E. A. Timofeev, Balls in sequence spaces,
Model. Anal. Inform. Sist., 2012, Volume 19, Number 2, 109–114

https://www.mathnet.ru/eng/mais223

Модел. и анализ информ. систем. Т. **19**, № **2** (2012) 109–114

УДК 519.987

Шары в пространствах последовательностей

Тимофеев Е.А. [1]

Ярославский государственный университет им. П.Г. Демидова

e-mail: timofeevea@gmail.com

получена 14 января 2012 года

Ключевые слова: энтропия, непараметрическая оценка, шар, мера Бернулли

Предлагается новая метрика на пространстве правосторонних бесконечных последовательностей над конечным алфавитом. Введенная в задаче оценивания энтропии дискретных стационарных процессов, эта метрика обладает рядом интересных свойств. Например, мера шара является разрывной при любом двоично-рациональном значении $\log r$, где r – радиус шара.

Введение

Точность непараметрических оценок энтропии [5] зависит от выбора метрики на пространстве Ω правосторонних последовательностей символов из конечного алфавита. В этой работе вводится новая метрика ρ_1 (см. (3)), которая обладает следующими свойствами:

1. ρ_1 эквивалентна обычной метрике, опирающейся на номер первых несовпадающих символов (см. (2)).

2. Для симметричной меры Бернулли мера шара в Ω, как функция радиуса, разрывна на всюду плотном множестве.

3. Обратная функция к мере шара непрерывна и постоянна почти всюду.

Подчеркнем, что поиск новых метрик обусловлен следующим результатом, доказанным в [5]: степенная точность оценки достигается в том случае, когда усредненная по пространству обратная функция к мере шара является гладкой (кусочно непрерывно дифференцируемой).

[1]Работа выполнена при поддержке гранта Правительства РФ по постановлению №220, договор 11.G34.31.0053.

1. Оценки энтропии

В этом разделе приведем краткое описание задачи непараметрического оценивания энтропии стационарного процесса. Стационарный процесс будем представлять как инвариантную относительно сдвига меру μ на пространстве $\Omega = \mathcal{A}^{\mathbb{N}}$, точками которого являются бесконечные правосторонние последовательности символов из конечного алфавита $\mathcal{A}$, где $\mathbb{N} = \{1, 2, \dots\}$. Таким образом, случайную последовательность, порожденную мерой μ, будем рассматривать как точку в пространстве Ω распределенную по мере μ.

Пусть задана некоторая метрика $\rho(\boldsymbol{x}, \boldsymbol{y})$ на пространстве Ω, которая билипшицево эквивалентна метрике (2). Пусть заданы $n+1$ точек $\boldsymbol{\xi_0}, \dots, \boldsymbol{\xi_n}$ в пространстве Ω, которые независимы и распределены по мере μ. Тогда оценка энтропии [2] определяется следующим образом:

$$h_n = -\left[\frac{1}{(n+1)\log n}\sum_{j=0}^{n}\log\left(\min_{i:i\neq j}\rho(\boldsymbol{\xi_i}, \boldsymbol{\xi_j})\right)\right]^{-1}. \tag{1}$$

В [5] эта оценка модифицирована на усредненные логарифмы расстояний разности между k-й и $(k+1)$-й ближайшими точками.

Наиболее важной характеристикой оценки h_n является ее точность (эффективность) $\mathbf{E}(h_n - h)^2$, где h – энтропия, а n – число наблюдений.

Напомним, что

$$\mathbf{E}(h_n - h)^2 = \mathbf{D}h_n + (\mathbf{E}h_n - h)^2,$$

где $\mathbf{D}h_n$ – дисперсия оценки, а $\mathbf{E}(h_n - h)$ – смещение.

В [5] показано, что $\mathbf{D}h_n = O(n^{-c})$ для довольно широкого класса мер и метрик ($c > 0$ – некоторая константа). Для метрики (2) в [3] эта оценка улучшена до максимально возможного порядка — $\mathbf{D}h_n = O(n^{-1})$.

Как это часто бывает в дискретных задачах, нахождение смещения является намного более трудным. Так, для симметричных мер Бернулли в [4] показано, что смещение является периодической функцией с периодом пропорциональным $\log n$. Для марковских мер смещение имеет аналогичное поведение [6], если логарифмы переходных вероятностей рационально соизмеримы.

Итак, точность оценки энтропии зависит от свойств метрики.

2. Метрики в пространстве $\Omega = \mathcal{A}^{\mathbb{N}}$

Пусть $\Omega = \mathcal{A}^{\mathbb{N}}$ – пространство правосторонних последовательностей символов из конечного алфавита $\mathcal{A}$. Пусть $\boldsymbol{x} = (x_1, x_2, \dots)$ и $\boldsymbol{y} = (y_1, y_2, \dots)$ – две точки из Ω. Определим следующие две метрики на Ω:

$$\rho_0(\boldsymbol{x}, \boldsymbol{y}) = \theta^{-\min\{k: x_k \neq y_k\}}, \tag{2}$$

$$\rho_1(a\boldsymbol{x}, b\boldsymbol{y}) = \begin{cases} \theta^{-1}\rho_1(\boldsymbol{x}, \boldsymbol{y}), & a = b; \\ \theta^{-\lambda(-\log_\theta \rho_1(\boldsymbol{x}, \boldsymbol{y}))}, & a \neq b; \end{cases} \tag{3}$$

где $\theta > 1$ and $\lambda(t) = \min\{1, t/2\}$.

Подчеркнем, что метрика (3) билипшицево эквивалентна метрике (2), т.е.

$$\rho_0(\boldsymbol{x}, \boldsymbol{y}) \le \rho_1(\boldsymbol{x}, \boldsymbol{y}) \le \theta \rho_0(\boldsymbol{x}, \boldsymbol{y}). \tag{4}$$

Отметим, что согласно терминологии [1], ρ_1 является *слабой* метрикой (или near-metric), поскольку неравенство треугольника для нее выполняется с некоторой константой $C > 1$.

Поскольку каждая точка $\boldsymbol{x}$ имеет бесконечное число координат, то для прикладных вычислений нужно ограничить число координат, используемых при расчетах. Для этого определяется *усечение* метрики, которое использует только m первых координат точки.

Определим *усечение* $\rho^{(m)}$ метрики ρ следующим образом:

$$\begin{aligned} &\rho^{(0)}(\boldsymbol{x}, \boldsymbol{y}) = 1; \\ &\rho^{(m)}(a\boldsymbol{x}, b\boldsymbol{y}) = \begin{cases} \theta^{-1} \rho^{(m-1)}(\boldsymbol{x}, \boldsymbol{y}), & a = b; \\ \theta^{-\lambda(-\log_\theta \rho^{(m-1)}(\boldsymbol{x}, \boldsymbol{y}))}, & a \ne b. \end{cases} \end{aligned} \tag{5}$$

Для упрощения вычислений введем

$$\alpha(\boldsymbol{x}, \boldsymbol{y}) = -\log_\theta \rho_1(\boldsymbol{x}, \boldsymbol{y})), \quad \alpha^{(m)}(\boldsymbol{x}, \boldsymbol{y}) = -\log_\theta \rho_1^{(m)}(\boldsymbol{x}, \boldsymbol{y})). \tag{6}$$

$$\alpha^{(m)}(a\boldsymbol{x}, b\boldsymbol{y}) = \begin{cases} \alpha^{(m-1)}(\boldsymbol{x}, \boldsymbol{y}), & a = b; \\ \lambda(\alpha^{(m-1)}(\boldsymbol{x}, \boldsymbol{y})), & a \ne b. \end{cases} \tag{7}$$

Подставив (6) в (3), получим

$$\alpha^{(m)}(a\boldsymbol{x}, b\boldsymbol{y}) = \begin{cases} \alpha^{(m-1)}(\boldsymbol{x}, \boldsymbol{y}) + 1, & a = b; \\ \min\{1, \alpha^{(m-1)}(\boldsymbol{x}, \boldsymbol{y})/2\}, & a \ne b. \end{cases} \tag{8}$$

Подчеркнем, что $-1/\log_\theta \rho_0(\boldsymbol{x}, \boldsymbol{y}))$ также является метрикой на Ω, а $1/\alpha(\boldsymbol{x}, \boldsymbol{y})$ не является даже слабой метрикой.

3. Мера шара

Рассмотрим симметричную (с равновероятными символами) меру Бернулли μ на пространстве Ω с метрикой (3) и $\mathcal{A} = \{0, 1\}$. Обозначим открытый шар радиуса r с центром в точке x через $B(\boldsymbol{x}, r) = \{\boldsymbol{y} \in \Omega \ : \ \rho_1(\boldsymbol{x}, \boldsymbol{y}) < r\}$.

Определим функцию $f(t)$, положив

$$f(t) = \mu(B(\boldsymbol{x}, \theta^{-t})).$$

Пусть $\boldsymbol{\xi}$ – случайная точка в Ω распределенная по мере μ. Рассмотрим случайную величину

$$\eta = \alpha(\boldsymbol{\xi}, \boldsymbol{x}).$$

Заметим, что $f(t)$ и η, очевидно, не зависят от $\boldsymbol{x}$, и $1 - f(t)$ является функцией распределения случайной величины η.

Справедлива следующая

Лемма 1. *η – дискретная случайная величина, принимающая значения* $\dfrac{2m+1}{2^k}+n$, *и имеющая следующее распределение:*

$$\begin{aligned} &P\left(\eta = \tfrac{2m+1}{2^k} + n\right) = \tfrac{1}{6}2^{-k-s(m)-n}, \\ &\qquad m = 0,1,\ldots,2^{k-1}-1, k = 1,2,\ldots, n = 0,1,\ldots, \\ &P\left(\eta = n\right) = \tfrac{1}{3}2^{-n}, \quad n = 1,2,\ldots, \end{aligned} \tag{9}$$

где через $s(m)$ обозначается число '1' в двоичном разложении числа m.

Доказательство. Из определения (3) для симметричной меры Бернулли получаем рекуррентное соотношение для функции $f(t)$

$$\begin{cases} f(t) = \frac{1}{2}f(t-1), & t \geq 0, \\ f(t) = \frac{1}{2} + \frac{1}{2}f(2t), & 0 \leq t < 1. \end{cases}$$

На интервале $0 \leq t < 1$ функция $f(t)$ является самоподобной и для нее выполняется так называемое условие "открытых множеств":

$$\begin{cases} f(t) = \frac{1}{2} + \frac{1}{2}f(2t), & 0 \leq t < 1/2. \\ f(t) = \frac{1}{2} + \frac{1}{4}f(2t-1), & 1/2 \leq t < 1. \end{cases} \tag{10}$$

Отсюда получаем, что

$$f(1-0) = \frac{2}{3}, \quad P(\eta = 1) = \frac{1}{6}.$$

Для распределения случайной величины η из (10) получаем

$$P\left(\eta = \frac{2m+1}{2^k}\right) = \begin{cases} \frac{1}{2}P\left(\eta = \frac{2m+1}{2^{k-1}}\right), & m < 2^{k-2}; \\ \frac{1}{4}P\left(\eta = \frac{2m+1-2^{k-1}}{2^{k-1}}\right), & 2^{k-2} \leq m < 2^{k-1}; \\ \frac{1}{2}P\left(\eta = \frac{2m+1-2^k}{2^{k-1}}\right), & m \geq 2^{k-1}. \end{cases}$$

Решив эти рекуррентные уравнения, получим (9).

Для завершения доказательства осталось показать, что η принимает только двоично-рациональные значения. Для этого нужно доказать равенство

$$\sum_{n=1}^{\infty}\frac{1}{3}2^{-n} + \sum_{n=0}^{\infty}\sum_{k=1}^{\infty}\sum_{m=0}^{\infty}\frac{1}{6}2^{-k-s(m)-n} = 1.$$

Последнее вытекает из легко проверяемого тождества

$$\sum_{m=0}^{2^k-1} 2^{-s(m)} = \left(\frac{3}{2}\right)^k.$$

□

Из леммы 1 получаем следующее утверждение

Утверждение 1. *Функция $\mu(B(x,r))$ разрывна при каждом двоично-рациональном значении $\log_\theta r$.*

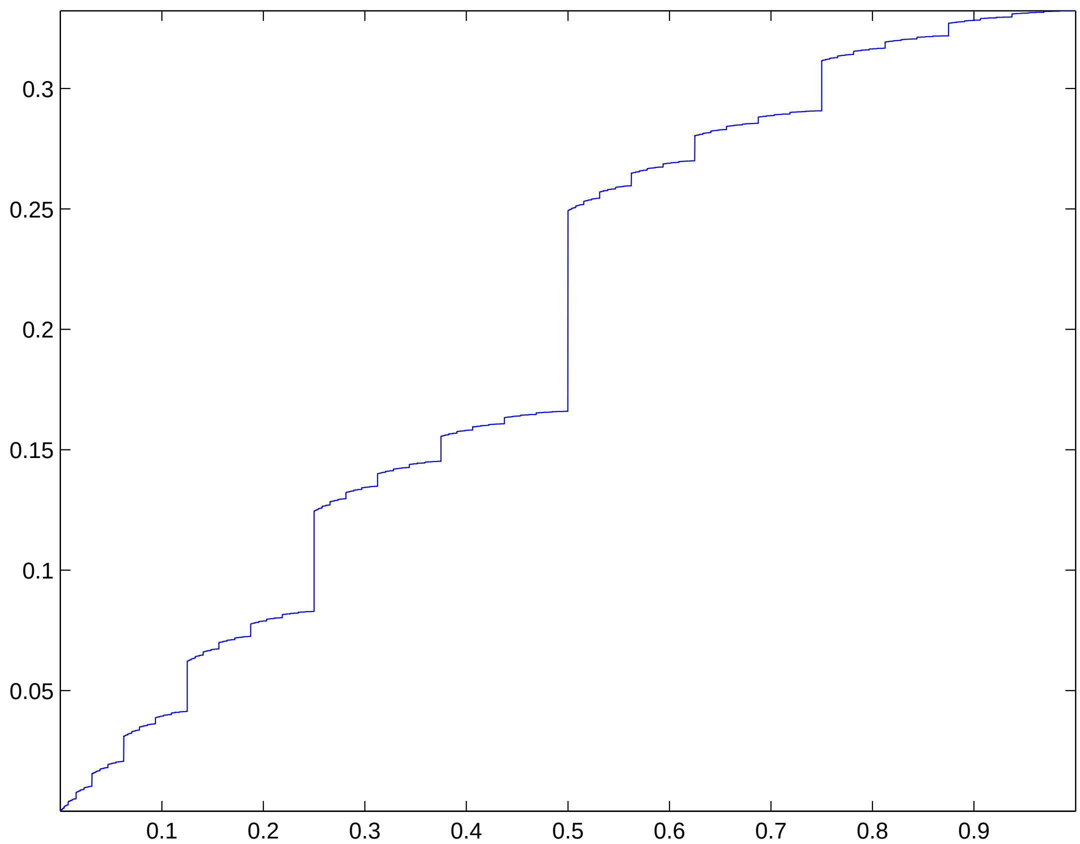

Рис. 1. Функция $1 - f(t) = 1 - \mu(B(x, \theta^{-t}))$ на интервале $0 \le t < 1$.

График функции $1 - f(t)$ на интервале $0 \le t < 1$ показан на рисунке 1. Отметим, что

$$f(t+1) = \frac{1}{2} f(t).$$

Определим обратную (обобщенную обратную) функцию $r = \nu(t)$ к функции $t = \mu(B(\boldsymbol{x}, r))$ следующим образом

$$\nu(t) \ = \sup\{r \ : \ \mu(B(\boldsymbol{x}, r)) < t\}. \tag{11}$$

Следует подчеркнуть, что участки постоянства функции $r = \nu(t, x)$ соответствуют разрывам функции $t = \mu(B(\boldsymbol{x}, r))$, поэтому справедливо следующее

Утверждение 2. *Функция $\nu(t, \boldsymbol{x})$ является непрерывной.*

Список литературы

[1] M. Deza, T. Deza, *Encyclopedia of Distances*, Springer, 2009.

[2] P. Grassberger, “Estimating the information content of symbol sequences and efficient codes”, *IEEE Trans. Inform. Theory*, **35** (1989), 669–675.

[3] A. Kaltchenko, N. Timofeeva, “Entropy Estimators with Almost Sure Convergence and an $O(n^{-1})$ Variance”, *Advances in Mathematics of Communications*, **2**:1 (2008), 1–13.

[4] A. Kaltchenko, N. Timofeeva, “Rate of convergence of the nearest neighbor entropy estimator”, *AEU – International Journal of Electronics and Communications*, **64**:1 (2010), 75–79.

[5] E. A. Timofeev, “Statistical Estimation of measure invariants”, *St. Petersburg Math. J.*, **17**:3 (2006), 527–551.

[6] E. A. Timofeev, “Bias of a nonparametric entropy estimator for Markov measures”, *Journal of Mathematical Sciences*, **176**:2 (2011), 255–269.

Balls in Sequence Spaces

Timofeev E.A.

Keywords: entropy, nonparametric statistic, ball, Bernoulli’s measure

We introduce a new metric on a space of right-sided infinite sequences drawn from a finite alphabet. Emerging from a problem of entropy estimation of a discrete stationary ergodic process, the metric is important on its own part and exhibits some interesting properties. For example, the measure of a ball is discontinuous at every binary rational value of $\log r$, where r is the radius.

Сведения об авторе:
Тимофеев Евгений Александрович,
Ярославский государственный университет им. П.Г. Демидова,
доктор физико-математических наук,
профессор кафедры теоретической информатики