Модел. и анализ информ. систем. Т. **20**, № **2** (2013) 178–185

UDC 519.987

Algorithm for Efficient Entropy Estimation

Timofeev E. A.[1]

P.G. Demidov Yaroslavl State University,
Sovetskaya str., 14, Yaroslavl, 150000, Russia

e-mail: TimofeevEA@gmail.com

received April 15, 2013

Keywords: entropy, nonparametric statistic, metric, ball, Bernoulli's measure

We consider the problem of the nonparametric entropy estimation of a stationary ergodic process. Our approach is based on the nearest-neighbor distances. We propose a broad class of metrics on the space $\Omega = A^{\mathbb{N}}$ of right-sided infinite sequences drawn from a finite alphabet A. The new metric has a parameter which is a non-increasing function. We apply this metrics to nearest-neighbor entropy estimators. We prove that, under certain conditions, the estimators has a small variance. We show that a special selection of the metric parameters reduction of the estimator's bias. The article is published in the author's wording.

Introduction

The paper studies the problem of the estimation of the entropy (entropy rate) of information sources with a finite state space. For our purposes, an information source, or a stationary process, is a shift-invariant ergodic measure μ on the space $\Omega = A^{\mathbb{N}}$ of right-sided infinite sequences drawn from a finite alphabet A, where $\mathbb{N} = \{1, 2, \dots\}$. Thus, an infinite random sequence generated by μ is viewed as a point in Ω chosen randomly with respect to μ.

We propose a broad class of metrics on Ω. The new metric has a parameter which is a non-increasing function. We apply this metric to the nearest-neighbor entropy estimator.

It is proved that, under certain conditions, the estimator has a small variance.

We describe a fast algorithm for finding the nearest-neighbor entropy estimator.

It is also proved that, under certain parameters of our metric, the estimator has a small bias.

We describe an effective algorithm for finding these parameters.

[1] Supported by the Russian government project 11.G34.31.0053

1. Problem Statement

Let $\Omega = A^{\mathbb{N}}$ and μ be a shift-invariant ergodic probability measure on Ω. Let $\boldsymbol{\xi_0}, \boldsymbol{\xi_1}, \ldots, \boldsymbol{\xi_n}$ be independent random variables taking values in Ω and identically distributed with a common law μ.

We want to evaluate the entropy of the measure μ.

In addition we impose the following restrictions on the measure:

$$\exists a, b > 0 \ : \ \mu(C_n(\boldsymbol{x})) \leq be^{-an}, \quad \forall n > 0, \quad a.e.\ \boldsymbol{x} \in \Omega, \tag{1}$$

where by

$$C_s(\boldsymbol{x}) = \{\boldsymbol{y} \in \Omega : y_1 = x_1, \ldots, y_s = x_s)\}$$

we denote cylinders in the space Ω.

Let $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots)$ be a point in Ω chosen randomly with respect to μ. Recall that the entropy h of a measure μ is defined as follows

$$h = -\lim_{n\to\infty} \frac{1}{n} \mathbf{E} \log \mu(C_n(\xi)), \tag{2}$$

here and throughout the paper, all logarithms are to base e, i.e., natural.

2. Metrics on $\Omega = A^{\mathbb{N}}$

In this section we introduce a wider class of so-called [1] *weak* metrics, for which the triangle inequality holds with some constant $C > 1$.

Let $\mathcal{A} = \{1, 2, \ldots, A\}$ and suppose that A is even.

Let $\boldsymbol{x} = (x_1, x_2, \ldots)$ be a point in $\Omega = \mathcal{A}^{\mathbb{N}}$, then by $a\boldsymbol{x}$ denote the point $(a, x_1, x_2, \ldots)$, $a \in \mathcal{A}$.

The class of metrics ρ is defined as follows:

$$\rho(\boldsymbol{x}, \boldsymbol{y}) = e^{-\alpha(\boldsymbol{x},\boldsymbol{y})}, \tag{3}$$

where the function $\alpha(\boldsymbol{x}, \boldsymbol{y})$ is defined as follows:

$$\alpha(a\boldsymbol{x}, b\boldsymbol{y}) = \begin{cases} \alpha(\boldsymbol{x}, \boldsymbol{y}) + 1, & a = b; \\ \lambda_{a,b}(\alpha(\boldsymbol{x}, \boldsymbol{y})), & a \neq b. \end{cases} \tag{4}$$

Auxiliary functions $\lambda_{a,b}(t)$ are non-decreasing and

$$0 \leq \lambda_{a,b}(t) = \lambda_{b,a}(t) \leq 1.$$

Moreover the family of functions $\lambda_{a,b}(t)$ is chosen such that the set of functions

$$S_a = \{\lambda_{a,1}(t), \ldots, \lambda_{a,a-1}(t), \lambda_{a,a+1}(t), \ldots, \lambda_{a,A}(t)\}$$

does not depend on a ($S_1 = S_2 = \cdots = S_A$) and $|S_a| = A - 1$.

In other words, the functions $\lambda_{a,b}(t)$ may be interpreted as the edge coloring of the complete graph with A vertices by $A - 1$ colors. Note that such coloring exists only for even A.

Take the following well-known coloring:

$$\begin{aligned}
&\lambda_1 = \lambda_{0,1} = \lambda_{2,A-1} = \lambda_{3,A-2} = \cdots = \lambda_{A/2,A/2+1};\\
&\ldots\\
&\lambda_k = \lambda_{0,k} = \lambda_{k-1,k+1} = \cdots = \lambda_{1,2k-1} = \lambda_{2k,A-1} = \cdots = \lambda_{A/2+k-1,A/2+k};\\
&\ldots\\
&\lambda_{A-1} = \lambda_{0,A-1} = \lambda_{1,A-2} = \lambda_{2,A-3} = \cdots = \lambda_{A/2-1,A/2}.
\end{aligned} \tag{5}$$

Take

$$\lambda_k(t) = \phi\left(\frac{k}{A} + \frac{1}{A}\phi^{-1}(t)\right), \quad k = 1, 2, \ldots, A-1, \tag{6}$$

where $\phi(t)$, is non-increasing defined on the interval $(0, 1]$ such that

$$\begin{aligned}
&\phi(1) = 0;\\
&\phi\left(\tfrac{t}{A}\right) = \phi(t) + 1.
\end{aligned} \tag{7}$$

We stress that $\phi(t)$ is an arbitrary function on the interval $\left(\frac{1}{A}\right., 1]$ (non-increasing and $0 \leq \phi(t) \leq 1$).

If $\phi(t) = 0$, $\frac{1}{A} < t \leq 1$, we obtain the well-known metric, and denote it by ρ_0.

If $\phi(t) = 1$, $\frac{1}{A} \leq t \leq 1$, we obtain the metric $\frac{1}{A}\rho_0$.

We stress also that the metric ρ with an arbitrary function $\phi(t)$ is bi-Lipschitz equivalent to the metric ρ_0, i.e.

$$e^{-1}\rho_0(\boldsymbol{x}, \boldsymbol{y}) \leq \rho(\boldsymbol{x}, \boldsymbol{y}) \leq \rho_0(\boldsymbol{x}, \boldsymbol{y}). \tag{8}$$

Therefore, ρ is a weak metric [1] (or near-metric), i.e. the triangle inequality holds with some constant $C > 1$.

While each point $\boldsymbol{x}$ has infinitely many coordinates, for any practical calculations, we need to limit the number of coordinates which are used for calculation. We make it by introducing a truncation of the metric that uses only the first m coordinates of the points.

We define $\rho^{(m)}$, *the truncation of the metric* ρ, as follows:

$$\rho^{(m)}(\boldsymbol{x}, \boldsymbol{y}) = e^{-\alpha^{(m)}(\boldsymbol{x},\boldsymbol{y})}, \tag{9}$$

where

$$\begin{aligned}
&\alpha^{(0)}(\boldsymbol{x}, \boldsymbol{y}) = 0;\\
&\alpha^{(m)}(a\boldsymbol{x}, b\boldsymbol{y}) = \begin{cases} \alpha^{(m-1)}(\boldsymbol{x}, \boldsymbol{y}) + 1, & a = b;\\ \lambda_{a,b}(\alpha^{(m-1)}(\boldsymbol{x}, \boldsymbol{y})), & a \neq b. \end{cases}
\end{aligned} \tag{10}$$

Proposition 1. *The set of values of* $\alpha^{(m)}(\boldsymbol{x}, \boldsymbol{y})$ *is*

$$\Phi_m = \left\{\phi\left(\frac{N}{A^m}\right), \ N = 1, 2, \ldots, A^m\right\}. \tag{11}$$

Proof. The proof is by induction on m.

For $m = 0$, there is nothing to prove.

Using definition (10), we get

$$\Phi_m = \bigcup_{k=1}^{A-1} \lambda_k(\Phi_{m-1}) \ \cup \{\Phi_{m-1} + 1\}\,. \tag{12}$$

By the inductive assumption and (6), we obtain

$$\bigcup_{k=1}^{A-1} \lambda_k(\Phi_{m-1}) = \left\{ \phi\left(\frac{k}{A} + \frac{N}{A^m}\right),\ N = 1, 2, \ldots, A^{m-1},\ k = 1, 2, \ldots, A-1 \right\}.$$

Using (7), we obtain

$$\{\Phi_{m-1} + 1\} = \left\{ \phi\left(\frac{N}{A^m}\right),\ N = 1, 2, \ldots, A^{m-1} \right\}.$$

Substituting in (12), we get (11). □

In all practical calculations we also need to limit the number of parameters that are used for finding the values of the function $\phi(t)$. We make it by introducing a truncation of the function $\phi(t)$.

We define $\phi_l(t)$, *the truncation of the function* $\phi(t)$, as follows:

$$\begin{array}{ll} \phi_l(t) = \phi\left(\frac{N}{A^l}\right), & \frac{N-1}{A^l} < t \le \frac{N}{A^l},\quad N = A^{l-1}+1, A^{l-1}+2, \ldots, A^l; \\ \phi_l\left(\frac{t}{A}\right) = \phi_l(t) + 1, & 0 < t \le 1. \end{array} \tag{13}$$

It should be stressed that the functions $x = \phi_l(t)$ and $t = \phi_l^{-1}(x)$ have intervals of constancy and discontinuities (a constancy interval of one function corresponds to a discontinuity of the other).

Corollary 1. *Let $\phi(t) = \phi_l(t)$, then the set of values of $\alpha^{(m)}(\boldsymbol{x}, \boldsymbol{y})$ is*

$$\bigcup_{k=0}^{m-l} \{\Phi_l + k\}\,.$$

3. Nearest Neighbor Entropy Estimator

Suppose $n+1$ points $\boldsymbol{\xi_0}, \ldots, \boldsymbol{\xi_n}$ are given by its first m coordinates. Therefore we must to use the truncation $\rho^{(m)}$ of the metric (3) - (7).

Suppose the truncation of the function (7) is given by the truncation $\phi_l(t)$, $l < m$, and $\phi_l(t)$ is given by parameters

$$\beta_i = \phi\left(\frac{A^{l-1} + i}{A^l}\right), \quad i = 1, 2, \ldots, A^l - A^{l-1} - 1. \tag{14}$$

Note that

$$1 \ge \beta_1 \ge \beta_2 \ge \cdots \ge \beta_{A^l - A^{l-1} - 1} \ge 0. \tag{15}$$

In addition to the points and metrics we use an auxiliary parameter k that serves to control the applicability of evaluation. The estimators obtained for different values of k are estimates of one and the same magnitude.

The nearest neighbor estimator $\eta_n^{(k)}(\rho^{(m)})$ of the inverse entropy $1/h$ is defined as follows [4]:

$$\eta_n^{(k)}(\rho^{(m)}) = k\left(r_n^{(k)}(\rho^{(m)}) - r_n^{(k+1)}(\rho^{(m)})\right), \tag{16}$$

where

$$r_n^{(k)}(\rho^{(m)}) = \frac{1}{n+1}\sum_{j=0}^{n} \max_{i:i\neq j}{}^{(k)} \alpha^{(m)}(\boldsymbol{\xi_i}, \boldsymbol{\xi_j}), \tag{17}$$

and $\max^{(k)}\{X_1, \ldots, X_N\} = X_k$, if $X_1 \geq X_2 \geq \cdots \geq X_N$.

Applying Corollary 1, we obtain

Corollary 2. *The estimator $\eta_n^{(k)}(\rho^{(m)})$ is a linear function of the parameters β_i, $i = 1, 2, \ldots, A^{l-1} - 1$.*

4. Properties of Estimator

In this section, we consider statistical properties of the estimator (16).

Proposition 2. *Let $\boldsymbol{\xi_0}$, …, $\boldsymbol{\xi_n}$ be $n+1$ independent points in the space Ω chosen randomly with respect to μ and $k = O(\log n)$, then*

$$\lim_{n\to\infty} \frac{\mathbf{E}r_n^{(k)}(\rho)}{\ln n} = \frac{1}{h}.$$

The proof follows from (8) and Proposition 8 in [4].

Proposition 3. *Suppose Condition* (1) *holds, then there exists a constant $c > 0$ such that*

$$\mathbf{E}r_n^{(k)}(\rho) - \mathbf{E}r_n^{(k)}(\rho^{(m)}) = O(n^{-1})$$

for $m > c\ln n$.

This Proposition can be proved just as in [2].

Using (6), we obviously have the following property:

$$1 > \lambda_1(t_1) > \lambda_2(t_2) > \cdots > \lambda_{A-1}(t_{A-1}), \quad \forall t_1, t_2, \ldots t_{A-1} \geq 0. \tag{18}$$

If we combine this property with Theorem 2 from [2], we get

Proposition 4. *There exists a constant C such that*

$$\mathbf{D}r_n^{(k)}(\rho^{(m)}) \leq C\frac{m^2k^2}{n}.$$

Thus, for determining the accuracy of $\eta_n^{(k)}(\rho^{(m)})$ we must found the bias. By Proposition 3 it is sufficient to find bias for $m = \infty$.

The application of Theorem 1 [4] yields the following

Proposition 5.

$$\mathbf{E}r_n^{(k)}(\rho) = \frac{n!}{(k-1)!(n-k)!}\int_0^1 \chi(t)\, t^{k-1}(1-t)^{n-k}\, dt, \tag{19}$$

where

$$\chi(t) = \int_\Omega \nu(t,\omega)\, d\mu(\omega), \tag{20}$$

and by $x = \nu(t,\omega)$ we denote the generalized inverse function of $t = \mu(B(\omega, e^{-x}))$, where by $B(\omega, r)$ we denote an open ball of radius r centered at x.

5. Determination Parameters of the Metric

We will choose the parameters of the metric so that to minimize the average deviation.

Since by Corollary 2 the estimator is a linear function of the parameters, we obtain a quadratic optimization problem on the simplex.

Problems involving estimation under linear inequality constraints arise often in statistics [3]. The traditional use for constrained estimation is for nonparametric regression with shape restrictions.

Thus, by Corollary 2, we get

$$\eta_n^{(k)}(\rho^{(m)}) = \eta_n^{(k)}(\rho_0^{(m)}) + \sum_{i=1}^{A^l - A^{l-1}-1} \beta_i R_{n,i}^{k,m}. \tag{21}$$

We will minimize the function

$$F(\beta) = \frac{1}{n-k+1}\sum_{j=k}^{n}\left(\eta_j^{(k)}(\rho^{(m)}) - \overline{\eta_j^{(k)}}(\rho^{(m)})\right)^2, \tag{22}$$

where

$$\overline{\eta_n^{(k)}}(\rho^{(m)}) = \frac{1}{n-k+1}\sum_{j=k}^{n}\eta_j^{(k)}(\rho^{(m)}). \tag{23}$$

Substituting (21) in (22), (23), we obtain that $F(\beta)$ is a quadratic form of its parameters. Therefore, the problem of minimization of this function is reduced to the problem of minimizing a quadratic form on the simplex (15).

6. Algorithm for Estimator Calculation

The algorithm consists of two stages.

1. Choose a part of the given strings.

   Find the coefficients in the linear expansion (21).

   Find the minimum of positive quadratic form (22) on the simplex (15).

2. Find the estimator (16) for the remainder part of the given strings with the parameters found in the first stage.

Since the property (18) holds, we can design a fast algorithm for the k-nearest neighbor search. The time complexity of the search is $O(m)$.

Let $\boldsymbol{X_1}, \ldots, \boldsymbol{X_N}$ be N words $\boldsymbol{X_i} = x_{i1} \ldots x_{im}$, $x_{ij} \in A$.

For a given word $\boldsymbol{X_0}$ and integer k, we want to find the following:

$$\max_{1 \le i \le n} {}^{(k)}\alpha^{(m)}(\boldsymbol{X_i}, \boldsymbol{X_0}),$$

The dictionary ($\boldsymbol{X_1}, \ldots, \boldsymbol{X_N}$) will be stored as a trie T. The root is associated with the empty string.

To each node i, we assign an auxiliary parameter $d(i)$ such that $d(i)$ is the number of descendant leaves in the subtrie with the root i. Thus, at each node i we store a list $C(i)$ of all children and a parameter $d(i)$.

Note that in our trie T all leaves are at level m. Thus, the required memory is $O(AN)$.

Proposition 6. *Suppose a k-nearest string belongs to the subtrie τ with the root r. Let r_0, r_1, ..., r_{A-1} be the children of r. We enumerate the node r_j by the index of the corresponding function λ_j (see* (5)*). Then the k-nearest string belongs to the subtrie with the root r_i such that*

$$\sum_{j=0}^{i} d(r_j) \ge k, \quad \sum_{j=0}^{i-1} d(r_j) < k.$$

7. Symmetric Bernoulli Measure

We show that for symmetric Bernoulli measures (with A equally probable symbols) we can find a function $\phi(t)$ of the metric (3) - (7) such that $\eta_n^{(k)}(\rho)$ is unbiased.

The function $\nu(t, \omega)$ is independent of the point ω for the symmetric Bernoulli measure. Therefore, $\chi(t) = \nu(t, \omega)$ and is defined by the following recursive equations:

$$\chi(t) = \begin{cases} \chi(At) + 1, & t \le \frac{1}{A}; \\ \phi\left(\frac{k}{A} + \frac{1}{A}\phi^{-1}(\chi(At-k))\right), & \frac{k}{A} < t \le \frac{k+1}{A}; \\ \quad k = 1, 2, \ldots, A-1. \end{cases} \tag{24}$$

It is easy to verify that the solution to these equations is $\chi(t) = \phi(t)$.

We claim that for

$$\chi(t) = \phi(t) = -\frac{\ln t}{\ln A}$$

$\eta_n^{(k)}(\rho)$ is unbiased.

Substituting $\chi(t)$ in (19), we obtain

$$\mathbf{E}r_n^{(k)}(\rho) = -\frac{n!}{(k-1)!(n-k)!\ln A}\int_0^1 \ln(t)\, t^{k-1}(1-t)^{n-k}\, dt.$$

Applying [5, 4.253.1], we get

$$\mathbf{E}r_n^{(k)}(\rho) = \frac{\psi(n+1) - \psi(k)}{\ln A}, \tag{25}$$

where $\psi(t) = \frac{d}{dt} \ln \Gamma(t)$ is the digamma function.

Hence,

$$\mathbf{E}\eta_n^{(k)}(\rho) = k\frac{\psi(k+1) - \psi(k)}{\ln A} = \frac{1}{\ln A}. \tag{26}$$

References

1. Deza M., Deza T. *Encyclopedia of Distances*, Springer, 2009.
2. *Kaltchenko A., Timofeeva N.* Entropy Estimators with Almost Sure Convergence and an $O(n^{-1})$ Variance //Advances in Mathematics of Communications. 2008. V. 2, №1. P. 1–13.
3. Silvapulle, M.J., Sen, P.K. Constrained statistical inference: Inequality, order and shape restrictions, John Wiley & Sons, USA. 2005.
4. Timofeev E.A. *Statistical Estimation of measure invariants* // St. Petersburg Math. J. 2006. **17**, №3. P. 527–551.
5. Градштейн И.С., Рыжик И.М. Таблицы интегралов, сумм, рядов и произведений. М.: Наука, 1971 (Gradshteyn I.S., Ryzhik I.M. Tablitsy integralov, summ, ryadov i proizvedeniy. Moskva: Nauka, 1971 [in Russian]).

Алгоритм эффективного оценивания энтропии

Тимофеев Е.А.

Ярославский государственный университет им. П. Г. Демидова
150000 Россия, г. Ярославль, ул. Советская, 14

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

Рассматривается задача непараметрического оценивания энтропии стационарного эргодического процесса. Применяется подход, основанный на нахождении расстояний до ближайших точек. Предложен довольно большой класс метрик на пространстве $\Omega = A^{\mathbb{N}}$ правосторонних бесконечных последовательностей над конечным алфавитом A. Новая метрика имеет параметр – невозрастающую функцию. Доказано, что при некоторых ограничениях предлагаемая оценка имеет малую дисперсию. Показано, что специальный выбор параметров позволяет уменьшить смещение. Описан алгоритм для выбора таких параметров. Статья публикуется в авторской редакции.

Сведения об авторе: **Тимофеев Евгений Александрович**,
Ярославский государственный университет им. П.Г. Демидова,
д-р физ.-мат. наук, профессор