УДК 519.2: 004.421.5

Палуха В.Ю.

ОБ ОДНОМ ПОДХОДЕ К ПОСТРОЕНИЮ ИНФОРМАТИВНЫХ ПРИЗНАКОВ ДЛЯ СТАТИСТИЧЕСКОГО РАСПОЗНАВАНИЯ ГЕНЕРАТОРОВ СЛУЧАЙНЫХ И ПСЕВДОСЛУЧАЙНЫХ ЧИСЕЛ

Белорусский государственный университет, Минск

Введение и постановка задачи

Современная криптография невозможна без случайных и псевдослучайных последовательностей $x_1, x_2, ... \in V = \{0,1\}$. Практическую значимость имеют генераторы последовательностей, близких по своим свойствам к равномерно распределённой случайной последовательности (РРСП). Гипотезу о том, что выходная последовательность генератора $\{x_t\}$ является равномерно распределённой, будем обозначать H_* = {{x_t} есть РРСП}.

При проведении испытаний средств криптографической защиты информации возникают задачи, которые в [1] обозначены как задачи S1, S2, S3. Математической сущностью этих практических задач является задача статистического распознавания генераторов случайных и псевдослучайных последовательностей, т.е. задача отнесения (классификации) наблюдаемой выходной последовательности генератора $x_1, ..., x_T \in V = \{0,1\}$ некоторой конечной длительности T к одному из L $(2 \le L < +\infty)$ классов Ω_1, …, Ω_L. Множество классов генераторов определяется спецификой задачи.

Как известно, для решения задачи классификации необходимо построение пространства информативных признаков. В данной статье предлагается подход к построению признаков, а также продемонстрировано применение построенных признаков для распознавания генераторов псевдослучайных чисел с помощью дискриминантного анализа [2].

1. Подход к построению информативных признаков

Дадим краткое описание подхода к построению информативных признаков, предложенное авторами в [3]. Будем предполагать, что выходная последовательность генератора $x_t \in V$ является случайной последовательностью на некотором вероятностном пространстве. Разобьём последовательность x_1, x_2, …, x_T на $l = [T / n]$ непересекающихся фрагментов (n-грамм) длины n $X^{(1)}, X^{(2)}, ..., X^{(l)}$, $X^{(k)} = (x_{(k-1)n+1}, \ldots, x_{kn}) \in V_n$ (если $nl < T$, то x_{nl+1}, …, x_T не рассматриваем). Пусть наблюдается некоторая статистика $a(n) = f(X)$, где $f(\cdot): V_n \to \square$ – некоторая борелевская функция, при различных длинах фрагмента $n \in [n_-, n_+]$, $1 \le n_- < n_+$; $a_*(n) = E_{H_*}\{a(n)\}$ – математическое ожидание этой

статистики при истинной гипотезе H_*. В качестве признака предлагается использовать уклонение статистики $a(n)$ от математического ожидания:

$$\rho = \frac{1}{n_+ - n_- + 1} \sum_{n=n_-}^{n_+} |\, a(n) - a_*(n) \,|. \qquad (1.1)$$

Признак можно модифицировать, взяв нормированное уклонение:

$$\rho = \frac{1}{n_+ - n_- + 1} \sum_{n=n_-}^{n_+} \frac{|\, a(n) - a_*(n) \,|}{|\, a_*(n) \,|}. \qquad (1.2)$$

2. 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 $t \in \mathbb{Z}$. Многомерная (n-мерная) энтропия Шеннона для фрагмента длины n равна:

$$h(n) = -\sum_{i_1,\ldots,i_n} p_{i_1,\ldots,i_n} \ln p_{i_1,\ldots,i_n}\,. \qquad (2.1)$$

Теорема 2.1. *Если справедлива гипотеза* H_*, *то*

$$h_*(n) = h(n)\big|_{H_*} = n \ln 2 = n h_*(1)\,. \qquad (2.2)$$

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 частотные 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 (2.1):

$$\hat{h}(n) = -\sum_{i=0}^{2^n - 1} \hat{p}_i(n) \ln \hat{p}_i(n). \qquad (2.3)$$

Теорема 2.2. *Если справедлива гипотеза* H_*, *то при* $l \to \infty$ *распределение вероятностей статистики* $\hat{h}(n)$ *сходится к нормальному распределению вероятностей* $\mathrm{N}(a_*(n), \sigma_*^2(n))$ *с математическим ожиданием* $a_*(n)$ *и дисперсией* $\sigma_*^2(n)$*:*

$$a_*(n) = h_*(n),\ \sigma_*^2(n) = \sum_{i,j=0}^{2^n - 1} (\ln p_i(n) + 1) \cdot \sigma_{ij}(n) \cdot (\ln p_j(n) + 1),$$

где $\sigma_{ij}(n) = \begin{cases} p_i(n)\cdot(1-p_i(n)), i=j; \\ -p_i(n)\cdot p_j(n), i\neq j. \end{cases}$

В качестве статистики $a(n)$ будем использовать оценку (2.3), математическое ожидание которой при истинной гипотезе H_* определяется (2.2). На основании статистик $\{\hat{h}(n): n_- \le n \le n_+\}$ построим информативные признаки согласно (1.1):

$$\rho = \frac{1}{n_+ - n_- + 1}\sum_{n=n_-}^{n_+} |\hat{h}(n) - h_*(n)|. \qquad (2.4)$$

Признаки (2.4) имеют асимптотически нормальное распределение при $l \to \infty$. Это следует из теоремы о функциональном преобразовании нормально распределённых величин [2] и теоремы 2.2. Опишем эксперимент, в котором для распознавания генераторов применялся признак (2.4).

Пусть Ω_1 и Ω_2 – самосжимающие [4] генераторы с характеристическими многочленами $f_1(x) = x^{32} + x^{16} + x^7 + x^2 + 1$ и $f_2(x) = x^{63} + x^{54} + x^{44} + x^{20} + 1$. Значения параметров: $n_- = 1$, $n_+ = 20$. В результате применения байесовского решающего правила [2] получена оценка безусловной вероятности ошибки 0,22.

3. Признаки на основе рангов матриц

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

$$A^{(k)} = (a_{ij}^{(k)}) = \begin{pmatrix} x_1^{(k)} & \ldots x_n^{(k)} \\ \cdots & \cdots \cdots \\ x_{(n-1)n+1}^{(k)} & \ldots x_{n^2}^{(k)} \end{pmatrix} \in V^{n\times n}. \qquad (3.1)$$

Ранг этой матрицы отражает наличие функциональной зависимости в последовательности. Если $rank(A^{(k)}) = r < n$, то в матрице $A^{(k)}$ имеется $n - r$ линейно зависимых над V_n строк. Обозначим ранг матрицы (3.1) $r^{(k)} = rank(A^{(k)}) \in \{0, 1, \ldots, n\}$. Пусть наблюдается l фрагментов, т.е. $T = l \cdot n^2$. Определим статистику, имеющую смысл среднего относительного ранга:

$$v(n) = \frac{1}{nl}\sum_{k=1}^{l} r^{(k)}, \qquad (3.2)$$

Теорема 3.1. *При верной гипотезе H_* математическое статистики* (*3.2*) *имеет вид [1]:*

$$E_{H_*}\{v(n)\} = v_*(n) = \frac{1}{n}\sum_{j=0}^{n} q_{nj}\, j, \qquad (3.3)$$

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

На основании статистик $\{v(n) : n_- \le n \le n_+\}$ построим информативные признаки согласно (1.1):

$$\rho = \frac{1}{n_+ - n_- + 1} \sum_{n=n_-}^{n_+} |v(n) - v_*(n)|. \qquad (3.4)$$

Кратко опишем эксперимент, в котором для распознавания генераторов применялся признак (3.4). Пусть Ω_1 – прореживающий генератор [5] с порождающим многочленом 63-й степени и управляющим многочленом $f(x) = x^{13} + x^8 + x^5 + x^3 + 1$, Ω_2 – самосжимающий генератор [4] с примитивным многочленом 63-й степени. Значения параметров: $n_- = 1$, $n_+ = 2236$. Генераторы оказались хорошо разделимы при помощи построенного признака, поэтому применялось линейное решающее правило. Эксперимент даёт оценку безусловной вероятности ошибки, равную 0.

4. Признаки на основе определителей 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 $(n, T \in \square, \quad T \square n^2)$:

$$\beta(1) = \frac{1}{T} \sum_{t=0}^{T-1} x_{t+1}^2, \quad \beta(2) = \frac{1}{l} \sum_{t=0}^{l-1} (D_t^{(2)})^2, \; D_t^{(2)} = \begin{vmatrix} x_{4t+1} & x_{4t+2} \\ x_{4t+3} & x_{4t+4} \end{vmatrix}, ...,$$

$$\beta(n) = \frac{1}{l} \sum_{t=0}^{l-1} (D_t^{(n)})^2, \; D_t^{(n)} = \begin{vmatrix} x_{n^2t+1} & x_{n^2t+2} & \dots & x_{n^2t+n} \\ x_{n^2t+n+1} & x_{n^2t+n+2} & \dots & x_{n^2t+2n} \\ \dots & \dots & \dots & \dots \\ x_{n^2t+(n-1)n+1} & x_{n^2t+(n-1)n+2} & \dots & x_{n^2t+n^2} \end{vmatrix}; \quad l = \left[\frac{T}{n^2} \right]. \qquad (4.1)$$

Теорема 4.1. *При верной гипотезе H_* математическое ожидание величины $(D_t^{(n)})^2$ равно*

$$E_{H_*}\{(D_t^{(n)})^2\} = \beta_*(n) = \frac{n!}{2^n} \sum_{k=0}^{n} \frac{C_n^{n-k}}{2^k} (-1)^{k-1} (k-1), \; n = 2,3,..., \; t = 1,2,\dots \qquad (4.2)$$

Заметим, что при увеличении n значение $\beta(n)$ растёт экспоненциально. Поэтому на практике удобно использовать не $\beta(n)$, а $\ln\beta(n)$. На основании статистик $\{\beta(n) : n_- \le n \le n_+\}$ и их математических ожиданий (4.2) построим информативные признаки согласно (1.2):

$$\rho = \frac{1}{n_+ - n_- + 1} \sum_{n=n_-}^{n_+} \frac{|\ln \beta(n) - \ln \beta_*(n)|}{|\ln \beta_*(n)|}. \qquad (4.3)$$

Продемонстрируем применение признака (4.3). Пусть Ω_1 – генератор Макларена – Марсальи [1], в котором генератором G_1 является регистр сдвига с многочленом $f_1(x) = x^{63} + x^{54} + x^{44} + x^{20} + 1$, а генератором G_2 является регистр сдвига с многочленом $f_2(x) = x^{13} + x^8 + x^5 + x^3 + 1$, глубина памяти K = 64, мощность алфавита |A| = 256; Ω_2 – самосжимающий генератор [6] с многочленом $f(x) = x^{63} + x^{54} + x^{44} + x^{20} + 1$. Заданы значения параметров $n_- = 1$, $n_+ = 50$. В результате применения байесовского решающего правила [2] получена оценка безусловной вероятности ошибки 0,065.

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 используемого фрагмента. Построенные согласно этому подходу признаки универсальны, т.к. не используют специфических свойств генераторов, и поэтому их можно применять для распознавания различных типов генераторов. Проиллюстрировано применение построенных признаков для распознавания прореживающего и самосжимающего генераторов, самосжимающих генераторов с различными параметрами, самосжимающего генератора и генератора Макларена – Марсальи.

Список использованных источников:

1. Математические и компьютерные основы криптологии / Ю.С. Харин [и др.]. – Минск: Новое знание, 2003. – 382 с.
2. Харин, Ю.С. Математическая и прикладная статистика / Ю.С. Харин, Е.Е. Жук. – Минск: БГУ, 2004. – 272 с.
3. Харин, Ю.С. Информативные признаки для статистического распознавания криптографических генераторов / Ю.С. Харин, В.Ю. Палуха // Информатика. – 2013. – №3 (39). – С. 126 – 138.
4. 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.
5. Coppersmith, D. The shrinking generator / D. Coppersmith, Y. Krawchuk, Y. Mansour // Advanced in Cryptology: Proceedings of Crypto 93, LNCS 773. – 1994. – P. 22–39.

Palukha V.Y.

ON ONE APPROACH TO INFORMATIVE ATTRIBUTES BUILDING FOR THE STATISTICAL RECOGNITION OF RANDOM AND PSEUDORANDOM NUMBER GENERATORS

Belarusian State University, Minsk

Summary

An approach to construction of informative features is proposed to solve the problem of statistical recognition of cryptographic generators by output sequences. Features based on entropy, ranks and determinants of matrices and constructed in accordance to this approach are described. The results of computer experiments on recognition of shrinking, self-pressed and MacLaren – Marsaglia generators are presented.