О НЕКОТОРЫХ ЗАДАЧАХ СТАТИСТИЧЕСКОГО АНАЛИЗА ДИСКРЕТНЫХ ВРЕМЕННЫХ РЯДОВ

Костевич А.Л., Харин Ю.С.

Национальный научно-исследовательский центр прикладных проблем
математики и информатики БГУ, Минск, Беларусь
kostevich@bsu.by kharin@bsu.by

Стойкость существующих систем защиты информации определяется вероятностно-статистическими свойствами выходных последовательностей, используемых в системе криптографических преобразований. Наличие зависимостей символов и отклонений их вероятностной модели от дискретного равномерного распределения в выходных последовательностях может приводить к компрометации систем защиты информации. Поэтому большое внимание уделяется разработке системы статистических критериев для выявления типовых закономерностей в выходных последовательностях криптографических преобразований, что позволяет оценивать их стойкость, выявлять «слабые» криптопреобразования.

В настоящее время наиболее известными наборами статистических критериев для выявления типовых закономерностей в дискретных случайных последовательностях являются:

1) набор критериев Д.Э. Кнута [7];

2) набор критериев Дж. Марсальи «Diehard» [4];

3) набор критериев NIST SP 800-22 [6], разработанный лабораторией информационных технологий национального института стандартизации США (ITL NIST) и который применялся для оценивания качества выходных последовательностей алгоритмов блочного шифрования, представленных на конкурс AES;

4) набор критериев, который применялся для оценивания качества выходных последовательностей криптографических алгоритмов, представленных на конкурс NESSIE [5].

По результатам обзора статистических методов анализа дискретных временных рядов можно отметить, что следующие задачи являются одними из актуальных:

1) Разработка критериев, позволяющих обнаруживать широкий класс отклонений от модели дискретной равномерно распределенной случайной последовательности. В качестве таких «универсальных» критериев могут выступать критерии, основанные на различных мерах сложности последовательности (критерии Маурера, линейной и нелинейной сложности, критерии сложности Лемпеля-Зива).

2) Разработка методик принятия итогового решения о качестве анализируемой последовательности при применении набора критериев.

Конечная однородная цепь Маркова при наличии различных искажений определяет широкий класс типовых отклонений от модели дискретной равномерно распределенной случайной последовательности.

В докладе рассмотрен один из типов искажений классической модели наблюдения цепи Маркова: наличие в реализации цепи Маркова пропущенных значений [1].

Пусть на вероятностном пространстве $(\Omega, \mathcal{F}, \mathcal{P})$ определена однородная односвязная цепь Маркова $\{X_t\}$, $t = 1, 2, \ldots$ с пространством состояний $S(N) = \{0, 1, \ldots, N-1\}$, $N \geqslant 2$ и некоторыми вектором начальных вероятностей π и матрицей вероятностей одношаговых переходов P:

$$\begin{array}{ll} \pi = (\pi_i): & \pi_i = \mathsf{P}\{X_1 = i\}, \\ P = (p_{ij}): & p_{ij} = \mathsf{P}\{X_{t+1} = j \mid X_t = i\}, \quad i, j \in S(N). \end{array} \tag{1}$$

Пусть регистрируются случайная реализация ОЦМ X с пропущенными значениями и соответствующий ей вектор индикаторов "пропусков" M, разбитые на фрагменты:

$$X = (x_1, x_2, \ldots, x_n) = (X_{(1)}|\bar{X}_{(1)}|X_{(2)}|\ldots|\bar{X}_{(T-1)}|X_{(T)}), \tag{2}$$

$$M = (m_1, m_2, \ldots, m_n) = (M_{(1)}|\bar{M}_{(1)}|M_{(2)}|\ldots|\bar{M}_{(T-1)}|M_{(T)}),$$

причем если $m_t = 1$, то значение x_t наблюдается, если же $m_t = 0$, то значение x_t является пропущенным. Здесь T — число серий из единиц в векторе индикаторов "пропусков" M, $(T \geqslant 2)$; $X_{(t)} = (x_{(t),1}, x_{(t),2}, \ldots, x_{(t),M_t^*})$ — t-й фрагмент без "пропусков", соответствующий t-й серии из единиц $M_{(t)}$ в M; $\bar{X}_{(s)} = (\bar{x}_{(s),1}, \bar{x}_{(s),2}, \ldots, \bar{x}_{(s),\bar{M}_s^*})$ — s-й ненаблюдаемый фрагмент из пропущенных значений, соответствующий s-й серии из нулей $\bar{M}_{(s)}$ в векторе M.

Вектор индикаторов "пропусков" будем рассматривать в качестве параметра модели наблюдения (1), (2). Обозначим: $p_{ij}^{(k)} = \left(P^k\right)_{ij}$ — вероятность перехода $i \to j$ за k шагов, $i, j \in S(N)$.

Теорема 1. *Если найдется натуральное число M_0 такое, что $\rho = 1 - \min_{i,j \in S(N)} p_{ij}^{(M_0)} < 1$, $\bar{M}_-^* = \min \bar{M}_t^* \geqslant M_0$, то для стационарной ОЦМ справедлива следующая аппроксимация функции правдоподобия:*

$$L(\pi, P; X, M) = \prod_{t=1}^{T} L\left(\pi, P; X_{(t)}\right) + \varepsilon(\pi, P; X, M), \tag{3}$$

$$|\varepsilon(\pi, P; X, M)| \leqslant T\rho^{\left[(\bar{M}_-^*+1)/M_0\right]-1} + \mathcal{O}\left(T^2\rho^{2\left(\left[(\bar{M}_-^*+1)/M_0\right]-1\right)}\right).$$

где $L\left(\pi, P; X_{(t)}\right) = \pi_{x_{(t),1}} \prod_{s=1}^{M_t^*-1} p_{x_{(t),s}, x_{(t),s+1}}$ — функция правдоподобия для фрагмента $X_{(t)}$ реализации (2), $[a]$ обозначает целую часть числа a.

Следствие 1. *В асимптотике $T \to \infty$, $\bar{M}_-^* \to \infty$, $T\rho^{\bar{M}_-^*/M_0} \to 0$ относительная погрешность аппроксимации (3) стремится к нулю:*

$$|\varepsilon(\pi, P; X, M) / L(\pi, P; X, M)| \to 0.$$

К общим методам статистического анализа данных при наличии пропущенных значений относят метод маргинального максимального правдоподобия, методы "игнорирования" или заполнения пропущенных значений с последующим использованием известных методов для полных данных, а также метод, основанный на применении EM-алгоритма [8].

Так как функция правдоподобия параметров модели (1), (2) существенно нелинейна по параметрам, то нахождение оценки максимального правдоподобия матрицы вероятностей одношаговых переходов аналитически или с помощью численных методов является затруднительным. Аппроксимация (3) функции правдоподобия позволяет строить оценки матрицы вероятностей одношаговых переходов и исследовать их свойства [1], строить критерии проверки гипотез о значении матрицы P [2], исследовать вопрос сходимости EM-алгоритма [3].

В [1] так же предложено семейство статистических оценок матрицы вероятностей одношаговых переходов, построенное с использованием разложений матричных функций в ряд, доказаны состоятельность и асимптотическая несмещенность предложенных оценок.

В докладе также обсуждается вопрос принятия итогового решения о качестве анализируемой последовательности при применении набора критериев.

Список литературы

[1] Kharin Yu., Kostevich A.L. Statistical inferences on finite Markov chains under missings // Problems of Discrete Mathematics. Vol. 5, 2002, p. 30-43.

[2] Kharin Yu.S., Kostevich A.L. Statistical Inferences on Transition Probabilities of Markov Chains under Missings // Simulation in Industry'2000: 12th European Simulation Symposium, p. 641-645, 2000.

[3] Kostevich A.L. On Estimation of Transition Probabilities of Markov Chain under Missings by the EM algorithm // Computer Data Analysis and Modeling / Robustness and Computer Intensive Methods: Proceedings of the 6th International Conference, Vol. 1, 2001, p. 245-250.

[4] Marsaglia G. Keynote Address: A Current View of Random Number Generators // Proceedings of the 16th Symposium on the Interface "Computer Science and Statistics". — Elsevier, 1985.

[5] NESSIE Report. List of General NESSIE Test Tools // NES/DOC/SAG/WP2/D03/1, http://www.cryptonessie.org

[6] NIST Special Publication 800-22 A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications, 2000.

[7] Кнут Д.Э. Искусство программирования. Т. 2: Получисленные алгоритмы. — 3-е изд., доп. — Вильямс, 2000. — 830 с.

[8] Литтл Р.Дж., Рубин Д.Б. Статистический анализ данных с пропусками. М.: Финансы и Статистика, 1991.