

РАЗРАБОТКА СТАТИСТИЧЕСКИХ КРИТЕРИЕВ НА БАЗЕ ПРЕДИКТОРОВ ДЛЯ ТЕСТИРОВАНИЯ БИНАРНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ В ЗАДАЧАХ КРИПТОГРАФИИ

А. В. Шилкин

*НИИ прикладных проблем математики и информатики
Минск, Беларусь
E-mail: shilkinanton@gmail.com*

Рассматривается критерий тестирования бинарных последовательностей, разработанный в рамках непараметрического подхода проверки гипотезы о чистой случайности на базе предикторов. Предложены стратегии прогнозирования и приведена мощность статистического критерия на базе предиктора максимального правдоподобия для параметрической модели цепи Маркова в предложенных стратегиях прогнозирования.

Ключевые слова: проверка гипотез, гипотеза о чистой случайности.

ВВЕДЕНИЕ

Рассмотрим задачу проверки гипотезы H_0 о чистой случайности, т. е. о том, что последовательность описывается моделью независимых симметричных испытаний Бернулли. Проверка данной гипотезы для широкого класса альтернатив (в криптографии такая задача является актуальной при оценке качества физических генераторов случайных числовых последовательностей) в рамках классических (параметрических) статистических методов является затруднительной в силу большого числа возможных альтернатив. Поэтому в последнее время активно развивается непараметрический подход на основе универсальных методов сжатия. Однако сложность алгоритмической записи методов сжатия затрудняет поиск вероятностных характеристик, необходимых для построения статистических критериев, чем вызвано использование в тестах Маурера и Лемпеля – Зива [1] статистических оценок нужных параметров.

В [2] приведена схема построения критерия для проверки гипотезы H_0 , которая позволяет в рамках непараметрического подхода строить тест на базе любого универсального предиктора. Использование универсальных предикторов позволяет построить статистические критерии проверки гипотезы H_0 против широкого класса альтернатив H_1 . В качестве такого класса альтернатив могут выступать, например, стационарные эргодические марковские источники конечного порядка, любой из которых может быть выявлен с использованием универсального для этого класса предиктора Лемпеля – Зива (непараметрический подход из [2] адаптируется на случай использования предиктора Лемпеля – Зива в [3]); предиктор на основе шаблонов SPM является универсальным для последовательностей со свойством сильного перемешивания (применение подхода из [2] в случае использования предиктора SPM приведено в [4]); предиктор на основе контекстных деревьев CTW является универсальным для древовидных источников с ограниченной длиной памяти (адаптация подхода из [2] в случае использования предиктора CTW приведена в [5]). Отметим, что тра-

диционный подход проверки гипотезы H_0 против набора параметрических альтернатив заключается в построении сложной процедуры множественной проверки гипотез, в то время как использование непараметрического подхода на базе универсальных предикторов позволяет строить один статистический критерий для выявления набора параметрических альтернатив.

В данной статье для подхода из [2] рассмотрены стратегии прогнозирования и приведена мощность критерия на базе предиктора максимального правдоподобия для модели цепи Маркова 1-го порядка с дважды стохастической (д.с.) матрицей вероятностей одношаговых переходов (в. о. п.).

КРИТЕРИЙ ПРОВЕРКИ ГИПОТЕЗ НА БАЗЕ ПРЕДИКТОРОВ

Пусть наблюдается бинарный временной ряд $X_1^t = X_1, X_2, \dots, X_t$, $X_i \in A = \{0, 1\}$, который описывается набором условных вероятностей $\{P\{X_i | X_1^{i-1}\}\}$ из некоторого класса моделей M . Предиктор указывает способ построения статистических оценок неизвестных вероятностей $\{\hat{P}\{X_i | X_1^{i-1}\}\}$ и строит прогноз значения следующего наблюдения согласно максимуму оценки условной вероятности:

$$\hat{X}_i = \arg \max_{a \in A} \hat{P}\{a | X_1^{i-1}\}. \quad (1)$$

Для каждого $i = 1, 2, \dots, t-1$ будем строить прогноз для X_i по правилу (1), затем вычислять индикатор успешного прогноза $Y_i = I\{\hat{X}_i = X_i\}$.

Определение. Будем называть предиктор *подходящим* для класса моделей M , если $\arg \max_{a=0,1} \hat{P}(a | X_1^i) - \arg \max_{a=0,1} P(a | X_1^i) \xrightarrow{i \rightarrow \infty} 0$ по вероятности для любого набора условных вероятностей $\{P(a | X_1^i)\}$ из класса M .

Использование подходящего предиктора достаточно для построения оптимального прогноза: $P\{Y_{i+1} = 1\} - \max_a P(a | X_1^i) \xrightarrow{i \rightarrow \infty} 0$, по вероятности.

Если для тестируемой выборки верна гипотеза H_0 о случайности, то последовательность индикаторов $\{Y_i\}$ также будет описываться гипотезой H_0 о чистой случайности, поскольку невозможно предсказать чисто случайную последовательность. Если же для тестируемой выборки верна гипотеза H_1 (т. е. существуют $x_1, \dots, x_{i-1}, x_i \in A$ что $P\{x_i | x_1^{i-1}\} = 1/2 + \varepsilon_{x_1, \dots, x_{i-1}}$ и $0 < |\varepsilon_{x_1, \dots, x_{i-1}}| < 1/2$) и предиктор является подходящим, то последовательность индикаторов успеха прогноза будет иметь некоторое (зависящее от альтернативы и предиктора) совместное распределение, но со следующими маргинальными вероятностями:

$$H_1^{(Y)} : P\{Y_i = 1\} = \frac{1}{2} + \varepsilon, \text{ причем } \varepsilon = \varepsilon(i) > 0 \text{ при } i \rightarrow \infty.$$

В качестве статистического критерия для проверки H_0 против уже альтернативы $H_1^{(Y)}$ (и исходной альтернативы H_1) в [2] предложен следующий критерий:

$$\text{принять } \begin{cases} H_0, & \text{если } 2\sqrt{t} (S_t - \frac{1}{2}) < \Phi^{-1}(1-\alpha), \\ H_1, & \text{иначе} \end{cases} \quad S_t = \frac{1}{t} \sum_{i=1}^t Y_i, \quad (2)$$

где $\Phi(\cdot)$ – функция распределения стандартного нормального закона, α – уровень значимости.

В [2] показано, что критерий (4) имеет заданный уровень значимости и, в случае использования подходящего для H_1 предиктора, является состоятельным.

СТРАТЕГИИ ПРОГНОЗИРОВАНИЯ

Критерий проверки гипотезы о чистой случайности, построенный на базе предикторов, состоит в построении прогнозов по тестируемой последовательности и оценке успешности прогнозирования. Если статистика S_t вероятности успеха прогноза статистически значимо отличается от вероятности «случайного» прогноза ($\varepsilon = 0$), это означает, что в последовательности имеются зависимости, и она не может быть признана случайной.

При построении прогнозов можно выделить следующие операции: обучение предиктора, т. е. построение оценок вероятностей условных переходов $\{\hat{P}\{X_i | X_1^{i-1}\}\}$, прогнозирование с использованием полученных оценок, обновление оценок с учетом вновь поступивших данных. Можно предложить следующие стратегии использования предикторов для построения прогнозов.

Стратегия С1 состоит в обучении предиктора по первой части последовательности и прогнозировании второй части. Пусть наблюдается последовательность X_1^t и выбраны положительные числа m, n ($m+n=t$). Первая часть последовательности X_1^m используется для построения оценок $\{\hat{P}\{X_{m+1} | X_1^m\}\}$. Вторая часть X_{m+1}^n используется для построения последовательности прогнозов и построения статистики S_n критерия (2). Заметим, что X_{m+1}^n не используется для пересчета условных вероятностей, т. е. $\{\hat{P}\{X_{i+1} | X_1^i\}\} = \{\hat{P}\{X_{m+1} | X_1^m\}\}, i = \overline{m+1, t}$. Отметим, что стратегия 1 является удобной с точки зрения исследования теоретических свойств построенного критерия, поскольку обучающая и прогнозируемая части выборки независимы. Кроме того, стратегия 1 позволяет ограничить объем памяти и вычислительных ресурсов для построения предиктора, что достигается заданием фиксированного объема обучающей части выборки.

Стратегия С2 состоит в последовательном построении прогнозов и переобучении предиктора с учетом вновь поступивших данных. Стратегия 2 является предпочтительной с точки зрения надежности полученных результатов, поскольку для прогнозирования используется вся выборка, и статистика S_t критерия (2) построена по максимально возможному числу наблюдений. Тем не менее стратегия 2 является более сложной для исследования теоретических свойств, так как оценки условных вероятностей в результате переобучения изменяются.

Стратегия С3. Если в тестируемом физическом ГСЧП произошла поломка, то требуется эффективный способ обнаружения изменений в вероятностных характеристиках выходной последовательности. Предлагается использовать предикторы в режиме скользящего окна: обучение предиктора происходит внутри окна длины w , после построения прогноза окно сдвигается на 1 и происходит переобучение предиктора. Таким образом, для построения прогнозов в каждый момент времени используется набор условных вероятностей вида $\{\hat{P}\{X_{i+1} | X_1^i\}\} = \{\hat{P}\{X_{i+1} | X_1^{i-w+1}\}\}, i = \overline{w+1, t}$,

т. е. прогноз значения следующего символа строится по w предшествующим наблюдениям.

ИССЛЕДОВАНИЕ МОЩНОСТИ КРИТЕРИЯ

Важной задачей при построении статистических критериев является нахождение аналитического представления мощности. Рассмотрим в качестве альтернативы параметрическую модель цепи Маркова и предиктор максимального правдоподобия (МП) для нее. Приведем мощность критерия (2) на базе МП-предиктора в предложенных стратегиях прогнозирования.

Пусть ряд X_1^n описывается моделью цепи Маркова первого порядка с д. с. матрицей в. о. п. $P=(p_{ij})$, $p_{ij}=P\{X_t=j|X_{t-1}=i\}=1-P\{X_t=j|X_{t-1}\neq i\}=p=0.5+\varepsilon, 0<|\varepsilon|<0.5$.

Известно, что такая цепь Маркова имеет равномерное стационарное распределение вероятностей: $P\{X_t=i\}=\pi_i=0.5$.

Оптимальным для модели цепи Маркова является предиктор максимального правдоподобия, который для построения прогноза использует оценку матрицы в.о.п.:

$$\hat{X}_{t+1}=f(X_t)=\arg\max_j \hat{p}_{X_t,j}, \hat{p}_{ij}=\frac{n_{ij}(t)}{n_i(t)}, n_{ij}(t)=\sum_{k=1}^{t-1} I\{X_k=i, X_{k+1}=j\}, n_i(t)=\sum_{j=0,1} n_{ij}(t).$$

Пусть для оценки матрицы в. о. п. используется обучающая выборка объема m , пусть для прогнозирования используется выборка объема n , т. е. используется стратегия С1 прогнозирования. Обозначим $c_{i0}=\frac{(0.5-p_0)}{\sqrt{2p_0(1-p_0)}}$, $i=0,1$. Тогда верна

Теорема 1. При верной альтернативе, описываемой моделью цепи Маркова 1-го порядка с д. с. матрицей в. о. п. и в асимптотике $m, n \rightarrow \infty$ критерий (2) на базе МП-предиктора в стратегии С1 состоятельный, его мощность имеет вид

$$W_{m,n}=\sum_{k=1}^4 P\{\hat{P}^m \in M_k\} W(\cdot | \hat{P}^m \in M_k),$$

$$P\{\hat{P}^m \in M_1\} \approx (1-\Phi(c_{00}\sqrt{m}))\Phi(c_{10}\sqrt{m}), P\{\hat{P}^m \in M_2\} \approx \Phi(c_{00}\sqrt{m})\Phi(c_{10}\sqrt{m}),$$

$$P\{\hat{P}^m \in M_3\} \approx (1-\Phi(c_{00}\sqrt{m}))(1-\Phi(c_{10}\sqrt{m})), P\{\hat{P}^m \in M_4\} \approx \Phi(c_{00}\sqrt{m})(1-\Phi(c_{10}\sqrt{m})),$$

$$W(\cdot | \hat{P}^m \in M_k) \approx 1-\Phi(\Phi^{-1}(1-\alpha)\sigma_0/\sigma_1-\sigma_1^{-1}(\mu_1-\mu_0)), \mu_0=0.5, \sigma_0^2=(4n)^{-1},$$

$$\mu_1=\sum_{i=0,1} \pi_i p_{i,f^{(k)}(i)}, \sigma_1^2=n^{-1}(D\{Y_1\}+2\sum_{t=2}^{\infty} Cov\{Y_1 Y_t\})<\infty,$$

$$Cov\{Y_1 Y_t\}=\sum_{ij \in \{0,1\}^2} \pi_i p_{i,f^{(k)}(i)} (p_{f^{(k)}(i),j}^{(t-1)}-\pi_j) p_{j,f^{(k)}(j)},$$

$$D\{Y_1\}=\frac{1}{n} \sum_{i=0,1} \pi_i p_{i,f^{(k)}(i)} (1-\sum_{i=0,1} \pi_i p_{i,f^{(k)}(i)}), f^{(k)}(i)=\arg\max_{j=0,1} \hat{p}_{ij} \text{ при } \hat{P}=(\hat{p}_{ij}) \in M_k,$$

где через $p_{i,j}^{(t-1)}$ обозначена вероятность перехода из состояния i в сост. j за $t-1$ шаг.

Пусть матрица в. о. п. оценивается по всей предшествующей прогнозируемому символу выборке, т. е. используется стратегия С2 прогнозирования.

Теорема 2. При верной альтернативе, описываемой моделью цепи Маркова 1-го порядка с д. с. матрицей в. о. п. и в асимптотике $t \rightarrow \infty$ критерий (2) на базе МП-предиктора в стратегии С2 состоятельный, его мощность имеет вид

$$W_t \approx 1 - \Phi\left(\frac{\Phi^{-1}(1-\alpha)}{2\sqrt{t}\sigma} + \frac{\frac{1}{2}-\mu}{\sigma}\right), \mu \approx \frac{1}{2} + (\frac{1}{2}-p)\Phi(c_{00}\sqrt{t}) + (p-\frac{1}{2})\Phi(c_{10}\sqrt{t}) + O(t^{-\frac{1}{2}}),$$

$$\sigma^2 \approx \frac{1}{4t} - \frac{(\frac{1}{2}-p)^2}{t}(\Phi(c_{00}\sqrt{t}) - \Phi^2(c_{10}\sqrt{t})) + \frac{2}{t^2} \sum_{j>i} Cov\{Y_i, Y_j\} + O(t^{-\frac{3}{2}}) < \infty,$$

$$Cov\{Y_i, Y_j\} = \sum_{m=1}^4 \sum_{n=1}^4 \sum_{u=0,1} \pi_u p_{u,f^m(u)} \sum_{v=0,1} p_{v,f^n(v)} \cdot$$

$$\cdot \left(P\{\hat{P}^m \in M_m, \hat{P}^n \in M_n\} p_{f^m(u),v}^{(j-i)} - P\{\hat{P}^m \in M_m\} P\{\hat{P}^n \in M_n\} \pi_v \right),$$

где O – символ Ландау.

Пусть матрица в. о. п. оценивается по w предшествующим прогнозируемому символу наблюдениям, т. е. используется стратегия С3 прогнозирования.

Теорема 3. При верной альтернативе, описываемой моделью цепи Маркова 1-го порядка с д. с. матрицей в. о. п. и в асимптотике $t \rightarrow \infty$ критерий (2) на базе МП-предиктора в стратегии С3 состоятельный, его мощность имеет вид:

$$W_t \approx 1 - \Phi\left(\frac{\Phi^{-1}(1-\alpha)}{2\sqrt{t}\sigma} + \frac{\frac{1}{2}-\mu}{\sigma}\right), \mu \approx \frac{1}{2} + (\frac{1}{2}-p)\Phi(c_{00}\sqrt{w}) + (p-\frac{1}{2})\Phi(c_{10}\sqrt{w}) + O(w^{-\frac{1}{2}}),$$

$$\sigma^2 \approx \frac{1}{4t} - \frac{(0.5-p)^2}{t}(\Phi(c_{00}\sqrt{w}) - \Phi^2(c_{10}\sqrt{w})) + \frac{2}{t^2} \sum_{u=1}^{t-1} \sum_{v=u+1}^{\max\{u+w,t\}} Cov\{Y_u, Y_v\} + O(t^{-\frac{3}{2}}) < \infty,$$

где выражение для $Cov\{Y_u, Y_v\}$ приведено в теореме 2.

ЗАКЛЮЧЕНИЕ

В теоремах 1 – 3 установлена асимптотическая мощность критерия на базе предиктора максимального правдоподобия для модели цепи Маркова. Универсальные предикторы Лемпеля – Зива, *SPM* и *CTW* являются подходящими для более широкого класса моделей, который включает в себя модель цепи Маркова. В них используются другие статистики для оценки вероятностей условных переходов [3–5], из-за чего критерии на базе универсальных предикторов являются менее мощными, чем критерий на базе МП-предиктора, но асимптотически эквивалентными. Предлагается использовать мощность критерия на базе МП-предиктора в разных стратегиях прогнозирования в качестве оценки сверху для мощности критериев на базе универсальных предикторов.

ЛИТЕРАТУРА

1. NIST Special Publication 800-22: A statistical test suite for random and pseudorandom number generators for cryptographic applications. 2001.
2. Shilkin, A. V. On Approach to Randomness Testing on the base of the Universal Predictors / A. V. Shilkin, A. L. Kostevich // CDAM: Complex Stochastic Data and Systems. Proc. of the 8 Int. Conf. Minsk: Belarusian State University, 2007. Vol. 2. P. 256–259.
3. Шилкин, А. В. Анализ эффективности применения универсальных предикторов для проверки гипотезы о чистой случайности / А. В. Шилкин, А. Л. Костевич // Материалы III Междунар. науч. конф. «Сетевые компьютерные технологии». Минск: Издат. центр БГУ, 2007. С. 211–216.

4. Шилкин, А. В. О методике применения SPM-предиктора для тестирования бинарных последовательностей / А. В. Шилкин, В. И. Семенов // Материалы IV Междунар. конф. «Информационные системы и технологии». Минск: Акад. упр. при Президенте Респ. Беларусь, 2008. С. 90–95.

5. Шилкин А. В. Об использовании предиктора СТW для тестирования бинарных последовательностей / А. В. Шилкин, А. Л. Костевич // Материалы V Междунар. конф. форума «Информационные системы и технологии». Минск: издатель А. Н. Вараксин, 2009. Vol. 2. Р. 57–60.
