ОБ ИСПОЛЬЗОВАНИИ ПРЕДИКТОРА CTW ДЛЯ ТЕСТИРОВАНИЯ БИНАРНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

А.Л. Костевич, А.В. Шилкин

НИИ прикладных проблем математики и информатики
пр Независимости, 4, Минск, Беларусь
телефон: + (017) 2095549; e-mail: shilkinanton@gmail.com

Рассматривается критерий тестирования бинарных последовательностей, разработанный в рамках непараметрического подхода проверки гипотезы о чистой случайности на базе универсального CTW-предиктора.

Ключевые слова – Context Tree Weighting предиктор, статистическая проверка гипотез, универсальные предикторы.

1 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 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.

Приведенная в [2] схема построения критерия для проверки гипотезы H_0 позволяет в рамках непараметрического подхода строить тест на базе любого универсального предиктора. В [3] непараметрический подход из [2] адаптируется на случай использования предиктора Лемпеля – Зива. Адаптация подхода из [2] на случай использования Sampled Pattern Matching предиктора рассмотрена в [4].

В данной статье рассматривается применение Context Tree Weighting [5] предиктора, который является универсальным для древовидных источников с ограниченной длиной памяти.

2 КРИТЕРИЙ ПРОВЕРКИ ГИПОТЕЗЫ О ЧИСТОЙ СЛУЧАЙНОСТИ НА БАЗЕ УНИВЕРСАЛЬНЫХ ПРЕДИКТОРОВ

Приведем сначала предложенную в [2] схему построения статистического критерия проверки гипотезы о чистой случайности на базе любого универсального предиктора. Пусть наблюдается временной ряд $X_1, X_2, \ldots, X_t \in A\{0,1\}$, который описывается набором условных вероятностей $\{P\{X_t \mid X_{t-1},\ldots,X_1\}\}$ из некоторого класса моделей M. Пусть зарегистрированы наблюдения $x_1, x_2, \ldots, x_t$ и требуется сделать прогноз значения следующего символа X_{t+1}.

Если вероятностная модель последовательности известна, то оптимальный прогноз определяется максимумом соответствующей условной вероятности:

$$\hat{X}^*_{t+1} = \arg\max_{a \in A} P\{a \mid X_{t-1},\ldots,X_1\} \qquad (1)$$

При этом достигается минимальная средняя ошибка прогнозирования:

$$\pi^*_t = P\{\hat{X}^*_{t+1} \neq X_{t+1}\} \qquad (2)$$

Если вероятностная модель последовательности неизвестна, то должен быть построен предиктор $\{\hat{P}\{X_t \mid X_{t-1},\ldots,X_1\}\}$, прогноз строится аналогично (1), но уже с большей вероятностью ошибки прогноза:

$$\begin{aligned} \hat{X}_{t+1} &= \arg\max_{a \in A} \hat{P}\{a \mid X_{t-1},\ldots,X_1\} \\ \hat{\varepsilon}_t &= P\{\hat{X}_{t+1} \neq X_{t+1}\} \end{aligned} \qquad (3)$$

Предиктор (3) называется *универсальным* [6] для класса M, если при неизвестных $\{P\{X_t \mid X_{t-1},\ldots,X_1\}\}$ из этого класса *ошибка предсказания* сходится по вероятности к нулю при $t \to \infty$ для любого набора условных вероятностей из класса M. Под ошибкой предсказания в рамках непараметрического подхода рассматривается избыточность предиктора $r_t = \hat{\varepsilon}_t - \pi^*_t$ [4].

В [2] предлагается подход к тестированию гипотезы H_0, основанный на построении последовательности прогнозов и оценке эффективности прогнозирования *любого* универсального предиктора.

Пусть наблюдается выборка $X = (x_1,\ldots,x_n)$ объема n. Для каждого $t = 1,\ldots,n-1$ по первым t наблюдениям будем строить прогноз для x_{t+1} с

использованием предиктора (3), затем вычислять индикатор успешного прогноза: $Y_t = I\{\hat{X}_t = x_t\}$.

Если для тестируемой выборки верна гипотеза H_0 о случайности, то последовательность индикаторов $\{Y_t\}$ также будет описываться гипотезой H_0 о независимости и симметричности испытаний Бернулли. Если же для тестируемой выборки верна гипотеза H_1 и $\pi_t^* < 0.5$, то в случае универсального для H_1 предиктора последовательность индикаторов успеха прогноза будет иметь некоторое (зависящее от H_1 и предиктора) совместное распределение, но со следующими маргинальными вероятностями:

$$H_1^{(Y)}: \; P\{Y_t = 1\} = \frac{1}{2} + \varepsilon_t, \; \varepsilon_t > 0, \; t \to \infty$$

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

$$\textit{принять} \begin{cases} H_0, & \textit{если} \;\; 2\sqrt{n}\,(S - \frac{1}{2}) < \Phi^{-1}(1-\alpha), \\ H_1, & \textit{иначе} \end{cases} \qquad (4)$$

где $S = \frac{1}{n}\sum_{t=1}^{n} Y_t$ – доля успеха в прогнозах, $\Phi()$ – функция распределения стандартного нормального распределения, α – уровень значимости.

3 ОПИСАНИЕ CTW-ПРЕДИКТОРА

Строка $s = q_{l-1} q_{2-l} \cdots q_0$ является *суффиксом* строки $s' = q'_{l'-1} q'_{2-l'} \cdots q'_0$, если $l < l'$, $q_{-i} = q'_{-i}$, $i = \overline{0, l-1}$.

Суффиксное множество S – это набор строк $s(k), k = 1, 2, \ldots, |S|$, обладающее свойством *корректности* и *полноты*. Корректность означает, что ни одна строка из S не является суффиксом другой строки из S. Полнота означает, что любая полубесконечная строка $x_{-\infty}^{n} = \cdots x_{n-2} x_{n-1} x_n$ имеет суффикс во множестве S. Такой суффикс является единственным, поскольку множество S корректно.

Выберем и зафиксируем параметр D.

Источником ограниченной памяти древесной структуры назовем источник, которому соответствует суффиксное множество S такое, что $\forall s \in S$ длина $l(s) \le D$. Каждому суффиксу $s \in S$ соответствует параметр $\theta_s \in (0,1)$. Обозначим $\Theta_S = \{\theta_s, s \in S\}$.

Определим функцию $\beta_s(\cdot)$, которая выделяет суффикс из множества S для последовательности $x_{-\infty}^{n}$. Поскольку длина каждого суффикса не превосходит D, то только последние D символов последовательности определяют суффикс в S. Если источник сгенерировал последовательность $x_{-\infty}^{t-1}$, то

$$P\{X_t = 1 \mid x_{t-D}^{t-1}, S, \Theta_S\} = \theta_{\beta_s(x_{t-D}^{t-1})}, \forall t$$

Тогда вероятность того, что источник сгенерирует последовательность X_1^t равна

$$P\{X_1^t = x_1^t \mid x_{1-D}^{0}, S, \Theta_S\} = \prod_{i=1}^{t} P\{X_i = x_i \mid x_{i-D}^{i-1}, S, \Theta_S\}$$

Говорят, что все источники с одинаковым суффиксным множеством имеют одинаковую модель. Множество всех моделей с ограниченной памятью образует класс моделей C_D.

Вероятность последовательности из a нулей и b единиц, сгенерированной источником без памяти с параметром θ, равняется $(1-\theta)^a \theta^b$. Если взвесить эту вероятность по всем θ с использованием распределения Дирихле с параметром $(0.5, 0.5)$, то получится так называемая *оценка Кричевского-Трофимова*.

Вероятность Кричевского-Трофимова последовательности из $a \ge 0$ нулей и $b \ge 0$ единиц определяется как:

$$P_e\{a,b\} = \int_0^1 \frac{1}{\pi\sqrt{(1-\theta)\theta}} (1-\theta)^a \theta^b d\theta$$

При этом для $a \ge 0, b \ge 0$ верны формулы последовательного вычисления:

$$P_e(0,0) = 1, \; P_e(a+1,b) = \frac{a+0.5}{a+b+1}, P_e(a,b+1) = \frac{b+0.5}{a+b+1} \qquad (5)$$

Контекстное дерево T_D – это множество узлов, соответствующих бинарным строкам $s, 0 \le l(s) \le D$. В каждом узле $s \in \mathrm{T}_D$ хранится $a_s(x_1^t \mid x_{1-D}^0)$ – число нулей в последовательности x_1^t после суффикса s, $b_s(x_1^t \mid x_{1-D}^0)$ – число единиц.

Каждому узлу s ставится в соответствие взвешенная вероятность P_w^s:

$$P_w^s = \begin{cases} 0.5 P_e(a_s, b_s) + 0.5 P_w^{0s} P_w^{1s}, & 0 \le l(s) < D \\ P_e(a_s, b_s), & l(s) = D \end{cases} \qquad (6)$$

Таким образом, для определения вероятности $P_{CTW}(x_1^t \mid x_{1-D}^0)$ последовательности x_1^t требуется произвести подсчет частот a_s, b_s, $\forall s \in \mathrm{T}_D$, тогда

$$P_{CTW}(x_1^t \mid x_{1-D}^0) = P_w^{\lambda}(x_1^t \mid x_{1-D}^0) \qquad (7)$$

где λ – пустая строка.

Отметим, что (7) является взвешенной вероятностью по всем возможным моделям источника из C_D.

4 ИССЛЕДОВАНИЕ БИНАРНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ С ИСПОЛЬЗОВАНИЕМ CTW-ПРЕДИКТОРА

Прогноз значения следующего символа в момент времени t с использованием предиктора CTW строится согласно правилу (3) с использованием (7):

$$\hat{X}_{t+1} = \arg\max_{i=0,1}\{P_{CTW}(x_1^t \,\|\, i \,|\, x_{-D}^0)\} \qquad (8)$$

где символ $\|$ означает конкатенацию строк.

Таким образом, для построения последовательности индикаторов успеха прогнозов $\{Y_t\}$ по выборке $x_{-D}^0 \,\|\, x_1^n$ объема $n+D+1$ с использованием CTWпредиктора с параметром D в каждый момент времени $t=\overline{1,n}$ требуется выполнить следующие шаги:

1) обновить частоты a_s, b_s суффиксов $s_1 = x_{t-1}, \ldots, s_{D+1} = x_{t-D-1}\cdots x_{t-1}$;

2) пересчитать (5), (7) согласно новым значения частот;

3) построить прогноз (8) и вычислить индикатор Y_t.

Заметим, что (5) представляет собой последовательное произведение малых чисел, что вызывает вопрос точности вычислений для больших n. Тем не менее, модификация (6) и использование логарифмирования ([5]) позволяет эффективно реализовать вычисление вероятности последовательности (7) и построение прогнозов (8). Таким образом, CTW-предиктор, так же как и предикторы Лемпеля – Зива и SPM ([3], [4]) допускает эффективную реализацию и может быть использован в режиме прогнозирования on-line.

Приведем результаты вычислительного эксперимента по оценке мощности и уровня значимости критерия (4) на базе CTW-предиктора с параметром $D=1$ в случае марковской альтернативы 1-го порядка с дважды стохастической (д. с.) матрицей вероятностей одношаговых переходов (в.о.п.) $P = \begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix}$. Нетрудно видеть, что выбранный класс цепей Маркова с д. с. матрицей в.о.п. принадлежит однопараметрическому экспоненциальному семейству с параметром p, причем стационарное распределение числа нулей и единиц в последовательности равномерно. В этом случае равномерно наиболее мощным критерием для проверки гипотезы $H_0 : p = 0.5$ против двусторонней альтернативы $H_1 : p \neq 0.5$ будет критерий знакоперемен [7]. На рис. 1 приведена оценка мощности критерия на базе CTW-предиктора (○) и оценка мощности критерия знакоперемен (сплошная) для $n=\overline{10,500}$ $p=0.4$, число экспериментов 10^3. Можно видеть, что CTW-предиктор проигрывает в мощности оптимальному критерию знакоперемен. Это объясняется тем, что при построении прогнозов CTW-предиктор во взвешенной сумме (7), кроме модели цепи Маркова 1-го порядка учитывает так же возможность наличия модели источника без памяти.

Приведем результаты эксперимента с обучением предиктора по некоторой части выборки объема m, и последующим прогнозированием следующей части выборки объема n без обновления частот в контекстном дереве. Мощность критерия на базе предиктора CTW примет вид (◊) при $m=n=500$.

Оценка уровня значимости $\alpha = 0.05$ представлена (×).

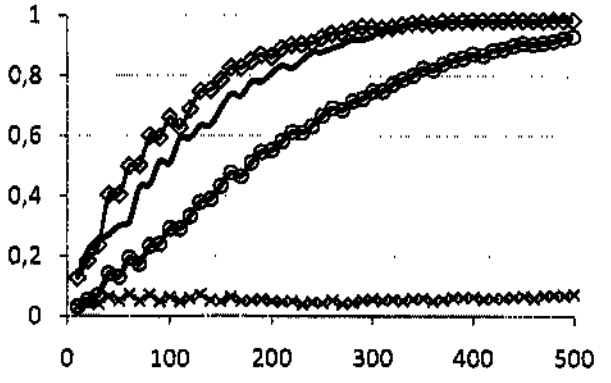

Рис.1. Сравнение оценок мощности критериев.

Из рис. 1 видно, что критерий на базе CTW-предиктора выдерживает заданный уровень значимости и является состоятельным.

5 ПРИМЕНЕНИЕ ПРЕДИКТОРА CTW ДЛЯ ПОИСКА РАЗЛАДКИ

В отличие от предикторов Лемпеля – Зива и SPM, CTW-предиктор допускает эффективную реализацию в оконном режиме. Это достигается за счет заранее известного параметра D (и, соответственно, вида контекстного дерева T_D), а также за счет того, что для пересчета модифицированной вероятности последовательности (6) требуется выполнить линейное от D число операций.

Рассмотрим следующую модель разладки. Пусть наблюдается выборка объемом $n=$ 20000 бит. Первые 5000 бит являются последовательностью н. о. р. с. в. Бернулли с вероятностью успеха $p=0.5+\varepsilon$, следующие 5000 бит являются последовательностью н. о. р. с. в. Бернулли с вероятностью успеха $p=0.5-\varepsilon$, затем 5000 бит является цепью Маркова 1-го порядка с д.с. матрицей в.о.п. с параметром $p=0.5+\varepsilon$, наконец, последние 5000 бит является цепью Маркова 1-го порядка с д.с. матрицей в.о.п. с параметром $p=0.5-\varepsilon$.

Можно видеть, что для каждой части такой последовательности существуют оптимальные критерии – критерий знаков [7] для первых двух фрагментов, критерий знакоперемен для последних. Однако принятие решение по всей выборке с использованием данных критериев приведет к мощности на уровне значимости. Поэтому для поиска такого рода разладки проведем эксперимент с использованием скользящего окна.

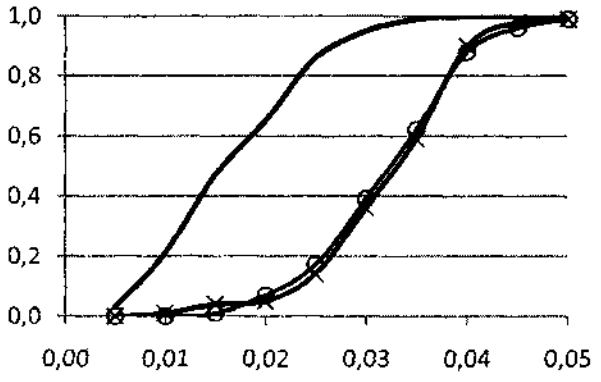

Рис.2. Сравнение оценок мощности для поиска разладки. Длина окна 2000.

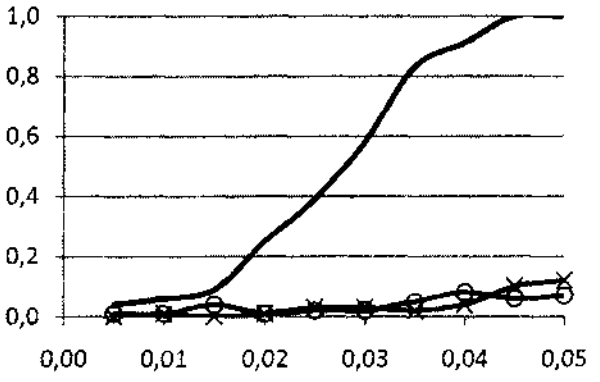

Рис.3. Сравнение оценок мощности для поиска разладки. Длина окна 200.

Пусть используется скользящее окно длины m по пересекающимся интервалам. Внутри окна ведется подсчет статистик критериев знаков, знакоперемен, принимается решение о принятии H_0 после чего окно сдвигается на единицу. Легко видеть, что каждый критерий применятся $n-m$ раз ($n=20000$), и, согласно теории множественной проверки гипотез, i-й тест должен быть использован с индивидуальным уровнем значимости $\alpha_i = \alpha/(n-m)$ Если гипотеза H_0 отвергается в одном из тестов, то принимается общее решение о том, что верна гипотеза H_1.

По наблюдаемой последовательности строится последовательность индикаторов успеха прогнозов с использованием предиктора CTW, который так же действует в оконном режиме (т.е. прогноз x_t строится по x_1^t, если $t \le m$, по x_{t-m+1}^t, иначе). Решение принимается по всей последовательности индикаторов длиной 20000.

На рис.2 приведены результаты оценки мощности в случае $m=2000, \varepsilon = 0.005, 0.01, ..., 0.05$. Результаты оценки мощности для $m=200, \varepsilon = 0.005, 0.01, ..., 0.05$ приведены на рисунке 3. Из рис. 2, 3 видно, что мощности критерия знаков (○) и критерия знакоперемен (×) эквивалентны, при этом использование предиктора (сплошная) в режиме скользящего окна дает значительный выигрыш в мощности.

ЛИТЕРАТУРА

[1] NIST Special Publication 800-22: A statistical test suite for random and pseudorandom number generators for cryptographic applications. – 2001.

[2] Kostevich A.L., Shilkin A.V. On Approach to Randomness Testing on the base of the Universal Predictors // Computer Data Analysis and Modeling: Complex Stochastic Data and Systems. Proceedings of the Eighth International Conference «Computer Data Analysis and Modeling: Complex Stochastic Data and Systems» – Minsk: Belarusian State University, 2007. – Vol. 1. – P. 256–259.

[3] Костевич А.Л., Шилкин А.В. Анализ эффективности применения универсальных предикторов для проверки гипотезы о чистой случайности // Материалы III Международной научной конференции «Сетевые компьютерные технологии» – Минск: издательский центр БГУ. – 2007. – С. 211-216

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

[5] F.M.J. Willems, Y.M. Shtarkov, Tj.J. Tjalkens, The Context Tree Weighting Method : Basic properties // IEEE Trans. on Inform. Theory – 1995, – P. 653-664.

[6] J. Suzuki. Universal prediction and universal coding // Systems and Computers, 2003. – Vol. 34(6). – P. 1–11.

[7] Боровков А.А. Математическая статистика. – М.: Наука, Главная редакция физико-математической литературы. · 1984. –472 с.