omega_1, \ldots, \Omega_L$. Необходимо определить, с помощью какого из этих генераторов была получена последовательность, т.е. отнести последовательность к одному из L классов $\Omega_1, \ldots, \Omega_L$, используя наблюдаемую реализацию $x_1, x_2, \ldots, x_T \in V$ некоторой конечной длины T. Генераторы могут быть разделены на классы различными способами: 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.

В качестве примера приведём описания прореживающего и самосжимающего генераторов.

Пусть LFSR-генератор с порождающим многочленом степени L_1 порождает «элементарную» двоичную последовательность $\{a_t\}$, а LFSR-генератор с порождающим многочленом степени L_2 порождает 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 и был предложен в [2].

Самосжимающий генератор является вариантом прореживающего и был предложен в [3]. 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. Если первый бит 0, то оба бита игнорируются и делается ещё

два такта.

2. Построение информативных признаков генераторов и алгоритм распознавания

Разбиваем последовательность $x_1, x_2, \ldots$ на фрагменты $X^{(1)}, X^{(2)}, \ldots \in V^n$ длины n: $X^{(k)} = (x_{(k-1)n+1}, \ldots, x_{kn})$. Пусть наблюдается некоторая статистика $a_k(n) = f(X^{(k)})$, где n – длина фрагмента – изменяется в заданных пределах: $n \in [n_-,\ n_+], \quad 1 \le n_- < n_+$. Пусть также нам известно математическое ожидание этой величины $a_*(n) = E_{H_0}\{a_k(n)\}$ в случае справедливости некоторой гипотезы H_0. Тогда в качестве признака можно взять значение, например, l-метрики отклонения от математического ожидания:

$$\rho^p = \frac{1}{n_+ - n_- + 1} \sqrt[p]{\sum_{n=n_-}^{n_+} |a(n) - a_*(n)|^p}. \tag{2.1}$$

Опишем применение метода дискриминантного анализа для решения задачи распознавания в случае двух классов: $L = 2$. Пусть по входной последовательности построен вектор признаков. Обозначим этот вектор через x.

Пусть у нас имеется m_l обучающих реализаций, принадлежащих классу $\Omega_l, \quad l \in \{1, 2\}$. Построим оценки математического ожидания вектора признаков и ковариационной матрицы для каждого из классов. Оценкой для математического ожидания будет выборочное среднее:

$$\hat{\mu}_l = \frac{1}{m_l} \sum_{i=1}^{m_l} x_i^{(l)}, \tag{2.2}$$

а оценкой для ковариационной матрицы будет выборочная ковариационная матрица:

$$\hat{\Sigma}_l = \frac{1}{m_l - 1} \sum_{i=1}^{m_l} (x_i^{(l)} - \hat{\mu}_l)(x_i^{(l)} - \hat{\mu}_l)^T, \tag{2.3}$$

Априорные вероятности классов $\pi_l = P\{x \in \Omega_l\}$ полагаем равными $\frac{1}{2}$.

Байесовское решающее правило имеет вид [4]:

$$d = d_0(x) = \arg \min_{l \in \{1,2\}} \{\ln |\hat{\Sigma}_l| + (x - \hat{\mu}_l)^T \hat{\Sigma}_l^{-1} (x - \hat{\mu}_l) - 2 \ln \pi_l\}, \tag{2.4}$$

где оценки $\hat{\mu}_l$ и $\hat{\Sigma}_l$ вычисляются по формулам (2.2) и (2.3) соответственно, $l \in \{1, 2\}$.

3. Признаки на основе энтропии

Разбиваем наблюдаемый ряд $x_1, x_2, \ldots$ на фрагменты $X^{(1)}, X^{(2)}, \ldots \in V^n$ длины n. Пусть $p_{i_1,\ldots,i_n} = P\{x_{t+1} = i_1, \ldots, x_{t+n} = i_n\}$ – распределение вероятностей n-граммы, которое не зависит от $t \in N$. Многомерная энтропия Шеннона фрагмента длины n равна

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

В качестве признака можно взять значение $H(n)$. Вычислить его теоретическое значение довольно просто. При верной гипотезе H_0 $H(n) = nH(1) = n \ln 2$. Однако классический метод вычисления многомерной энтропии требует огромных вычислительных затрат. Выходом является вычисление многомерной энтропии в пространстве мер Хэмминга [6].

Используя k-й фрагмент $X^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)}) \in V^n$ выходной последовательности и вес Хэмминга $W(x_1, \ldots, x_n) = \sum\limits_{i=1}^{n} x_i$, определим целочисленную случайную величину

$$w_1^{(k)} = w_1(X^{(k)}) = W(X^{(k)}) = \sum_{i=1}^{n} x_i \in \{0, 1, \ldots, n\}.$$

Пусть $p_i(n) = P\{w_1(X) = i\},\ X \in V^n$. Тогда энтропия Шеннона определяется по формуле

$$H(n) = E\{-\ln p(n)\} = -\sum_{i=0}^{n} p_i(n) \ln p_i(n). \tag{3.1}$$

При верной гипотезе $H_0 = \{x_t$ есть РРСП$\}$

$$H(n) = -\sum_{i=0}^{n} \frac{C_n^i}{2^n} \ln \frac{C_n^i}{2^n}. \tag{3.2}$$

Пусть наблюдается l фрагментов. Оценим энтропию (3.1) по формуле

$$\hat{H}(n) = -\sum_{i=0}^{n} \hat{p}_i(n) \ln \hat{p}_i(n),$$

где $\hat{p}_i(n) = \frac{1}{l} \sum\limits_{k=1}^{l} \delta_{W(X^{(k)}),i}$ – относительная частота встречаемости веса i.

Пусть теперь n изменяется в заданных пределах: $n \in [n_-,\ n_+],\quad 1 \le n_- < n_+$. Отклонения траектории $\hat{H}(n)$ от теоретического значения (3.2), которое обозначим через $H_0(n)$, будем использовать в качестве информативного признака по принципу (2.1):

$$\rho^p = \frac{1}{n_+ - n_- + 1} \sqrt[p]{\sum_{n=n_-}^{n_+} |\hat{H}(n) - H_0(n)|^p}. \tag{3.3}$$

Приведём результаты компьютерных экспериментов.

Пусть Ω_1 и Ω_2 – классы последовательностей, полученных при помощи самосжимающего генератора псевдослучайных двоичных последовательностей [3] с порождающими примитивными многочленами порождающего регистра сдвига $x^{32} + x^{16} + x^7 + x^2 + 1$ и $x^{63} + x^{54} + x^{44} + x^{20} + 1$ соответственно.

В качестве признака возьмём значение l-метрики (3.3). Таким образом, имеем одномерное пространство признаков.

Для каждого из классов Ω_1 и Ω_2 сгенерируем 20 последовательностей одинаковой длины $T = 10^7$, которые отличаются начальным заполнением порождающего регистра сдвига. Две последовательности от каждого многочлена идут на обучение, на остальных восемнадцати будем тестировать алгоритм распознавания.

Зададим следующие параметры: $n_- = 1, n_+ = 1000, p = 1$. Получены следующие значения оценок параметров (2.2), (2.3):

$$\hat{\mu}_1 = 0.006235850057, \quad \hat{\Sigma}_1 = (0.237335814 \cdot 10^{-4}); \hat{\mu}_2 = 0.003387342675, \quad \hat{\Sigma}_2 = (0.1931854597 \cdot 10^{-6}).$$

Данные, которые использовались для обучения, отображены в таблице 3.1.

Таблица 3.1 – Данные для обучения

Ω_1	Ω_2
0.009680670909	0.003698136385
0.002791029206	0.003076548966

Применим решающее правило (2.4). Оно принимает следующий вид:

$$d = \begin{cases} 1, & x \in [0;\ 0.00236188) \cup (0.004366052;\ \infty); \\ 2, & x \in [0.00236188;\ 0.004366052]. \end{cases}$$

Результаты распознавания представлены в таблице 3.2. Гистограммы значений построенных признаков, которые использовались в процессе распознавания, приведены на рисунке 3.1.

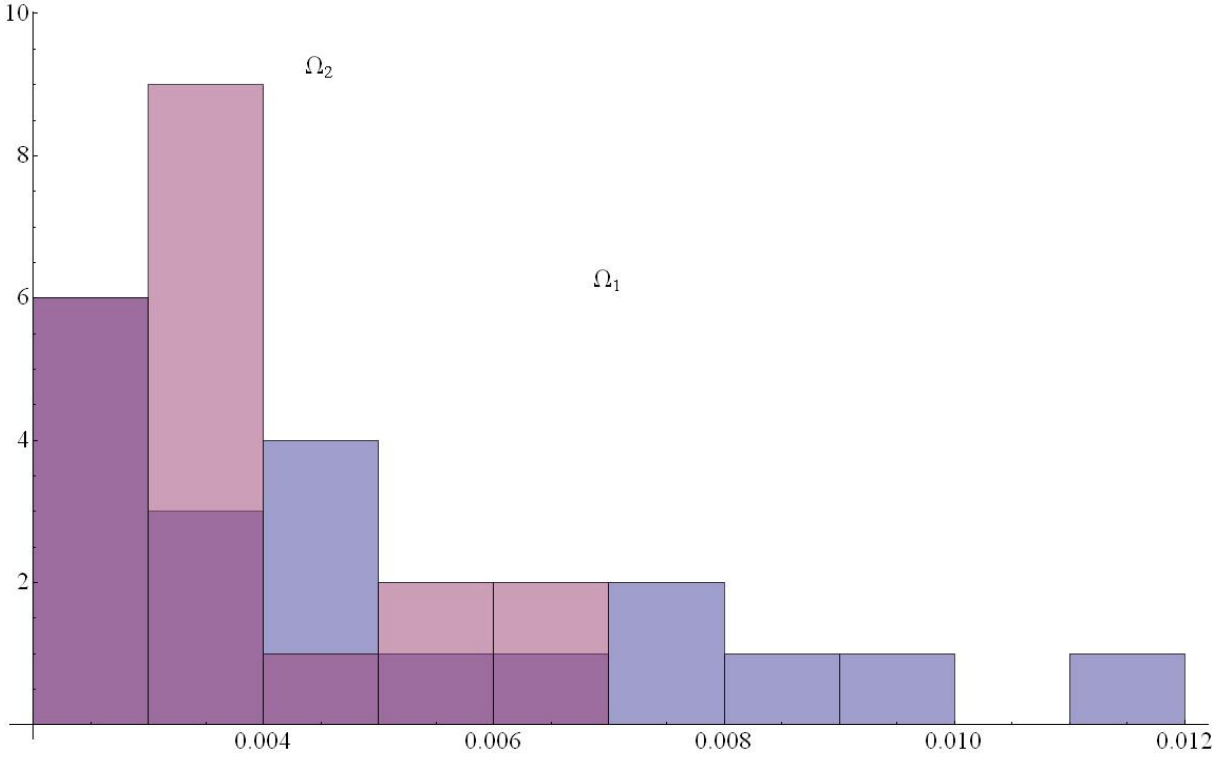

Рисунок 3.1 – Гистограммы значений признаков классов Ω_1 и Ω_2

Таблица 3.2 – Результаты распознавания

Класс	**Число решений в пользу Ω_1**	**Число решений в пользу Ω_2**	**Доля верных решений**
Ω_1	10	8	56 %
Ω_2	6	12	67 %

4. Признаки на основе рангов матриц

Разбиваем последовательность $x_1, x_2, \ldots$ на фрагменты $X^{(1)}, X^{(2)}, \ldots \in V^{n^2}$ длины n^2: $X^{(k)} = (x_{(k-1)n^2+1}, \ldots, x_{kn^2})$. Используя k-й фрагмент $X^{(k)} = (x_1^{(k)}, \ldots, x_{n^2}^{(k)})$ выходной последовательности, построим $(n \times n)$ – матрицу

$$A^{(k)} = a_{ij}^{(k)} = \begin{pmatrix} x_1^{(k)} & x_2^{(k)} & \ldots & x_n^{(k)} \\ x_{n+1}^{(k)} & x_{n+2}^{(k)} & \ldots & x_{2n}^{(k)} \\ \ldots & \ldots & \ldots & \ldots \\ x_{(n-1)n+1}^{(k)} & x_{(n-1)n+2}^{(k)} & \ldots & x_{n^2}^{(k)} \end{pmatrix} \in V^{n \times n},$$

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

Ранг матрицы отражает наличие функциональной зависимости в последовательности. Если $\mathrm{rank}(A^{(k)}) = \rho < n$, то в матрице $A^{(k)}$ имеется $n - \rho$ линейно зависимых над V^n строк. В нашем случае это означает, что найдена зависимость элементов последовательности от предыдущих:

$$x^{(k)}_{(i-1)n+j} = \sum_{m=1}^{\rho} \alpha_m x^{(k)}_{(m-1)n+j}, \quad i \in \{\rho+1, \ldots, n\},\ j \in \{1, \ldots, n\},\ \alpha_m \in V = \{0, 1\}.$$

Пусть наблюдается l фрагментов, т.е. $T = l \cdot N^2$. Определим статистику

$$v(n) = \frac{1}{nl} \sum_{k=1}^{l} r^{(k)}, \tag{4.1}$$

имеющую смысл среднего относительного ранга. При верной гипотезе $H_0 = \{x_t$ есть РРСП$\}$ математическое ожидание статистики (4.1) имеет вид

$$E_{H_0}\{v(n)\} = \frac{1}{n} \sum_{j=0}^{n} q_{nj} j, \tag{4.2}$$

где $q_{nj} = P_{H_0}\{r^{(k)} = j\} = 2^{j(2n-j)-n^2} \prod_{i=0}^{j-1} \frac{(1-2^{i-n})^2}{1-2^{i-j}}, \quad j \in \{1, \ldots, n\}.$ [5]

Пусть теперь n изменяется в заданных пределах: $n \in [n_-,\ n_+], \quad 1 \le n_- < n_+$. Отклонения траектории $v(n)$ от математического ожидания (4.2) также можно использовать в качестве информативного признака. В качестве признака возьмём значение l-метрики по (2.1):

$$\rho^p = \frac{1}{n_+ - n_- + 1} \sqrt[p]{\sum_{n=n_-}^{n_+} |v(n) - E\{v(n)\}|^p}. \tag{4.3}$$

Приведём результаты компьютерных экспериментов.

Пусть Ω_1 – класс последовательностей, полученных при помощи прореживающего генератора псевдослучайных двоичных последовательностей [2] с порождающим примитивным многочленом 63-й степени порождающего регистра сдвига и порождающим многочленом $x^{13}+x^8+x^5+x^3+1$ управляющего регистра сдвига; Ω_2 – класс последовательностей, полученных при помощи самосжимающего генератора псевдослучайных двоичных последовательностей [3] с порождающим примитивным многочленом 63-й степени порождающего регистра сдвига.

В качестве признака возьмём значение l-метрики (4.3). Таким образом, имеем одномерное пространство признаков.

Для каждого из классов Ω_1 и Ω_2 сгенерируем 30 последовательностей по следующему принципу. В качестве порождающих многочленов порождающих регистров сдвига используем три примитивных многочлена: $x^{63} + x + 1, x^{63} + x^{54} + x^{44} + x^{20} + 1, x^{63} + x^{60} + x^{22} + x^{18} + x^8 + x^5 + 1$. Для каждого многочлена в каждом из классов сгенерируем 10 последовательностей одинаковой длины $T = 10^7$, которые отличаются начальным заполнением порождающего регистра сдвига. Одна последовательность от каждого многочлена идёт на обучение, на остальных девяти будем тестировать алгоритм распознавания. Таким образом, для каждого из классов имеем 3 обучающие реализации.

Зададим следующие параметры: $n_- = 1, n_+ = 2236, p = 1$. Получены следующие значения оценок параметров (2.2), (2.3):

$$\hat{\mu}_1 = 0.001575833409, \quad \hat{\Sigma}_1 = (0.1120998871 \cdot 10^{-8});$$

$$\hat{\mu}_2 = 0.0001741820352, \quad \hat{\Sigma}_2 = (0.706244773 \cdot 10^{-12}).$$

Данные, которые использовались для обучения, отображены в таблице 4.1.

Таблица 4.1 – Данные для обучения

Ω_1	Ω_2
0.001614450173	0.0001745742966
0.001558124403	0.0001747545677
0.00155492565	0.0001732172414

Можно применить метод дискриминантного анализа, описанный в разделе 2. Однако вид гистограммы подсказывает, что можно применить более простое линейное решающее правило

$$d = \begin{cases} 1, & x > \frac{\hat{\mu}_1+\hat{\mu}_2}{2} = 0.008750077221; \\ 2, & x \leq \frac{\hat{\mu}_1+\hat{\mu}_2}{2} = 0.008750077221. \end{cases}$$

Результаты распознавания представлены в таблице 4.2. Гистограммы значений построенных признаков, которые использовались в процессе распознавания, приведены на рисунке 4.1.

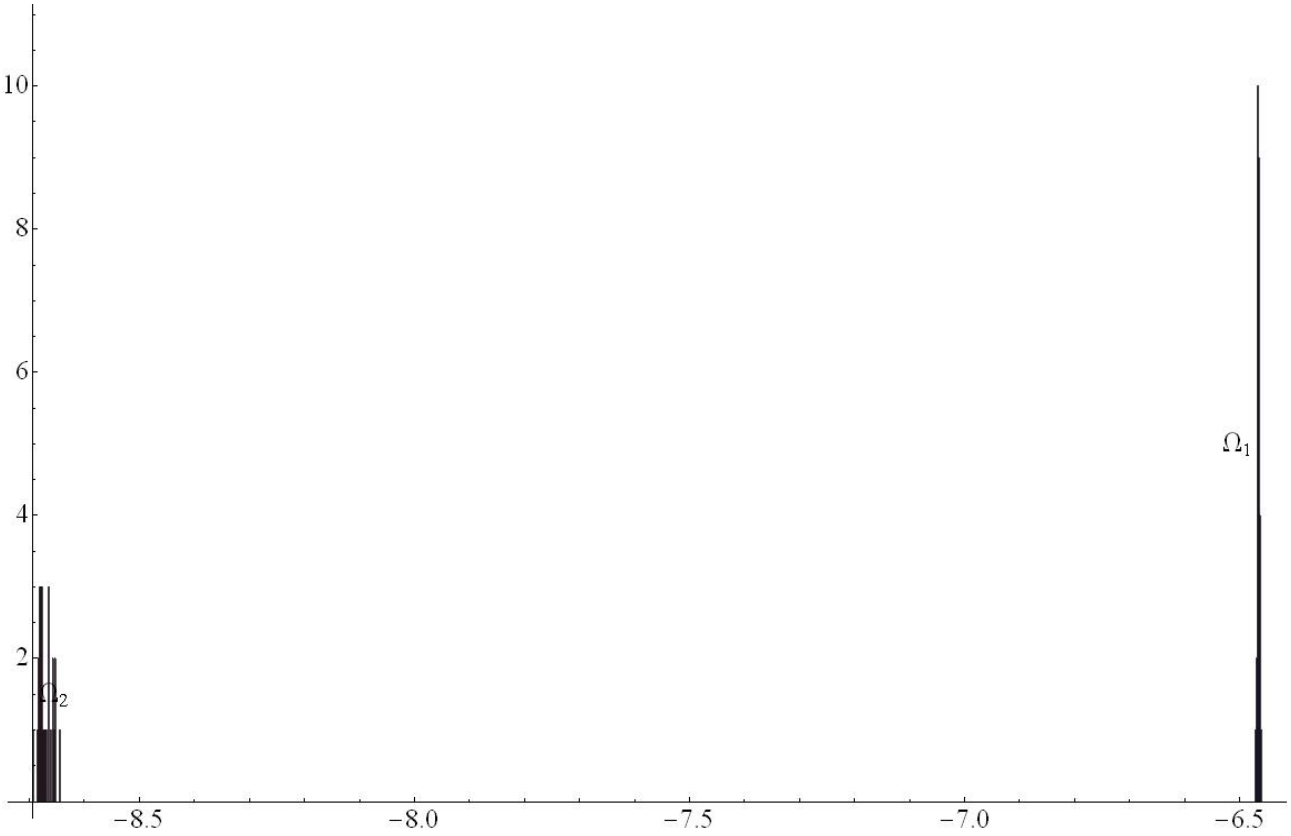

Рисунок 4.1 – Гистограммы значений признаков классов Ω_1 и Ω_2 в логарифмической шкале

Таблица 4.2 – Результаты распознавания

Класс	**Число решений в пользу Ω_1**	**Число решений в пользу Ω_2**	**Доля верных решений**
Ω_1	27	0	100 %
Ω_2	0	27	100 %

Как показал эксперимент, данный признак удобно применять для распознавания описанных выше генераторов – прореживающего и самосжимающего. Число верных решений составило 100 %.

5. Заключение

Для решения актуальной задачи криптоанализа – распознавания криптографических генераторов – разработан принцип построения информативных признаков. Этот принцип учитывает динамику изменения характеристик генераторов при изменении длины используемого фрагмента. Построенные по этому принципу признаки универсальны, т.к. не используют специфических свойств генераторов, и поэтому их можно применять для распознавания различных типов генераторов. Продемонстрировано применение одного из признаков для распознавания прореживающего и самосжимающего генераторов, и другого для распознавания самосжимающих генераторов с различными параметрами. Несмотря на хорошие результаты распознавания, даже в этих случаях исследование может быть продолжено. На момент публикации статьи не исследована зависимость результатов распознавания от длины последовательности и размера обучающей реализации. Таким образом, исследование, ставшее темой статьи, можно продолжать в различных направлениях – как применения этих признаков для распознавания других типов генераторов, поиска новых алгоритмов распознавания, так и анализа уже полученных экспериментальных результатов.

Список литературы

1. Харин, Ю. С. Математические основы криптологии / Ю. С. Харин, В. И. Берник, Г. В. Матвеев // Минск: БГУ, 1999. – 319 с.
2. Coppersmith, D. The shrinking generator / D. Coppersmith, Y. Krawchuk, Y. Mansour // Advanced in Cryptology: Proceedings of Crypto 93, LNCS 773. – 1994. – P. 22–39.
3. Meier, W. Analysis of pseudo random sequences generated by cellular automata / W. Meier, O. Staffelbach // D.W. Davies, editor, Advances in Cryptology. – Eurocrypt '91, Springer-Verlag, Berlin, 1992.– P. 186–199.
4. Харин, Ю. С. Практикум на ЭВМ по математической статистике / Ю. С. Харин, М. Д. Степанова // Минск: Университетское, 1987. – 304 с.
5. Харин, Ю. С. Математические и компьютерные основы криптологии / Ю. С. Харин, В. И. Берник, Г. В. Матвеев, С. В. Агиевич // Минск: Новое знание, 2003. – 382 с.
6. Куликов, С. В. Оценка энтропии биометрических образов через переход к дискретному представлению с ассиметричным распределением меры Хемминга / С. В. Куликов // Труды научно-технической конференции кластера пензенских предприятий, обеспечивающих безопасность информационных технологий. – Пенза, 2012. – Т. 8. – С. 18–21.

Палуха Владимир Юрьевич, студент 5 курса факультета прикладной математики и информатики Белорусского государственного университета, инженер-программист Научно-исследовательского института прикладных проблем математики и информатики, fpm.paluha@bsu.by

Харин Юрий Семёнович, член-корреспондент Национальной Академии наук Республики Беларусь, 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