(сплошная тонкая линия) в зависимости от длины выборки n. Как видно, значение α_+, попадает в 95%-й доверительный интервал (пунктир).

Вторая серия экспериментов (рис. 2) проводилась для сравнения мощности процедуры (1) (сплошная линия) с процедурой Бонферрони (длинный пунктир), построенной по критериям шаблонов длины 3 по непересекающимся интервалам. Индивидуальный уровень значимости процедуры Бонферрони выбирался согласно [1] равным $\alpha_c = 0.019428/8 = 0.0024285$. Как видно, процедура Бонферрони проигрывает по мощности построенной двухэтапной процедуре на альтернативе $\mathbf{H}_1 : \mathbf{P}\{X_t = 1\} = 0.53$.

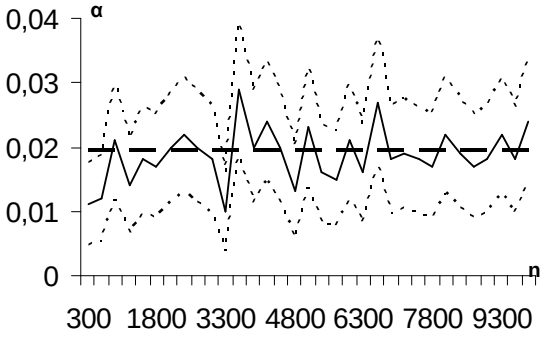

Рис. 1. Оценка вероятности ошибки первого рода построенной процедуры

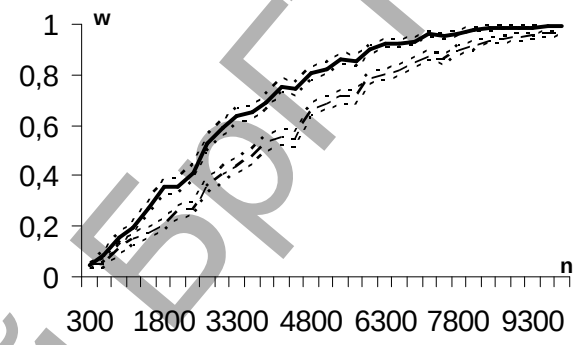

Рис. 2. Сравнение мощности построенной процедуры и процедуры Бонферрони

Литература

1. Aickin M., Gensler H. Adjusting for Multiple Testing When Reporting Research Results: The Bonferroni vs Holm Methods // Public Health Briefs, 1996, v.86, №5.

2. Lehmann E. L., Romano J. P. Generalizations of the familywise error rate // Annals of Statistics, 2005, v. 33, № 3, P. 1138-1154.

ПОСТРОЕНИЕ КРИТЕРИЯ СОГЛАСИЯ ДЛЯ ЦЕПЕЙ МАРКОВА БОЛЬШОЙ СВЯЗНОСТИ

Костевич А. Л., Шилкин А.В., БГУ, Минск

Введение

Многие задачи криптографии, например генерация ключей, предварительный анализ стойкости криптографических примитивов, требуют применения статистических критериев для обнаружения отклонений от модели независимых симметричных испытаний Бернулли. Одной из важных моделей отклонений является наличие марковской зависимости большого порядка связности *L*=32, 64, 128 и алфавитом мощности *N*. Применение классических методов выявления марковской зависимости [1] является невозможным ввиду большого числа параметров $N^L(N-1)$.

В данной статье предлагается подход к построению критерия согласия о значении матрицы вероятностей одношаговых переходов цепи Маркова большой связности, основанный на замене задачи проверки гипотезы согласия задачей прогнозирования реализации цепи Маркова. Работоспособность предлагаемого подхода иллюстрируется на примере MTD–модели цепи Маркова большой связности [2].

Построение критерия согласия

Пусть $\mathrm{X}=(X_{-m},\ldots,X_0,\ldots,X_n)$ реализация *L*-связной цепи Маркова с алфавитом 1,...,*N* и неизвестной матрицей вероятностей одношаговых переходов $P^0=(p^0_{i_1,i_2,\ldots,i_L,j})$, о которой выдвинута простая гипотеза $H_0: P^0=\hat{P}$, где $\hat{P}=(\hat{p}_{i_1,i_2,\ldots,i_L,j})$ — оценка матрицы P^0, построенная каким-либо способом по первому фрагменту выборки $(X_{-m},\ldots,X_0)$. Для проверки гипотезы H_0 по второму фрагменту выборки $(X_1,\ldots,X_n)$ для каждого наблюдения будем строить его прогноз по предыдущим *L* наблюдениям с использованием принципа максимального правдоподобия и индикатор успеха прогноза:

$$\hat{X}_t=\arg\max_{j\in\{1,\ldots,N\}}\hat{P}_{X_{t-L},\ldots,X_{t-1},j};\ Y_t=I\{\hat{X}_t=X_t\} \qquad (1)$$

В результате формируется выборка Y. Исходя из (1), имеет место следующее свойство случайных величин Y_t:

$$P\{Y_t=y_t\mid Y_{t-1},\ldots,Y_1,X_{-L+1},\ldots,X_n\}=P\{Y_t=y_t\mid X_{t-1},\ldots,X_{t-L}\},\quad t=1,2,\ldots,n.$$

Теорема 1. *Логарифмическая функция правдоподобия для выборки* Y *равна:*

$$l\{\mathrm{Y}\mid\mathrm{X},P\}==\sum_{i_1,\ldots,i_L}n^{(1)}_{i_L,\ldots,i_1,j^P_{\max}}\ln P_{i_L,\ldots,i_1,j^P_{\max}}+\sum_{i_1,\ldots,i_L}n^{(0)}_{i_L,\ldots,i_1,j^P_{\max}}\ln(1-P_{i_L,\ldots,i_1,j^P_{\max}}),$$

$$j^P_{\max}=\arg\max_{j\in\{0,1,\ldots,N-1\}}P_{i_L,\ldots,i_1,j},$$

$$n^{(1)}_{i_L,\ldots,i_1,j^P_{\max}}=\sum_{t=1}^{n}I\{X_{t-L}=i_L,\ldots,X_{t-1}=i_1,X_t=j^P_{\max}\}\times I\{Y_t=1\},$$

$$n^{(0)}_{i_L,\ldots,i_1,j^P_{\max}}=\sum_{t=1}^{n}I\{X_{t-L}=i_L,\ldots,X_{t-1}=i_1,X_t\neq j^P_{\max}\}\times I\{Y_t=0\}.$$

Рассмотрим, как изменится функция правдоподобия при внесении искажений в матрицу P^0.

Теорема 2. *Пусть* $P^1_{i_L,\ldots,i_1,j}=P^0_{i_L,\ldots,i_1,j}(1+\varepsilon_{i_L,\ldots,i_1,j})$. *Тогда*

$$l\{\mathrm{Y}\mid\mathrm{X},P^1\}-l\{\mathrm{Y}\mid\mathrm{X},P^0\}=\sum_{i=0}^{1}\sum_{i_1,\ldots,i_L}[n^{(i)}_{i_1,\ldots,i_L,j^{P^1}_{\max}}\ln(\delta_{i,0}+$$

$$+(-1)^{\delta_{i,0}}P^0_{i_1,\ldots,i_L,j^{P^1}_{\max}}(1+\varepsilon_{i_1,\ldots,i_L,j^{P^1}_{\max}}))-n^{(i)}_{i_1,\ldots,i_L,j^{P^0}_{\max}}\ln(\delta_{i,0}+(-1)^{\delta_{i,0}}P^0_{i_1,\ldots,i_L,j^{P^0}_{\max}})]$$

где $\delta_{i,j}=I\{i=j\}$ *— символ Кронекера.*

Рассмотрим проверку гипотезы H_0 о чистой случайности исходной бинарной выборки X против альтернативной гипотезы H_1 о наличии в ней марковской зависимости порядка *L*. В случае верной H_0 последовательность прогнозов будет являться последовательностью независимых испытаний Бернулли с $p=P(Y_t=1)=0.5$. В случае верной H_1 будет выполняться $p=0.5+\varepsilon$, $\varepsilon>0$. Тогда проверку исходных гипотез можно заменить на проверку гипотезы о значении вероятности успеха прогноза $p=P(Y_t=1)$: $H_0: p=0.5$ против альтернативы $H_1: p>0.5$. Для проверки такой гипотезы будем использовать критерий хи–квадрат:

принимается $\begin{cases} H_0, & \chi^2 < \Delta, \\ H_1, & \text{иначе}, \end{cases} \quad \chi^2 = \sum_{i=0}^{1} \frac{(\nu_i - n/2)^2}{n/2}, \quad \nu_i = \sum_{t=1}^{n} I\{Y_t = i\},$

где Δ — квантиль уровня α хи-квадрат распределения с 1 степенью свободы.

Вычислительный эксперимент

Для проведения вычислительного эксперимента в качестве цепи Маркова большого порядка связности использовалась бинарная MTD–модель [2] связности 32: $p_{i_1,\ldots,i_L,j} = \sum_{u\in\{1,\ldots,32\}} \lambda_u q_{i_u j}$, где $\lambda = (\lambda_u)$ — вектор весовых коэффициентов $(0 \le \lambda_u < 1, \quad \sum_u \lambda_u = 1)$, $Q = (q_{ij})$ — матрица вероятностей одношаговых переходов. В экспериментах использовались следующие значения:

$$Q = \begin{bmatrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{bmatrix}, \ \lambda = (0.1, 0.0039, 0.0039, \ldots, 0.0039, 0.0039, 0, 0.0039, 0.1454)$$

В качестве алгоритма, оценивающего параметры MTD-цепи Маркова, был использован EM-алгоритм [3], который применялся к первому фрагменту выборки объема *m*=4000. По второму фрагменту выборки, объем которого изменялся от 100 до 10000, по методу Монте–Карло были построены оценки вероятности ошибки первого рода (см. рис. 1) и мощности (см. рис. 2). Из рис. 1, 2 можно видеть, что значения оценки вероятности ошибки первого рода колеблются возле уровня значимости критерия $\alpha = 0.05$, а мощность стремится к единице с ростом объема второго фрагмента. Отметим, что на результаты эксперимента влияет выбор начального приближения EM-алгоритма.

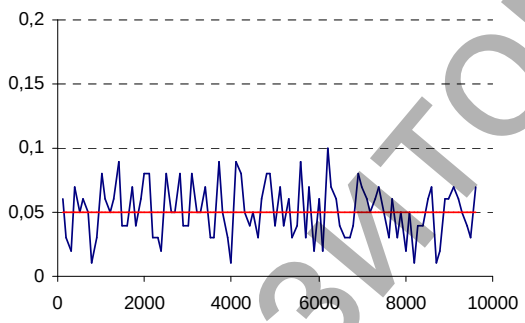

Рис. 1

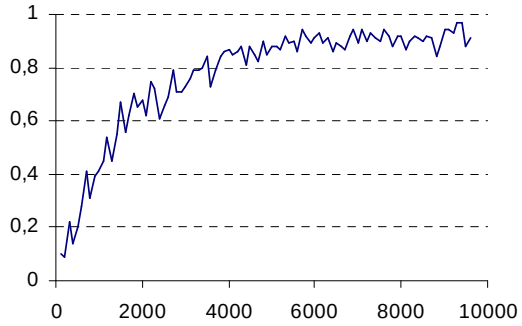

Рис. 2

Эксперимент иллюстрирует, что критерий обладает заданным уровнем значимости и является состоятельным. Отметим, что предложенный метод был успешно применен к выборке объема порядка 2^{14}, в то время как использование классических методов проверки гипотезы согласия потребовало бы выборку объема порядка 2^{34}.

Литература

1. Billingsley P. Statistical Methods in Markov Chains // The Annals of Mathematical Statistics. — 1961. — Vol. 32. — p. 12-40.
2. Raftery A. E. A model for high-order Markov chains // Journal of the Royal Statistical Society, Ser. B — 1985. — Vol. 3. — p. 528–539.
3. Костевич А.Л. EM–алгоритм оценивания параметров MTD–модели цепи Маркова большого порядка связности // Современные прикладные задачи и технологии обучения в математике и информатике (Брест): сборник научных статей международной конференции. — Минск: БГУ, 2004. — с. 98–103.