Модел. и анализ информ. систем. Т. **20**, № **6** (2013) 174–178

УДК 519.987

Построение оценки энтропии для специальной метрики и произвольной функции

Тимофеева Н.Е.

НОУ ВПО "Институт управления", Ярославский филиал
Россия, г. Ярославль, ул. Некрасова, д. 52

e-mail: net0807@mail.ru

получена 24 октября 2013

Ключевые слова: энтропия, непараметрическая оценка, смещение, метрика, оптимизация

В статье предлагается обобщение оценки энтропии, предложенной в работе [1]. Для построения оценки сначала выбирается метрика на пространстве последовательностей. Эта метрика строится по матрице, которую можно интерпретировать как реберную раскраску полного графа с петлями. Обобщение состоит в том, что вместо логарифма в оценке энтропии применяется похожая функция, которая может быть произвольной на заданном интервале.

Предлагаемая функция не является монотонной, поэтому задача оптимизации среднего отклонения, которая является задачей квадратичной оптимизации, решается на всем пространстве, а не на симплексе.

Основные свойства оценки, такие как, асимптотическая несмещенность и степенное убывание дисперсии, доказываются аналогичным образом.

1. Постановка задачи

Обозначим через $\Omega = \mathcal{A}^{\mathbb{N}}$ пространство правосторонних бесконечных последовательностей символов из конечного алфавита $\mathcal{A}$.

Пусть даны $\boldsymbol{\xi_0}, \boldsymbol{\xi_1}, \ldots, \boldsymbol{\xi_n}$ – независимые случайные точки в Ω, одинаково распределенные по мере μ. Будем считать, что μ – инвариантная относительно сдвига эргодическая вероятностная мера на Ω.

Требуется оценить энтропию меры μ.

Напомним, что энтропия h меры μ определяется следующим образом:

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

где случайная точка $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots)$ распределена по мере μ, а через

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

обозначим цилиндры в пространстве Ω.

2. Метрика на пространстве последовательностей

Пусть $\mathcal{A} = \{0, 1, \ldots, A-1\}$ и будем считать, что A четно.

Обозначим через σ симметричную матрицу размером $A \times A$ с нулевыми элементами по диагонали ($\sigma_{i,i} = 0$)

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

Отметим, что каждая строка матрицы σ является перестановкой первой строки.

Другими словами, матрицу σ можно интерпретировать как реберную раскраску полного графа с петлями на A вершинах A цветами. Заметим, что такая раскраска существует только при четных A.

Определим метрику d_σ, положив

$$d_\sigma(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{\infty} \sigma_{x_i,y_i} A^{-i}. \tag{3}$$

Отметим, что метрика d_σ билипшицево эквивалентна метрике ρ_0, т.е.

$$\rho_0(\boldsymbol{x}, \boldsymbol{y}) \le d_\sigma(\boldsymbol{x}, \boldsymbol{y}) \le A\rho_0(\boldsymbol{x}, \boldsymbol{y}), \tag{4}$$

где метрика ρ_0 определяется как

$$\rho_0(\boldsymbol{x}, \boldsymbol{y}) = \max\{A^{-k} : x_k \ne y_k\}. \tag{5}$$

Следовательно, d_σ является слабой метрикой [2], т.е. неравенство треугольника выполняется только с некоторой константой.

Определим *сечение* метрики d_σ как

$$d_\sigma^{(m)}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{m} \sigma_{x_i,y_i} A^{-i} + A^{-m}. \tag{6}$$

3. Оценки энтропии

Пусть заданы:

1. $\boldsymbol{\xi_0}, \boldsymbol{\xi_1}, \ldots, \boldsymbol{\xi_n}$ – точки в пространстве Ω;
2. k – вспомогательный параметр, который служит для контроля применимости (оценки, полученные для различных значений k, являются оценками одной и той же величины);
3. $\phi(t)$ – функция, определенная на полуинтервале $(0, 1]$ такая, что

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

Оценка $\eta_n^{(k)}(d_\sigma, \phi)$ величины, обратной к энтропии $1/h$, определяется следующим образом:

$$\eta_n^{(k)}(d_\sigma, \phi) = \frac{k}{\ln A}\left(r_n^{(k)}(d_\sigma, \phi) - r_n^{(k+1)}(d_\sigma, \phi)\right), \tag{8}$$

где

$$r_n^{(k)}(d_\sigma, \phi) = \frac{1}{n+1}\sum_{j=0}^{n} \phi\left(\min_{i:i\neq j}{}^{(k)} d_\sigma(\boldsymbol{\xi_i}, \boldsymbol{\xi_j})\right), \tag{9}$$

и $\min^{(k)}\{X_1, \ldots, X_N\} = X_k$, если $X_1 \le X_2 \le \cdots \le X_N$.

Ясно, что оценка линейно зависит от функции $\phi(t)$. Отметим, что функция $\phi(t)$ является произвольной на интервале $\left(\frac{1}{A}, 1\right)$.

В прикладных вычислениях $n+1$ точек $\boldsymbol{\xi_0}, \ldots, \boldsymbol{\xi_n}$ заданы своими первыми m координатами, поэтому вместо метрики (6) нужно применять сечение $d_\sigma^{(m)}$.

4. Нахождение функции $\phi(t)$

Пусть l – некоторое целое число. Зададим функцию $\phi(t)$ $M = A^l - A^{l-1} - 1$ параметрами β_i, $i = 1, 2, \ldots, M$, положив

$$\phi(t) = \beta_i, \quad \frac{i + A^{l-1}}{A^l} < t \le \frac{i + A^{l-1} + 1}{A^l}, \quad i = 1, 2, \ldots, M. \tag{10}$$

Подчеркнем, что оценка $\eta_n^{(k)}(d_\sigma, \phi)$ и статистика $r_n^{(k)}(d_\sigma, \phi)$ являются линейными функциями от параметров β_i, $i = 1, 2, \ldots, M$.

Для параметрической функции $\phi(t)$ имеем

$$\phi\left(\min_{i:i\neq j}{}^{(k)} d_\sigma(\boldsymbol{\xi_i}, \boldsymbol{\xi_j})\right) = U_{i,k}\beta_{I_{i,k}} + V_{i,k},$$

где $U_{i,k}, V_{i,k}$ не зависят от β.

Подставляя в (8), получим

$$\eta_n^{(k)}(d_\sigma^{(m)}, \phi) = \frac{1}{n+1} D_{n,k} + \frac{1}{n+1}\sum_{s=1}^{M} C_{n,k,s}\beta_s. \tag{11}$$

Величины $D_{n,k}, C_{n,k,s}$ не зависят от β и задаются как

$$D_{n,k} = \frac{k}{(n+1)\ln A}\sum_{i=0}^{n} V_{i,k} - V_{i,k+1},$$

$$C_{n,k,s} = \frac{k}{(n+1)\ln A}\left(\sum_{i:I_{i,k}=s} U_{i,k} - \sum_{i:I_{i,k+1}=s} U_{i,k+1}\right).$$

Параметры β_i выберем такими, при которых достигается минимум среднеквадратичного отклонения

$$F(\beta) = \frac{1}{n-k+1}\sum_{j=k}^{n}\left(\eta_j^{(k)}(d_\sigma^{(m)}, \phi) - \overline{\eta_j^{(k)}}(d_\sigma^{(m)}, \phi)\right)^2, \tag{12}$$

где

$$\overline{\eta_n^{(k)}}(d_\sigma^{(m)},\phi)=\frac{1}{n-k+1}\sum_{j=k}^{n}\eta_j^{(k)}(d_\sigma^{(m)},\phi). \tag{13}$$

5. Алгоритм нахождения оценки энтропии

Алгоритм состоит из двух этапов.

1. Выбираем параметр $l<m$ и часть заданных точек $\boldsymbol{\xi_0},\boldsymbol{\xi_1},\ldots,\boldsymbol{\xi_{n_0}}$, $n_0<n$.

   Находим параметры β_i функции $\phi(t)$ (10), как описано в предыдущем разделе.

2. Находим оценку (8) энтропии по оставшимся заданным точкам с функцией $\phi(t)$, найденной на первом этапе.

Трудоемкость алгоритма равна $\mathcal{O}(mn^2)$ и не зависит от размера алфавита A.

Отметим, что при использовании экономных алгоритмов из работы [1] трудоемкость будет равна $\mathcal{O}(m^2An)$.

6. Свойства оценки

Нетрудно видеть, что метрика

$$\rho(\boldsymbol{x},\boldsymbol{y})=A^{-\phi(d_\sigma(\boldsymbol{x},\boldsymbol{y}))}$$

билипшицево эквивалентна метрике ρ_0:

$$A^{-\phi_1}\rho_0(\boldsymbol{x},\boldsymbol{y})\le\rho(\boldsymbol{x},\boldsymbol{y})\le A^{1-\phi_0}\rho_0(\boldsymbol{x},\boldsymbol{y}),$$

где

$$\phi_0=\min_{1/A\le t\le 1}\phi(t),\quad \phi_1=\max_{1/A\le t\le 1}\phi(t).$$

Следовательно, следующие леммы и утверждения доказываются так же, как в [1].

Параметр k будем считать фиксированным.

Лемма 1. *Пусть мера μ удовлетворяет условию*

$$\exists a,b>0\ :\ \mu(C_n(\boldsymbol{x}))\le be^{-an},\quad \forall n>0, \tag{14}$$

для почти всех $\boldsymbol{x}\in\Omega$, и пусть функция $\phi(t)$ задана параметрически (10).

Тогда существуют такие константы c_1,c_2, что верно следующее неравенство:

$$0<\mathbf{E}r_n^{(k)}(d_\sigma,\phi)(\rho)-\mathbf{E}r_n^{(k)}(d_\sigma^{(m)},\phi)\le c_1n^{-1},\ for\ m\ge l+c_2\ln n. \tag{15}$$

Утверждение 1. *Пусть $\boldsymbol{\xi_0},\ldots,\boldsymbol{\xi_n}$ – $n+1$ независимых точек в пространстве Ω, распределенных по мере μ, тогда*

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

Следствие 1. *Пусть выполнены условия леммы 1, тогда*

$$\lim_{n\to\infty} \frac{\mathbf{E} r_n^{(k)}(d_\sigma^{(m)}, \phi)}{\ln n} = \frac{1}{h}.$$

Аналогично доказательству теоремы 2 из [3] получаем

Утверждение 2. *Пусть $r_n^{(k)}(d_\sigma^{(m)}, \phi)$ – статистика, определенная в (9), тогда выполняется следующее неравенство*

$$\mathbf{D} r_n^{(k)}(d_\sigma^{(m)}, \phi) \le \frac{m^2(km+1)^2 A^2}{4(n+1)}. \tag{16}$$

Список литературы

1. Timofeev E.A. Algorithm for Efficient Entropy Estimation // Modeling and Analysis of Information Systems. T. 20, No 2. 2013. P. 112–119.

2. Deza M., Deza T. *Encyclopedia of Distances*. Springer, 2009.

3. *Kaltchenko A., Timofeeva N.* Entropy Estimators with Almost Sure Convergence and an $O(n^{-1})$ Variance // Advances in Mathematics of Communications. 2008. V. 2, No 1. P. 1–13.

Construction of an Entropy Estimator with a Special Metrics and an Arbitrary Function

Timofeeva N.E.

Institute of Managment,
150040, Yaroslavl, Nekrasova Str, 52

Keywords: entropy, nonparametric, optimization, bias, metrics

The paper proposes a generalization of entropy as in [1]. At first, to constract the estimator, we select the metrics on the space of sequances. This metrics is based on a matrix that can be interpreted as an edge coloring of a complete graph with loops. A generalization consists in that instead of using the logarithm in the estimation of the entropy, we apply a similar function which may be arbitrary at the given range. The proposed function is not monotone, so the task of optimizing the average deviation which is a quadratic optimization problem, is solved in the whole space and not on the simplex. The main properties of the estimator, such as asymptotic unbiasedness and power decrease dispersion, are proved in a similar way.

Сведения об авторе:
Тимофеева Нина Евгеньевна,
НОУ ВПО "Институт управления", Ярославский филиал,
канд. физ.-мат. наук, зав. кафедрой информатики