ЦЕПЬ МАРКОВА УСЛОВНОГО ПОРЯДКА И ЕЕ ПРИМЕНЕНИЕ В ЗАДАЧАХ ЗАЩИТЫ ИНФОРМАЦИИ

Ю. С. Харин, М. В. Мальцев

НИИ прикладных проблем
математики и информатики
Минск, Беларусь
E-mail: kharin@bsu.by, maltsew@mail.ru

Для исследования свойств криптографических алгоритмов и генераторов псевдослучайных последовательностей разработана математическая модель цепи Маркова условного порядка. Построены статистические оценки параметров модели, исследованы их асимптотические свойства, на основе которых построен статистический тест для обнаружения отклонения от модели «чисто случайной» последовательности.

Ключевые слова: цепь Маркова, цепь Маркова условного порядка, базовый фрагмент памяти, оценки максимального правдоподобия.

ВВЕДЕНИЕ

Цепи Маркова [1] широко используются для анализа дискретных временных рядов в таких областях, как генетика [2], экономика [3], защита информации [4]. К примеру, данная модель применяется для исследования регистров сдвига, являющихся базовыми элементами при построении многих криптографических генераторов [5].

Базовой моделью в таких исследованиях является цепь Маркова s-го порядка ($1 \le s < +\infty$) [6]. Однако число параметров $D = (N-1)N^s$ данной модели возрастает экспоненциально при увеличении s. Этот недостаток затрудняет использование цепей Маркова высоких порядков в конкретных приложениях, поскольку для статистического оценивания параметров требуется иметь реализацию не всегда доступной на практике длительности. Поэтому построен ряд «малопараметрических» моделей цепи Маркова высокого порядка, описываемых меньшим числом параметров, чем полносвязная цепь Маркова порядка s. Примерами таких моделей являются: модель Рафтери [7], цепь Маркова s-го порядка с r частичными связями [8], цепь Маркова переменного порядка [9]. К данному классу моделей относится и рассматриваемая в настоящей статье цепь Маркова условного порядка.

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ

Примем следующие обозначения: $\mathbf{N}$ – множество натуральных чисел; $A = \{0, 1, \ldots, N-1\}$ – множество состояний мощности $N \in \mathbf{N}$, $2 \le N < \infty$; $J_n^m = (j_n, j_{n+1}, \ldots, j_{m-1}, j_m) \in A^{m-n+1}$, $m, n \in \mathbf{N}$, $m \ge n$, – мультииндекс;

$\{x_t \in A\}$, $t \in \mathbf{N}$, – однородная цепь Маркова s-го порядка $(2 \le s < +\infty)$, заданная на вероятностном пространстве (Ω, F, P), с $(s+1)$-мерной матрицей вероятностей одношаговых переходов $P = (p_{J_1^{s+1}})$, $p_{J_1^{s+1}} = \mathrm{P}\{x_{t+s} = j_{s+1} | x_{t+s-1} = j_s, \ldots, x_t = j_1\}$, $J_1^{s+1} \in A^{s+1}$, $t \in \mathbf{N}$, и стационарным распределением $\pi(J_1^s) = \mathrm{P}\{x_t = j_1, \ldots, x_{t+s-1} = j_s\}$, $J_1^s \in A^s$, $t \in \mathbf{N}$; $I\{C\}$ – индикатор события C; $L \in \{1, 2, \ldots, s-1\}$, $K = N^L - 1$ – натуральные числа; $Q^{(1)}, \ldots, Q^{(M)}$ – семейство M $(1 \le M \le K+1)$ различных квадратных стохастических матриц порядка N: $Q^{(m)} = (q_{i,j}^{(m)})$, $0 \le q_{i,j}^{(m)} \le 1$, $\sum_{j \in A} q_{i,j}^{(m)} = 1$, $i, j \in A$, $1 \le m \le M$; $< J_n^m >= \sum_{k=n}^{m} N^{k-n} j_k \in \{0, 1, \ldots, N^{m-n+1} - 1\}$ – числовое представление мультииндекса J_n^m; $[F_k^l G_n^m] = J_1^{l-k+m-n+2} \in A^{l-k+m-n+2}$ – конкатенация мультииндексов F_k^l и G_n^m $(l \ge k,\ m \ge n)$; $A^{s+1}(r, J_1^l) = \{F_1^{s+1} \in A^{s+1} : f_1 = j_1, F_{r+2}^{r+l} = J_2^l\}$, $r + l \le s + 1$, – подмножество $(s+1)$-грамм с фиксированным начальным символом j_1 и фрагментом J_2^l, начиная с $(r+2)$-й позиции; $n^- = n - s$, $\nu_{s+1}(J_1^{s+1}) = \sum_{t=1}^{n^-} I\{X_t^{t+s} = J_1^{s+1}\}$, – частота $(s+1)$-граммы $J_1^{s+1} \in A^{s+1}$; $\nu(r, J_1^l) = \sum_{F_1^{s+1} \in A^{s+1}(r, J_1^l)} \nu_{s+1}(F_1^{s+1})$ – сумма частот $(s+1)$-грамм из множества $A^{s+1}(r, J_1^l)$.

Цепь Маркова s-го порядка $\{x_t \in A\}$, $t \in \mathbf{N}$, называется цепью Маркова условного порядка [10], если ее вероятности одношаговых переходов имеют следующее малопараметрическое представление:

$$p_{J_1^{s+1}} = \sum_{k=0}^{K} I\{< J_{s-L+1}^s >= k\} q_{j_{b_k}, j_{s+1}}^{(m_k)}, \qquad (1)$$

где $1 \le m_k \le M$, $1 \le b_k \le s - L$, $0 \le k \le K$, $\min_{0 \le k \le K} b_k = 1$. Последовательность элементов J_{s-L+1}^s, определяющая условие в формуле (1), называется базовым фрагментом памяти (БФП) случайной последовательности; L – длина БФП. В дальнейшем для удобства будем также использовать следующее обозначение для $q_{ij}^{(k)}$: $q(J_0^{L+1}) = \sum_{k=1}^{K} I\{< J_1^L >= k\} q_{j_0, j_{L+1}}^{(k)}$.

СТАТИСТИЧЕСКИЕ ОЦЕНКИ ПАРАМЕТРОВ

В [10] были получены статистические оценки параметров цепи Маркова условного порядка.

Если длина БФП L, значения $\{b_k\}$ и значения $\{m_k = k\}$, $k = 0, 1, \ldots, K$, заданы, то оценки максимального правдоподобия матриц вероятностей одношаговых переходов имеют вид

$$\hat{q}_{i,j}^{(k)} = \begin{cases} \sum\limits_{w \in A^L} I\{< w >= k\} \dfrac{\nu(r_k, [iwj])}{\nu(r_k, [iw])}, \text{ если } \nu(r_k, [iw]) > 0, \\ 1/N, \text{ если } \nu(r_k, [iw]) = 0, \end{cases} \quad (2)$$

где $r_k = s - b_k - L$.

Если истинные значения L и $\{m_k\}$ известны, то ОМП параметров $\{b_k\}$ имеют вид

$$\hat{b}_k = \arg\max_{1 \le b \le s-L} \sum_{i,j \in A} \nu(s - b - L, [iwj]) \ln(\hat{q}_{i,j}^{m_k}), \ k = 1,2,\ldots,K. \quad (3)$$

Оценки порядка цепи Маркова s и длины БФП L находим, решая задачу минимизации информационного функционала Байеса [11]:

$$(\hat{s}, \hat{L}) = \arg\min_{2 \le s \le \bar{S},\, 1 \le B \le \bar{L}} BIC(s, L),$$

$$BIC(s, L) = -\Big(\sum_{\substack{i,j \in A,\\ w \in A^L}} \sum_{k=0}^{K} I\{< w >= k\} \nu(s - \hat{b}_k - L, [iwj]) \ln \hat{q}_{i,j}^{(k)}\Big) + 2N^L \log n,$$

где $\bar{S} \ge 2$, $1 \le \bar{L} \le \bar{S} - 1$ – максимально допустимые значения параметров s и L, оценки $\hat{Q}^{(i)}, i = 1,\ldots,M$, и $\hat{b}_k, k = 0,\ldots,K$, вычисляются по формулам (2) и (3) соответственно.

ОБНАРУЖЕНИЕ ОТКЛОНЕНИЯ ОТ МОДЕЛИ «ЧИСТО СЛУЧАЙНОЙ» ПОСЛЕДОВАТЕЛЬНОСТИ

Задача обнаружения отклонения наблюдаемой последовательности от модели равномерно распределенной случайной последовательности (РРСП) [4] часто возникает при построении и оценке надежности систем защиты информации. На практике РРСП также называют «чисто случайной» последовательностью. Используя свойства построенных ранее оценок $\hat{Q}^{(1)},\ldots,\hat{Q}^{(M)}$, построим статистический тест, позволяющий решить эту задачу для цепи Маркова условного порядка.

Обозначим: $\bar{q}(J_0^{L+1}) = \sqrt{n^-}\left(\hat{q}(J_0^{L+1}) - q(J_0^{L+1})\right)$ – нормированное уклонение оценки $\hat{q}(J_0^{L+1})$, $Q(J_0^L) = \{j_{L+1} \in A : q(J_0^{L+1}) > 0\}$; $Q^-(J_0^L) = Q(J_0^L) \setminus \{i\}$, где i – некоторый элемент из $Q(J_0^L)$; $u = \sum\limits_{J_0^L \in A^{L+1}} |Q^-(J_0^L)|$; $\rho = \rho(n) = \sum\limits_{J_0^L \in A^{L+1}} \sum\limits_{j_{L+1} \in Q(J_0^L)} \dfrac{\pi(J_0^L)}{q(J_0^{L+1})} \bar{q}^2(J_0^{L+1})$.

Для оценок $\hat{Q}^{(1)},\ldots,\hat{Q}^{(M)}$ доказана состоятельность и асимптотическая нормальность.

Теорема 1 [10]. Если цепь Маркова условного порядка является стационарной, т. е. если выполнено условие эргодичности и начальное распределение совпадает со стационарным, то при $n - \propto$ оценки (2) являются состоятельными:

$$\hat{q}_{i,i}^{(k)} \xrightarrow{\text{P}} q_{i,j}^{(k)}, \ 1 \le k \le K.$$

Теорема 2. Если цепь Маркова условного порядка $\{x_t \in A\}, t \in \mathbf{N}$, является стационарной, то при $n - \propto$ случайные величины $\{\bar{q}(H_0^{L+1}) : H_0^{L+1} \in A^{L+2}\}$ распределены

в совокупности асимптотически нормально с нулевым асимптотическим математическим ожиданием и асимптотическими ковариациями:

$$\operatorname{cov}\left\{\overline{q}(H_0^{L+1}),\ \overline{q}(J_0^{L+1})\right\} = I\{H_0^L = J_0^L\} q(H_0^{L+1}) \frac{I\{h_{L+1} = j_{L+1}\} - q([H_0^L j_{L+1}])}{\pi(H_0^L)}.$$

Теорема 3. Если цепь Маркова условного порядка является стационарной, то при $n - \propto$ распределение статистики ρ стремится к стандартному χ^2- распределению с u степенями свободы.

Используя результат, полученный в теореме 3, построим тест для проверки гипотез о значении матриц вероятностей одношаговых переходов $Q^{(k)},\ k = 1,\ldots,K+1$: $H_0 = \{Q^{(1)} = Q_0^{(1)},\ldots,Q^{(K+1)} = Q_0^{(K+1)}\};\ H_1 = \overline{H}_0,$

$$\text{принимается} \quad \begin{cases} H_0 : \rho \le \delta, \\ H_1 : \rho > \delta, \end{cases}$$

где $\delta = G_m^{-1}(1-\alpha)$ – квантиль уровня $1-\alpha$ стандартного χ^2–распределения с u степенями свободы, $\alpha \in (0,1)$ – заданный уровень значимости.

Пусть теперь имеет место семейство контигуальных альтернатив: $H_0 = \{Q^{(1)} = Q_0^{(1)},\ldots,Q^{(K+1)} = Q_0^{(K+1)}\};\ H_1 = \{Q^{(1)} = Q_1^{(1)},\ldots,Q^{(K+1)} = Q_1^{(K+1)}\}$, $Q_1^{(k)} = Q_0^{(k)} + \frac{1}{\sqrt{n^-}}\Delta^{(k)}$, где $\Delta^{(k)} = (\Delta_{i,j}^{(k)}),\ i,j \in A,\ k = 1,\ldots,K+1$ – квадратные матрицы порядка N, такие, что $\sum_{j\in A} \Delta_{i,j}^{(k)} = 0,\ \sum_{i,j\in A} \left(\Delta_{i,j}^{(k)}\right)^2 > 0,\ k = 1,\ldots,K+1.$

Теорема 4. Если цепь Маркова является стационарной и справедлива альтернатива $H_1 = \{Q^{(1)} = Q_1^{(1)},\ldots,Q^{(K+1)} = Q_1^{(K+1)}\}$, то при $n - \propto$ распределение статистики ρ стремится к нецентральному χ^2-распределению с u степенями свободы и параметром нецентральности

$$\lambda^2 = \sum_{J_0^L \in A^{L+1}} \sum_{j_{L+1} \in Q(J_0^L)} \frac{\pi(J_0^L)}{q_0(J_0^{L+1})} \Delta^2(J_0^{L+1}),$$

$$\text{где } \Delta(J_0^{L+1}) = \sum_{k=1}^{K+1} I\{< J_1^L = k >\} \Delta_{j_0, j_{L+1}}^{(k)}.$$

Тест для проверки гипотез о значении матриц вероятностей одношаговых переходов в случае контигуальных альтернатив имеет вид, аналогичный тесту, построенному выше,

$$\text{принимается} \quad \begin{cases} H_0 : \rho \le \delta, \\ H_1 : \rho > \delta, \end{cases} \tag{4}$$

$\delta = G_m^{-1}(1-\alpha)$, α – уровень значимости.

Теорема 4 позволяет найти мощность построенного выше теста.

Следствие. В условиях теоремы 4 мощность теста (4) при $n - \propto$ и уровне значимости $\alpha \in (0,1)$ стремится к величине $w = 1 - G_{u,\lambda^2}\left(G_u^{-1}(1-\alpha)\right)$, где $G_{u,\lambda^2}() -$ функция нецентрального χ^2-распределения с u степенями свободы и параметром нецентральности λ^2.

Замечание. При $q_{ij}^{(k)} = 1/N,\ \forall k = 1,\ldots,K+1$, имеем тест проверки гипотез для обнаружения отклонения наблюдаемой последовательности $x_t \in A$ от модели РРСП: $H_0' = \{x_t \in A$ есть РРСП, т.е. $q_{i,j}^{(k)} = 1/N, \forall i,j \in A, k = 1,2,\ldots,K\}$; $H_1' = \{x_t \in A$ есть цепь Маркова условного порядка с вероятностями одношаговых переходов

$$q_{i,j}^{(k)} = q_{i,j}^{(k)}(n) = \frac{1}{N} + \frac{\Delta_{i,j}^{(k)}(n)}{\sqrt{n}} > 0,$$

где матрицы $\Delta^{(k)} = (\Delta_{i,j}^{(k)}),\ i,j \in A,\ k = 1,\ldots,K+1$ имеют тот же смысл, что и ранее.

ЛИТЕРАТУРА

1. *Кемени, Дж.* Конечные цепи Маркова / Дж. Кемени, Дж. Снелл. М. : Наука, 1970. 272 с.

2. *Уотермен, М. С.* Математические методы для анализа последовательностей ДНК / М. С. Уотермен. М.: Мир, 1999. 350 с.

3. *Ching, W. K.* High-order Markov chain models for categorical data sequences / W. K. Ching, E. Fung, K. N. Michael // Wiley Periodicals, Inc. Naval Research Logistics. 2004. Vol. 51. P. 557–574.

4. *Харин, Ю. С.* Математические и компьютерные основы криптологии / Ю. С. Харин, В. И. Берник, Г. В. Матвеев, С. В. Агиевич. Минск: Новое знание, 2003. 381 с.

5. *Максимов, Ю. И.* О цепях Маркова, связанных с двоичными регистрами сдвига со случайными элементами / Ю. И. Максимов // Труды по дискретной математике. 1997. Т. 1. С. 203–220.

6. *Дуб, Дж.* Вероятностные процессы / Дж. Дуб. М., 1956. 605 с.

7. *Raftery, A. E.* A model for high-order Markov chains / A. E. Raftery // J. Royal Statistical Society. 1985. Vol. B-47, № 3. P. 528–539.

8. *Харин, Ю. С.* Цепь Маркова с частичными связями ЦМ(*s, r*) и статистические выводы о ее параметрах / Ю. С. Харин, А. И. Петлицкий // Дискретная математика. 2007. Т. 19, № 2. С. 109–130.

9. *Buhlmann, P.* Variable length Markov chains / P. Buhlmann, A. Wyner // The Annals of Statistics. 1999. Vol. 27, № 2. P. 480–513.

10. *Харин, Ю. С.* Алгоритмы статистического анализа цепей Маркова с условной глубиной памяти / Ю. С. Харин, М. В. Мальцев // Информатика. 2011. № 1. С. 34–43.

11. *Csiszar, I.* Consistency of the BIC order estimator / I. Csiszar, P. C. Shields // Electronic research announcements of the American mathematical society. 1999. Vol. 5. P. 123–127.