УДК 519.24 DOI: 10.24151/1561-5405-2020-25-2-175-182

О границах вероятностных аргументов при синтезе линейных сигнатур и статистических оценок

И.П. Кобяк

Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Беларусь

IPKobyak2012@mail.ru

Метод идентификации последовательностей счетом векторов состояний достаточно просто реализуется аппаратными средствами. Однако в последние годы основная масса статистических алгоритмов вытеснена из технических приложений так называемым сигнатурным анализом. Тем не менее существуют границы для вероятностных аргументов, в пределах которых один алгоритм имеет преимущество перед другим с точки зрения уровня пропуска ошибок. В представляемой работе для приложений математической статистики в задачах идентификации многомерных последовательностей сформирована методика определения границ аргументов, в пределах которых вероятность пропуска ошибки тем или иным методом меньше, чем у конкурирующего алгоритма. В качестве объектов сравнения выбраны алгоритмы формирования линейных сверточных кодов и аппаратного наблюдения *r*-разрядных векторов в стационарном и эргодическом случайном процессе. В результате выполненных исследований получены соотношения, позволяющие использовать для конкретной реализации процесса наиболее эффективный метод идентификации с точки зрения выборочной вероятности ошибки. Сфера применения результатов – хранение данных в компьютере или передача информации в канал связи.

Ключевые слова: вектор состояния; метод идентификации; сигнатурный анализ; формула Стирлинга; математическое ожидание; экстремум функции; линейная свертка

Для цитирования: Кобяк И.П. О границах вероятностных аргументов при синтезе линейных сигнатур и статистических оценок // Изв. вузов. Электроника. 2020. Т. 25. № 2. С. 175–182. DOI: 10.24151/1561-5405-2020-25-2-175-182

About Borders of Probabilistic Arguments at Synthesis Linear Signatures and Statistical Estimates

I.P. Kobyak

Belarusian State University of Informatics and Radioelectronics, Minsk, Belarusia

IPKobyak2012@mail.ru

Abstract: For applications of mathematical statistics in problems of identification of the multidimensional sequences the technique of delimitation of arguments within which the probability of the admission of a mistake by this or that method is less, than at the competing algorithm is created. As subjects to comparison algorithms of formation linear the svertochnykh of codes and hardware observation *r*- digit vectors in stationary and ergodic process are chosen. As a result of the executed researches the ratios allowing to use for concrete realization of process the most effective method of identification in terms of the selective probability of a mistake are received. Scope of results this data storage in the computer or information transfer in a communication channel.

Keywords: state vector; identification method; signature analysis; formula Stirlinga; population mean; function extremum; linear convolution.

For citation: Kobyak I.P. About borders of probabilistic arguments at synthesis of linear signatures and statistical estimates. *Proc. Univ. Electronics*, 2020, vol. 25, no. 2, pp. 175–182. DOI: 10.24151/1561-5405-2020-25-2-175-182

Введение. Принцип идентификации последовательностей счетом векторов состояний (СВС) хорошо известен и достаточно просто реализуется аппаратными средствами. Однако основная масса статистических алгоритмов в последние годы вытеснена из технических приложений методами линейной свертки, или так называемым сигнатурным анализом (СА) [1, 2]. Тем не менее исследование теоретических результатов в данной области показало, что при длине выборки *r*-разрядных векторов, равной *n*, графики плотности распределения вероятностей пропуска ошибки, характеризующие методы наблюдения векторов состояний и синтеза сигнатур, имеют общие точки. Иными словами, существуют границы для вероятностных аргументов, в пределах которых один алгоритм имеет преимущество перед другим с точки зрения уровня пропуска ошибки. Цель настоящей работы – исследование границ аргументов при формировании статистик СВС и сигнатур, в пределах которых тот или иной алгоритм имеет меньшее значение суммы составляющих интегральной вероятности, формируемых производящей функцией.

Постановка задачи и аналитическое решение. Пусть для статистики *k* вероятность пропуска ошибки методом СВС (*condition vector count, cvc*) определяется равенством

$$P_{cvc} = \frac{1}{m^n}\left[C_n^k (m-1)^{n-k} - 1\right], \quad (1)$$

где $m = 2^r$ – общее число видов векторов в выборке; C_n^k – биномиальный коэффициент; $k = \hat{p}(x_\omega)n$ – число наблюдаемых событий заданного вида x_ω; $\hat{p}(x_\omega)$ – выборочная вероятность.

Соотношение, аналогичное (1), при формировании линейной свертки МСА (*multichannel signature analysis, msa*) многоразрядных последовательностей имеет вид

$$P_{msa} = \frac{2^{rn-l} - 1}{m^n}, \quad (2)$$

где 2^{rn-l} – фактически число последовательностей, дающих одинаковую сигнатуру; $l = \log_2 rn$.

Определим границы интервалов вероятности $\hat{p}(x_\omega) = \frac{k}{n}$, в пределах которых один алгоритм предпочтительнее другого, формируя решение как функцию от статистического аргумента $k = k_0 \pm \Delta k$, $k_0 = \frac{n}{m}$ – точка экстремума функции (1) при $n = \infty$, т.е. с учетом отклонений аргументов Δk от математического ожидания. Постановка задачи с использованием соотношений (1) и (2) при этом будет сводиться к анализу гипотетического равенства

$$C_n^{k_0+\Delta k}(m-1)^{n-(k_0+\Delta k)} = 2^{rn-l}. \quad (3)$$

Совместное расположение графиков функций (1) и (2) на плоскости пропуска ошибки определяет два случая формирования общих точек. Так, при достаточно большом n и $k_0 \approx \frac{n}{m}$ для малых значений Δk графики функций имеют две общие точки, которые могут быть определены из равенства (3). При значительных отклонениях Δk от k_0 и $\Delta k > 2k_0$ существует одна общая точка (при $r > 1$).

Для решения поставленной задачи преобразуем левую часть соотношения (3), используя сокращенный вариант формулы Стирлинга [3, 4]:

$$n! = n^n\sqrt{2\pi n}.$$

Тогда

$$C_n^{k_0+\Delta k}(m-1)^{n-(k_0+\Delta k)} \approx \frac{(m-1)^{n-k_0-\Delta k}}{\left(\frac{1}{m}+\frac{\Delta k}{n}\right)^{k_0+\Delta k}\left(1-\frac{1}{m}-\frac{\Delta k}{n}\right)^{n-k_0-\Delta k}}\sqrt{\frac{n}{2\pi(k_0+\Delta k)(n-k_0-\Delta k)}}.$$

Прологарифмируем полученное равенство и заменим асимптотическое значение k_0 частным $\frac{n}{m}$. При этом, полагая, что $\ln\left(\frac{n}{m}+\Delta k\right) = \ln\frac{n}{m} + \ln\left(1+\frac{m\Delta k}{n}\right)$, получим

$$\ln\left[C_n^{k_0+\Delta k}(m-1)^{n-(k_0+\Delta k)}\right]=\left[n-\frac{n}{m}-\Delta k\right]\ln(m-1)-\frac{n}{m}\ln\left(1+\frac{m\Delta k}{n}\right)-\Delta k\ln\left(1+\frac{m\Delta k}{n}\right)+$$
$$+n\ln m-\left[n-\frac{n}{m}-\Delta k\right]\ln\left(m-1-\frac{m\Delta k}{n}\right)-\frac{1}{2}\ln 2\pi-\frac{1}{2}\ln\left(1+\frac{m\Delta k}{n}\right)+$$
$$+\ln m-\frac{1}{2}\ln n-\frac{1}{2}\ln\left(m-1-\frac{m\Delta k}{n}\right).$$

Упростим данное соотношение с учетом допущения $\Delta k \ll n$, $n\to\infty$:

$$n\ln m-\frac{1}{2}\ln 2\pi+\ln m-\frac{1}{2}\ln n-\frac{1}{2}\ln\left(1+\frac{m\Delta k}{n}\right)-\frac{1}{2}\ln\left(m-1-\frac{m\Delta k}{n}\right)\approx$$
$$\approx\ln\frac{m^{n+1}}{\sqrt{2\pi n}}-\frac{1}{2}\ln(m-1)\approx\ln\frac{m^{n+1}}{\sqrt{2\pi n(m-1)}}.$$

Отсюда

$$\ln\left[C_n^{k_0+\Delta k}(m-1)^{n-(k_0+\Delta k)}\right]\approx\ln\frac{m^{n+1}}{\sqrt{2\pi n(m-1)}}+\left(n-\frac{n}{m}-\Delta k\right)\ln(m-1)-\left(\frac{n}{m}+\Delta k\right)\times$$
$$\times\ln\left(1+\frac{m\Delta k}{n}\right)-\left(n-\frac{n}{m}-\Delta k\right)\ln\left(m-1-\frac{m\Delta k}{n}\right)\approx \qquad (4)$$
$$\approx\ln\frac{m^{n+1}}{\sqrt{2\pi n(m-1)}}-\left(\frac{n}{m}+\Delta k\right)\ln\left(1+\frac{m\Delta k}{n}\right)-\left(n-\frac{n}{m}-\Delta k\right)\ln\left(1-\frac{m\Delta k}{n(m-1)}\right).$$

Множители вида $\ln\left(1\pm\frac{m\Delta k}{n}\right)$, $\frac{m\Delta k}{n}<1$ заменим соответствующими рядами:

$$\ln\left(1+\frac{m\Delta k}{n}\right)\approx\sum_{\lambda=1}^{\infty}(-1)^{\lambda+1}\frac{1}{\lambda}\left(\frac{m\Delta k}{n}\right)^{\lambda}=\frac{m\Delta k}{n}-\frac{m^2\Delta k^2}{2n^2}+\frac{m^3\Delta k^3}{3n^3}-\ldots,$$
$$\ln\left[1-\frac{m\Delta k}{n(m-1)}\right]\approx\sum_{\lambda=1}^{\infty}-\frac{1}{\lambda}\left[\frac{m\Delta k}{n(m-1)}\right]^{\lambda}=-\frac{m\Delta k}{n(m-1)}-\frac{m^2\Delta k^2}{2n^2(m-1)^2}-\frac{m^3\Delta k^3}{3n^3(m-1)^3}-\ldots$$

Подстановка членов ряда в соотношение (4) и приведение подобных дает равенство вида

$$\ln\left[C_n^{k_0+\Delta k}(m-1)^{n-(k_0+\Delta k)}\right]\approx\ln\frac{m^{n+1}}{\sqrt{2\pi n(m-1)}}-\Delta k^2\frac{m}{2n}\left[\frac{(m-1)+1}{(m-1)}\right]+$$
$$+\Delta k^3\frac{m^2}{6n^2}\left[\frac{(m-1)^2-1}{(m-1)^2}\right]-\Delta k^4\frac{m^3}{3n^3}\left[\frac{(m-1)^3+1}{(m-1)^3}\right]+\ldots$$

В работе [4] показано, что в асимптотике погрешность аппроксимации логарифмических функций рядами незначительно увеличивается при использовании только двух первых членов. Таким образом, в худшем случае можно записать

$$C_n^{k_0\pm\Delta k}(m-1)^{n-(k_0\pm\Delta k)} \approx \frac{m^{n+1}}{\sqrt{2\pi n(m-1)}}\exp\left[-(\Delta k)^2\frac{m}{2n}\frac{m}{(m-1)}\right]$$

при $\Delta k = 1, 2, 3, \ldots, \Delta k << n$.

С учетом полученных преобразований постановка задачи поиска границ вероятностных аргументов в (3) принимает вид

$$\frac{m^{n+1}}{\sqrt{2\pi n(m-1)}}\exp\left[-(\Delta k)^2\frac{m}{2n}\frac{m}{(m-1)}\right] \approx 2^{rn-l}. \qquad (5)$$

Определим теперь точки пересечения двух графиков как отклонение от экстремума $\frac{n}{m}$, т.е. рассмотрим случай взаимного расположения графиков на плоскости пропуска ошибки при достаточно большом $n >> 1$, но не асимптотическом значении данного параметра.

Разделим гипотетическое равенство (5) на m^n и экспоненту левой части и прологарифмируем полученное соотношение. Тогда

$$\ln\frac{m}{\sqrt{2\pi n(m-1)}} \approx (\Delta k)^2\frac{m}{2n}\frac{m}{(m-1)} - l\ln 2.$$

Отсюда

$$(\Delta k)^2 \approx \frac{2n(m-1)}{m^2}\ln\frac{2^l m}{\sqrt{2\pi n(m-1)}}.$$

Тогда решениями общего уравнения (3) будут точки:

$$k \approx \frac{n}{m} \pm \sqrt{\frac{2n(m-1)}{m^2}\ln\frac{2^l m}{\sqrt{2\pi n(m-1)}}}. \qquad (6)$$

Иными словами, в диапазоне изменения аргумента в пределах

$$\frac{n}{m} - \Delta k \le k \le \frac{n}{m} + \Delta k \qquad (7)$$

использование линейной свертки предпочтительнее метода СВС. Для всех остальных значений аргумента статистический алгоритм точнее. Заметим, однако, что в (6) и (7) должно выполняться неравенство вида $\Delta k < \frac{n}{m}$, в противном случае графики функций будут иметь только одну общую точку справа от математического ожидания.

Используя левую часть равенства (5), определим значение функции в точке $k_0 - \Delta k = 0$. Следовательно, подставляя значение $\Delta k = \frac{n}{m}$, получаем

$$\frac{m^{n+1}}{\sqrt{2\pi n(m-1)}}\exp\left[-\frac{n^2}{m^2}\frac{m}{2n}\frac{m}{(m-1)}\right] \approx \frac{m^{n+1}}{\sqrt{2\pi n(m-1)}}\exp\left[-\frac{n}{2(m-1)}\right].$$

Однако при $k = 2k_0$ значение функции имеет вид

$$\frac{m^{n+1}}{\sqrt{2\pi n(m-1)}}\exp\left[-\frac{4n^2}{m^2}\frac{m}{2n}\frac{m}{(m-1)}\right] \approx \frac{m^{n+1}}{\sqrt{2\pi n(m-1)}}\exp\left[-\frac{2n}{m-1}\right].$$

Таким образом, при данном уровне аппроксимации график биномиального распределения в диапазоне аргумента $0 \le k \le 2k_0$ несимметричен относительно точки математического ожидания.

Метод снижения уровня вероятности ошибки при СА. Во многих практических задачах, решаемых разработчиками цифровой аппаратуры, методология СА в достаточно широком диапазоне аргумента k проигрывает методу СВС с точки зрения вероятности ошибки. В связи с этим далее будем учитывать возможность снижения уровня вероятности P_{msa} за счет увеличения длины сигнатурного регистра до значения $l + \chi$, определяющего равенство $P_{msa}=P_{cvc}$. При этом будем полагать, что распределение последовательностей по классам эквивалентностей для СА остается равномерным.

При данной постановке задачи имеет смысл рассматривать значения аргумента $k \ge 2np_0(x_\omega)$, так как данный диапазон вероятностей достаточно широк, а уровни P_{msa} и P_{cvc} в нем существенно различаются. Следовательно, из (5) при этом имеем

$$\frac{m^{n+1}}{\sqrt{2\pi n(m-1)}}\exp\left[-(\Delta k)^2\frac{m}{2n}\frac{m}{(m-1)}\right] \approx 2^{rn-(l+\chi)}. \qquad (8)$$

Выполнив ряд преобразований, из (8) получим

$$\chi \approx \frac{1}{\ln 2}\left[(\Delta k)^2\frac{m}{2n}\frac{m}{m-1}\right] - \frac{1}{\ln 2}\ln\frac{2^l m}{\sqrt{2\pi n(m-1)}}. \qquad (9)$$

Очевидно, что при $m = 2$ (что актуально для часто используемых бернуллиевских приложений) из (9) следует

$$\chi \approx \frac{1}{\ln 2}\left[\frac{2(\Delta k)^2}{n} - \ln\sqrt{\frac{2n}{\pi}}\right]. \qquad (10)$$

Из (10) находим, что расширение диапазона вероятностей в рамках постановки задачи (8) может быть реализовано начиная с отклонения значения k от корня k_0 (6), равного

$$\Delta k \approx \sqrt{\frac{n}{2}\ln\sqrt{\frac{2n}{\pi}}},$$

так как при этом значение $\chi = 0$.

Сравнение суммарных вероятностей в поддиапазонах аргумента $np(x_\omega)$. Рассматриваемый общий случай для Δk в соотношении (6) может быть использован в качестве базы для анализа полученных результатов с точки зрения суммарного числа значений функций, образующих классы перестановок оценок в поддиапазонах аргумента $np(x_\omega)$ (6).

Теорема. При биномиальном распределении последовательностей для

$$n \geq \frac{2\pi(m-1)}{r^2 m^2} e^9, \quad r \leq 4,$$

суммарная вероятность пропуска ошибки в пределах интервала значений (7) больше, чем сумма вероятностей, определенных за пределами указанного диапазона аргументов [5].

Доказательство. Обозначим сумму всех значений функции, принадлежащих диапазону $\frac{n}{m} \pm np(x_\omega)$ в (7), через S. Соответственно, сумма C, определяющая число реализаций выборки, принадлежащих вероятностям за пределами указанного интервала, будет равна $m^n - S$.

Для оценки отношения $\frac{C}{S}$ используем формулу среднеквадратического отклонения вида

$$\sigma_K = \sqrt{np_0(x_\omega)q_0(x_\omega)} = \sqrt{n}\sqrt{\frac{1}{m}\left(1-\frac{1}{m}\right)}.$$

Очевидно, что отклонение $3\sigma_K$ от математического ожидания не должно превышать величины $\frac{n}{m}$. В противном случае будет наблюдаться выход параметра в отрицательную область, не определенную для аргумента Δk. При этом имеем

$$\frac{n}{m} > 3\sqrt{n}\sqrt{\frac{1}{m}\left(1-\frac{1}{m}\right)},$$

откуда минимальное значение длины выборки будет равно

$$n > 9(m-1). \tag{11}$$

Если в качестве $3\sigma_K$ использовать значение Δk из формулы (6), то для выполнения условий сформулированной теоремы, очевидно, необходимо выполнение неравенства

$$\sqrt{\frac{2(m-1)}{m^2}\ln\frac{2^l m}{\sqrt{2\pi n(m-1)}}} \geq 3\sqrt{\frac{1}{m}\left(1-\frac{1}{m}\right)}.$$

Отсюда следует соотношение

$$2\ln\frac{2^l m}{\sqrt{2\pi n(m-1)}} \geq 9.$$

Для $n = \frac{2^l}{r}$ при $l = \log_2 rn$ имеем

$$n \geq \frac{2\pi(m-1)}{r^2 m^2} e^9. \qquad (12)$$

Подстановка значений $r \leq 4$ показывает, что соответствующие величины n в (12) удовлетворяют условию (11). Таким образом, сформулированная теорема доказана.

В частном случае, если $m = 2$, из (12) имеем $n > 0{,}5\pi e^9 \approx 12\,729$. При данной длине выборки отклонение от k_0 в (6) оказывается больше интервала $3\sigma_K$ со всеми вытекающими отсюда последствиями. Таким образом, уровень отношения $\frac{C}{S}$ зависит от длины выборки и $\frac{C}{S} < 1$ при равенстве n указанному в теореме значению.

Что касается больших значений разрядности последовательностей случайных событий, то для всех $r \geq 5$ прямая вероятности пропуска ошибки (2) проходит выше уровня функции, определяемого границами аргументов среднеквадратического отклонения. Соответственно, и преимущества сигнатурного интегрирования при этом сводятся к минимуму, однако в асимптотике сравнение алгоритмов интегрально оказывается в пользу метода СВС [1].

Заключение. Из анализа полученных результатов следует, что вероятность пропуска ошибки, определенная методом наблюдения заданного вектора в достаточно большой, но не асимптотической выборке при $r \leq 4$, определяет преимущество линейного сверточного кодирования перед статистической методологией. Однако для $r \geq 5$, а также с точки зрения поддиапазона аргументов $2np_0(x_\omega) \leq k \leq n$ преимущество метода СВС в области больших значений статистики возрастает весьма существенно.

Литература

1. ***Кобяк И.П.*** Сравнительная оценка достоверности методов сигнатурного анализа и счета состояний // Электронное моделирование. 1996. Т. 18. № 1. С. 58–62.

2. ***Бертсекас Д., Галлагер Р***. Сети передачи данных. М.: Мир, 1989. 544 с.

3. ***Риордан Дж.*** Комбинаторные тождества. М.: Наука, 1982. 255 с.

4. ***Яблонский С.В.*** Введение в дискретную математику. М.: Наука, 1986. 384 с.

5. ***Кобяк И.П.*** О границах вероятностных аргументов при синтезе линейных сигнатур и статистических аргументов // Информационные технологии и системы 2017: материалы междунар. науч. конф. (Беларусь, Минск, 25 окт. 2017 г.). Минск: БГУИР, 2017. С. 216–217.

Поступила в редакцию 18.11.2019 г.; после доработки 18.11.2019 г.; принята к публикации 28.01.2020 г.

Кобяк Игорь Петрович – кандидат технических наук, доцент кафедры электронных вычислительных машин Белорусского государственного университета информатики и радиоэлектроники (Беларусь, 220600, г. Минск, ул. П. Бровки, 6), IPKobyak2012@mail.ru

References

1. Kobyak I.P. Comparative assessment of reliability of methods of the signature analysis and account of states. *Elektronoye modelirovaniye = Engineering Simulation,* 1996, vol. 18, no. 1, pp. 58–62. (In Russian).

2. Bertsekas D., Gallager R. *Data transmission networks*. Moscow, World Publ., 1989. 544 p. (In Russian).

3. Riordan J. *Combinatory identities*. Moscow, Nauka Publ., 1982. 255 p. (In Russian).
4. Yablonsky S.V. *Introduction to discrete mathematics*. Moscow, Nauka Publ., 1986. 384 p. (In Russian).
5. Kobyak I.P. About borders of probabilistic arguments at synthesis of linear signatures and statistical arguments. Information technologies and systems 2017. *Proceeding of the International scientific conference.* Minsk, 2017, pp. 216–217. (In Russian).

Received 18.11.2019; Revised 18.11.2019; Accepted 28.01.2020.

Information about the author:

Igor P. Kobyak – Cand. Sci. (Eng.), Assoc. Prof. of the Electronic Computing Machines Department, Belarusian State University of Informatics and Radioelectronics (Belarus, 220600, Minsk, P. Brovki st., 6), IPKobyak2012@mail.ru