На правах рукописи

Тимофеева Нина Евгеньевна

ПОСТРОЕНИЕ ОЦЕНОК ЭНТРОПИИ СТАЦИОНАРНЫХ СЛУЧАЙНЫХ ПРОЦЕССОВ

Специальность 05.13.18 - математическое моделирование, численные методы и комплексы программ

АВТОРЕФЕРАТ

диссертации на соискание ученой степени кандидата физико-математических наук

Ярославль – 2010

Работа выполнена на кафедре теории функций и функционального анализа Ярославского государственного университета им. П.Г. Демидова

Научный руководитель –	доктор физико-математических наук, профессор Дольников Владимир Леонидович
Официальные оппоненты:	доктор физико-математических наук, профессор Соколов Валерий Анатольевич доктор физико-математических наук, доцент Райгородский Андрей Михайлович
Ведущая организация –	Учреждение Российской академии наук Институт программных систем имени А.К.Айламазяна РАН

Защита состоится "__" ____ ____ 2010 г. в ____часов на заседании диссертационного совета Д 212.002.05 при Ярославском государственном университете имени П.Г. Демидова по адресу: 150000, г. Ярославль, ул. Советская, д. 14.

С диссертацией можно ознакомиться в библиотеке Ярославского государственного университета имени П.Г. Демидова по адресу: 150000, г. Ярославль, ул. Полушкина роща, д. 1.

Автореферат разослан "____"_____ 2010 г.

Ученый секретарь диссертационного совета Глызин С.Д.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы.

Математическое понятие энтропии впервые ввел К.Шеннон в 1948 г. Он использовал его для измерения количества информации, показал, что энтропия определяет границу степени сжатия текста.

В 50-е гг. А.Н. Колмогоров перенес понятие энтропии на динамические системы и доказал, что энтропия есть инвариант динамической системы: изоморфные динамические системы имеют одинаковую энтропию ([11]). Он применил понятие энтропии для решения задачи об изоморфности сдвигов Бернулли.

В 1971 г. Д.Орнстейн доказал утверждение, обратное теореме Колмогорова: если энтропии сдвигов Бернулли одинаковы, то эти сдвиги изоморфны[12].

Эти исследования энтропии имеют важное теоретическое значение для математики, в частности, для информатики.

В прикладных задачах возникает проблема нахождения значения энтропии, поскольку она определяет границу сжатия информации. По экспериментальным данным можно только ценить энтропию. Существуют два основных подхода к построению оценок энтропии: построение эмпирической функции распределения и непараметрические оценки.

В первом подходе на прямую вычисляется математическое ожидание логарифма эмпирической функции распределения. Этот способ удобен, если число параметров, от которых зависит распределение, конечно. Впервые такой подход описан Г.П.Башариным[10].

Непараметрические оценки энтропии могут быть поделены на два больших класса. Первый класс использует алгоритмы сжатия данных Лемпеля-Зива[1]. Основываясь на результате Лемпеля-Зива, П.Грасбергер предложил свою первую оценку величины обратной к энтропии[2]. Используя результаты Д.С.Орнстейна и Б.Вейса, П.Шилдс доказал, что эта оценка не является сходящейся для общих эргодических процессов, но при наложении определенных условий будет состоятельной [9]. Основываясь на этом результате, И.Контояннис и Ю.М.Сухов показали состоятельность оценки для

более широкого класса стационарных эргодических процессов [3].

Второй класс оценок основан на "методе расстояния до ближайших соседей". Р.Л.Добрушин первым предложил оценку энтропии, использующую этот метод.[5]. Основная идея заключается в следующем: если ранее рассматривалась одна бесконечная последовательность, то при новом подходе рассматривается бесконечное число конечных последовательностей и вычисляется расстояние между ними. В.А.Ватутин и В.Г.Михайлов нашли смещение оценки Р.Л.Добрушина, оценили значение ее дисперсии и доказали состоятельность [6]. Позднее исследователи перенесли идею Р.Л.Добрушина на пространство случайных последовательностей. Опираясь на полученные результаты, Р.Бадии и А.Полити. предложили оценивать размерность Хаусдорфа в общем метрическом пространстве [7]. П.Биллингслей показал, что при выборе подходящей метрики энтропия совпадает с размерностью Хаусдорфа [8]. Основываясь на результатах Р.Бадии и А.Полити, П.Грассбергер ввел свою вторую оценку энтропии на пространстве случайных последовательностей [2]. П.Шилдс построил пример, который показывает, что оценка П.Грассбергера несостоятельна для общего эргодического процесса.Также он показал, что предлагаемая П.Грасбергером оценка состоятельна для неприводимых непериодических марковских цепей [9]. В работе В.В.Майорова и Е.А.Тимофеева было введено обобщение оценки П.Грасбергера[13].

Таким образом, теория энтропии является одной из активно развивающихся областей математики и находит применение во многих областях современной науки.

Цель работы. Построение новых статистических оценок энтропии. Исследование свойств этих оценок, а именно: смещенности, состоятельности, порядка убывания дисперсии. Построение эффективного алгоритма для вычисления энтропии динамических систем по их траекториям.

Методы исследования. В работе используются методы теории вероятности, теории динамических систем, теории информации, линейной алгебры.

Научная новизна. Основные результаты диссертации являются

новыми и состоят в следующем.

Построены три новые статистические оценки энтропии, основанные на методе "расстояния до ближайших соседей". Принципиальные преимущества новых оценок это: степенной порядок точности дисперсии, который уменьшен почти до границы Рао-Крамера; при оценивании энтропии динамических систем значения оценок не зависят от разбиения пространства (это подтверждается экспериментальными исследованиями); в ряде случаев показан степенной порядок убывания дисперсии.

Исследованы свойства этих оценок: смещенность, состоятельность. Доказано, что первая из предложенных оценок сходится μ-почти всюду. Описан прием уменьшения смещения оценки. Проведены вычислительные эксперименты, подтверждающие точность предложенных оценок.

В работе также приведен эффективный алгоритм вычисления оценки по экспериментальным данным.

Теоретическая и практическая ценность. Работа носит теоретический характер. Результаты диссертации могут быть полезны при изучении задач построения оценок энтропии, а также при изучении динамических систем.

Апробация диссертации. Основные диссертации докладывались на XI симпозиуме по проблемам избыточности в информационных и управляющих системах, г.Санкт-Петербург, 2007 г., на 45 ежегодной конференции университета штата Иллинойс США, 2007 г., на IEEE Information Theory Workshop, Lake Tahoe, California, USA, 2007 г., на Международной конференции "Компьютерные технологии в электротехнике и радиоэлектронике", Новосибирск, 2008г., на семинаре кафедры математического моделирования Ярославского государственного университета им. П.Г. Демидова.

Публикации. По теме диссертации опубликовано 10 научных работ, из них 5 в журналах из перечня ВАК. Список публикаций приведен в конце автореферата. Работы, написанные совместно с другими исследователями, выполнены в нераздельном соавторстве.

Структура диссертации. Диссертация состоит из введения и 4-х глав. Список литературы содержит 39 наименований. Общий объем

диссертации составляет 63 страницы.

СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Рассматривается пространство последовательностей $\Omega = \mathcal{A}^{\mathbb{N}}$. Пусть μ — эргодическая, инвариантная относительно сдвига мера на пространстве Ω. Пусть $\xi_1, \ldots, \xi_n$ — независимые, одинаково распределенные по мере μ случайные точки из Ω.

В прикладных задачах задано n слов длины m. В этом случае, считаем, что эти слова совпадают с первыми m символами точек $\xi_1, \ldots, \xi_n$ и будем обозначать эти слова через $\xi_1^{(m)}, \ldots, \xi_n^{(m)}$.

Требуется построить оценку величины обратной к энтропии. Эта замена энтропии на обратную величину сделана для удобства формулировки и доказательства теорем.

Глава 1.

В главе 1 на пространстве Ω рассматривается метрика

$$\rho_0(\mathbf{x}, \mathbf{y}) = \frac{1}{\min\{k : x_k \neq y_k\}}. \tag{1}$$

Предлагается следующая оценка:

$$\eta_n^{(k,m)} = \frac{1}{n \log n} \sum_{j=1}^{n} \left(\min_{i:i\neq j}{}^{(k)} \rho^{(m)}(\xi_i, \xi_j) \right)^{-1}, \tag{2}$$

где

$$\rho_0^{(m)}(\mathbf{x}, \mathbf{y}) = \max\left\{ \frac{1}{m}, \rho_0(\mathbf{x}, \mathbf{y}) \right\}. \tag{3}$$

Замечание. Подчеркнем, что предлагаемая оценка использует только первые m символов последовательностей $\xi_1, \ldots, \xi_n$.

Асимптотическая несмещенность оценки доказана в следующей теореме.

Теорема 1. *Пусть μ — инвариантная относительно сдвига борелевская вероятностная эргодическая мера, удовлетворяющая условию:*

$$\exists a > 0, b : \mu(C_s(\mathbf{x})) \leqslant e^{-as+b}, \quad \forall s > 0, \mathbf{x} \in \Omega \tag{4}$$

Тогда для всех $m \geqslant \frac{2}{a} \ln n$ *существует предел:*

$$\lim_{n\to\infty} \mathrm{E}\eta_n^{(k,m)} = \frac{1}{h}. \tag{5}$$

где $\eta_n^{(k,m)}$ *определена в (2).*

Отметим, что условия теоремы выполняются для гиббсовских мер.

По сравнению с ранее известным результатом П.Шилдса сходимость оценки (2) показана для более широкого класса мер.

Для любого $s \in \mathbb{R}$ и произвольной вещественной функции $\phi : \mathbb{R} \to \mathbb{R}$, определим ее конечную разность $\Delta\phi(s)$ как

$$\Delta\phi(s) = \phi(s) - \phi(s-1) \tag{6}$$

и k-ю разность как

$$\Delta^{(k)}\phi(s) = \Delta^{(k-1)}\phi(s) - \Delta^{(k-1)}\phi(s-1). \tag{7}$$

Обозначим:

$$\begin{aligned} \phi_m(n) &= \sum_{s=0}^{m-1} \left(1 - \int\limits_{\Omega} \left[1 - \mu(C_s(x))\right]^n \, d\mu(x)\right), \\ \phi(n) &= \sum_{s=0}^{\infty} \left(1 - \int\limits_{\Omega} \left[1 - \mu(C_s(x))\right]^n \, d\mu(x)\right). \end{aligned} \tag{8}$$

Лемма 1.

$$\mathrm{E}\eta_{n+1}^{(k,m)} = \frac{1}{\log n} \sum_{j=0}^{k-1} (-1)^j \binom{n}{j} \Delta^{(j)}\phi_m(n). \tag{9}$$

Лемма 2. *Пусть мера* μ *удовлетворяет условию (4). Тогда, для* $m \geq \frac{2}{a} \ln n$ *и заданного* k*, существует такая константа* C*, что следующее неравенство выполнено:*

$$\mathrm{E}\eta_{n+1}^{(k,\infty)} - \mathrm{E}\eta_{n+1}^{(k,m)} \leqslant Cn^{-1}.$$

Неравенство для параметра m, полученное в этой лемме, ранее было доказано П.Шилдсом для мер с условием свободной энергии. Условие леммы 2 является более широким.

В следующей теореме найдено ограничение сверху для дисперсии оценки.

Теорема 2. *Пусть μ удовлетворяет условиям теоремы 1. Тогда выполнено следующее неравенство:*

$$\mathrm{D}\eta_n^{(k,m)} \leqslant \frac{(2k-1)^2(m-1)^2}{4n(\log n)^2} \tag{10}$$

где $\eta_n^{(k,m)}$ определена в (2).

Замечание. Для $m = \mathcal{O}(\log n)$ и фиксированного параметра k оценка (10) равняется $\mathcal{O}(n^{-1})$.

Полученное неравенство для дисперсии в теореме 2 при $m = \mathcal{O}(\log n)$, улучшает ранее известный результат — $\mathcal{O}(n^{-c})$. Более того, полученный порядок убывания оценки по параметру n совпадает с порядком классической нижней границы Рао-Крамера для дисперсии.

Таким образом, предложенная в главе 1 оценка является наилучшей в настоящий момент по порядку убывания дисперсии по n и этот порядок, скорее всего, неулучшаем.

Из теорем 1 и 2 легко выводится состоятельность оценки.

Сходимость почти всюду была доказана П.Шилдсом [9] для марковских мер, а в данной работе получена сходимость для более широкого класса мер.

Теорема 3. *Пусть мера μ удовлетворяет условиям теоремы 1. Тогда оценка $\eta_n^{(k,m)}$, определенная в (2), сходится к $\frac{1}{h}$ μ-почти всюду.*

В работе проведен эксперимент для нескольких динамических систем, энтропия которых известна. Результаты эксперимента подтверждают эффективность оценки (2) для нахождения энтропии.

Глава 2. На пространстве Ω рассматривается метрика

$$\rho_1(\boldsymbol{x},\boldsymbol{y}) = \frac{\theta-1}{|\mathcal{A}|-1}\left|\sum_{k=1}^{\infty}\theta^{-k}x_k - \sum_{k=1}^{\infty}\theta^{-k}y_k\right|, \tag{11}$$

где параметр $\theta \geq |\mathcal{A}|$, $|\mathcal{A}|$ — число символов алфавита $\mathcal{A}$.

Построена следующая оценка энтропии:

$$\eta_n^{(k,m)} = \frac{k\left(r_n^{(k,m)} - r_n^{(k+1,m)}\right)}{\ln\theta}, \tag{12}$$

где

$$r_n^{(k,m)} = -\frac{1}{(n+1)}\sum_{j=1}^{n}\ln\left(\min_{i:i\neq j}{}^{(k)}\rho_1^{(m)}(\boldsymbol{\xi_i},\boldsymbol{\xi_j})\right), \tag{13}$$

метрика ρ_1 определена по формуле (11) и через $\rho_1^{(m)}(\boldsymbol{x},\boldsymbol{y})$ обозначается усечение метрики:

$$\rho_1^{(m)}(\boldsymbol{x},\boldsymbol{y}) = \max\left\{\frac{\theta-1}{|\mathcal{A}|-1}\theta^{-m}, \rho_1(\boldsymbol{x},\boldsymbol{y})\right\}. \tag{14}$$

Так как

$$\lim_{\theta\to\infty}\frac{\ln\rho_1(\boldsymbol{x},\boldsymbol{y})}{\ln\theta} = \rho_0(\boldsymbol{x},\boldsymbol{y})^{-1},$$

то метрику ρ_0 можно рассматривать как частный случай метрики ρ_1 при $\theta = \infty$. Поэтому при $\theta \to \infty$ оценка (12) совпадает с оценкой, рассмотренной в главе 1.

Подчеркнем, что предлагаемая оценка (12) использует только m первых координат заданных случайных точек Ω. Эта естественная модификация оценки из работы [14] позволила улучшить оценку дисперсии.

Через λ обозначим отображение $\lambda\ :\ \Omega \to [0,1]$

$$\lambda(\boldsymbol{\omega}) = \frac{\theta-1}{|\mathcal{A}|-1}\sum_{k=1}^{\infty}\theta^{-k}\omega_k. \tag{15}$$

При $\theta > |\mathcal{A}|$ это отображение будет вложением и изоморфизмом компакта Ω и компакта $\lambda(\Omega) \subset [0,1]$. При $\theta = |\mathcal{A}|$ это отображение будет метрическим изоморфизмом.

Отображение λ переносит меру μ на $\lambda(\Omega)$. Распространим полученную меру на весь отрезок $[0,1]$ и обозначим через $\widehat{\mu}(x)$ соответствующую меру отрезка $[0,x]$ (функцию распределения). Будем считать, что $\widehat{\mu}(x) = 0$ при $x < 0$, и $\widehat{\mu}(x) = 1$ при $x > 1$.

Пусть

$$\beta(x,r) = \widehat{\mu}(x+r) - \widehat{\mu}(x-r), \quad 0 \le r \le 1, \ -1 \le x \le 2. \tag{16}$$

Ясно, что

$$\beta(x,r) = \mu\left(\{\boldsymbol{\omega} \in \Omega \ : \ |\lambda(\boldsymbol{\omega})) - x| < r\}\right), \quad 0 \le r \le 1, \ -1 \le x \le 2. \tag{17}$$

Через $r = \nu(x,t)$ обозначим обратную (по r) функцию к функции $t = \beta(x,r)$.

Положим

$$\chi(t) = -\int_0^1 \ln \nu(x,t)\, d\widehat{\mu}(x). \tag{18}$$

Через $\preccurlyeq$ будем обозначать лексикографический порядок на пространстве Ω.

Нетрудно видеть, что метрика ρ_1 согласована с порядком, т.е.

$$\rho_1(\boldsymbol{x},\boldsymbol{y}) \le \rho_1(\boldsymbol{x},\boldsymbol{z}) \quad \forall \boldsymbol{x} \preccurlyeq \boldsymbol{y} \preccurlyeq \boldsymbol{z}. \tag{19}$$

Обозначим $\mathrm{E}\, r_n^{(k)} = \mathrm{E}\, r_n^{(k,\infty)}$, $\mathrm{E}\eta_n^{(k)} = \mathrm{E}\eta_n^{(k,\infty)}$.

Лемма 3.

$$\mathrm{E}r_n^{(k)} = \int_0^1 \int_0^1 \left[1 - \sum_{j=0}^{k-1} \binom{n}{j} \beta(x,u)^j \left(1 - \beta(x,u)\right)^{n-j}\right] d\widehat{\mu}(x) \frac{du}{u}. \tag{20}$$

Лемма 4.

$$\mathrm{E}r_n^{(k,m)} = \int_{\varepsilon_m}^1 \int_0^1 \left[1 - \sum_{j=0}^{k-1} \binom{n}{j} \beta(x,u)^j \left(1-\beta(x,u)\right)^{n-j}\right] d\widehat{\mu}(x) \frac{du}{u}, \tag{21}$$

где

$$\varepsilon_m = \frac{\theta - 1}{|\mathcal{A}| - 1}\, \theta^{-m}.$$

Лемма 5. *Пусть мера μ удовлетворяет условию*

$$\lim_{t\to 0} t \ln \nu(t,x) = 0, \ \forall x, \tag{22}$$

тогда

$$\mathrm{E}r_n^{(k)} = k\binom{n}{k} \int_0^1 \chi(t) t^{k-1} (1-t)^{n-k} dt. \tag{23}$$

$$\mathrm{E}\eta_n^{(k)} = \frac{k}{\ln \theta} \binom{n}{k} \int_0^1 \int_0^1 \beta(x,u)^k \left(1-\beta(x,u)\right)^{n-k} \, d\widehat{\mu}(x) \frac{du}{u}. \tag{24}$$

Теорема 4. *Пусть μ – инвариантная относительно сдвига борелевская вероятностная мера и $\xi_0, \ldots, \xi_n$ – независимые случайные точки в Ω распределенные по мере μ, тогда справедливо неравенство*

$$\mathrm{D}\, r_n^{(k,m)} \le \frac{(2k-1)^2 m^2}{4(n+1)} \ln^2 \theta, \tag{25}$$

где $r_n^{(k,m)}$ определено в (13).

Следствие 1. *Пусть μ – инвариантная относительно сдвига борелевская вероятностная мера и $\xi_0, \ldots, \xi_n$ – независимые случайные точки в Ω распределенные по мере μ, тогда справедливо неравенство*

$$\mathrm{D}\eta_n^{(k,m)} \le \frac{4k^4 m^2}{n+1}. \tag{26}$$

Правило выбора параметра m установлено в следующих утверждениях.

Лемма 6. *Пусть μ – инвариантная относительно сдвига борелевская вероятностная мера, удовлетворяющая условию*

$$\exists a > 0, C > 0 \ : \ \beta(x,t) \le Ct^a, \quad \forall t > 0, x \in [0,1]. \tag{27}$$

Тогда существует константа C_1 такая, что:

$$\mathrm{E}\, r_n^{(k,\infty)} - \mathrm{E}\, r_n^{(k,m)} \le C_1 n^k \theta^{-mk}.$$

Следствие 2. *Пусть μ – инвариантная относительно сдвига борелевская вероятностная мера, удовлетворяющая условию (27), тогда для $m \ge \frac{2}{a \ln \theta} \ln n$ и фиксированного k, существует константа C_2 такая, что:*

$$\mathrm{E}\eta_n^{(k,\infty)} - \mathrm{E}\eta_n^{(k,m)} \le C_2 n^{-1}.$$

Доказана сходимость оценки (12) почти всюду.

Теорема 5. *Пусть μ – инвариантная относительно сдвига борелевская вероятностная мера, удовлетворяющая условию (27), тогда,*

$$\lim_{n\to\infty} \left[\eta_n^{(k,m)} - \mathrm{E}\eta_n^{(k,\infty)} \right] = 0 \ \textit{почти всюду}$$

при $\frac{2}{a \ln \theta} \ln n < m \le C \ln n$, где C – некоторая константа.

Для симметричной меры Бернулли смещение оценки (12) является степенным по n.

$$\mathrm{E}\eta_n^{(k,\infty)} = \frac{1}{\ln |\mathcal{A}|} \left(1 - \frac{2 \ln 2 - 1}{n+1} \right), \tag{28}$$

т.е. при $n \to \infty$ математическое ожидание оценки стремится к $1/h$ ($h = |\ln \mathcal{A}|$ для симметрической меры Бернулли).

Итак, скорость сходимости оценки (12) для симметричной меры Бернулли при $\theta = |\mathcal{A}|$ равна $\mathcal{O}(n^{-1})$ при $m = \mathcal{O}(\ln n)$.

Таким образом, предложенная в главе 2 оценка энтропии для этой меры лучше, чем оценка из первой главы, у которой смещение равно $\mathcal{O}(1)$. Ее недостаток, по сравнению с оценкой (2) первой главы, — не доказан факт сходимости к $1/h$ почти всюду.

Глава 3.

В этой главе описан общий подход к построению эвристических оценок энтропии и исследованы их свойства.

Опишем основную идею предлагаемого подхода. Исследованная в главе 1 оценка содержит целочисленный параметр k. Новые оценки энтропии строятся как линейная комбинация этих оценок при различных значениях параметра k. Коэффициенты линейной комбинации подбираются так, чтобы формальный ряд по n смещения оценки имел заданное число нулевых случайных слагаемых.

Этот подход справедлив не только для метрики ρ_0, но и для произвольной метрики ρ, поэтому рассматривается произвольная метрика ρ на пространстве последовательностей.

Исследуемая оценка задается формулой

$$\eta_n^{(k,m)} = \sum_{i=1}^{k} a_i r_n^{(i,m)}, \tag{29}$$

где

$$a_i \triangleq (-1)^{i-1}(k-1)\binom{k-1}{i-1}, \tag{30}$$

а $r_n^{(k,m)}$ определена в формуле (13) и для простоты будем считать, что $m = \infty$. Для краткости будем обозначать $\eta_n^{(k)} = \eta_n^{(k,m)}$, $r_n^{(k)} = r_n^{(k,m)}$.

Выбор коэффициентов a_i объясняется в следующем утверждении.

Утверждение 1. *Пусть величина* $\mathrm{E}\eta_n^{(k)}$ *представлена формальным рядом по* n^{-1} *и коэффициенты* a_i *в формуле (29) определены согласно формуле (30). Тогда формальный ряд для* $\mathrm{E}\eta_n^{(k)}$ *имеет вид:*

$$\mathrm{E}\eta_n^{(k)} = \frac{1}{h} + \psi_{k-1} n^{-k+1} + \psi_k n^{-k} + \dots.$$

Положим

$$\phi(n) = \mathrm{E} r_n^{(1)}. \tag{31}$$

Сходимость оценки (29) показана в следующей теореме.

Теорема 6. *Предположим, что существуют константы $h > 0$ и $a > 0$ такие, что для некоторого $k = K_0 > 0$ верно*

$$\Delta^{(k)}\phi(n) = (-1)^{k-1}\frac{(k-1)!(n-k)!}{hn!}(1 + \mathcal{O}(n^{-a})). \tag{32}$$

Тогда константа h из (32) равна энтропии меры μ для всех $k < K_0$ и верно следующее равенство:

$$\lim_{n\to\infty} \mathrm{E}\eta_n^{(k)} = \frac{1}{h}.$$

Доказательство теоремы основано на двух, интересных на наш взгляд, леммах.

Лемма 7. *Если условие (32) выполнено для некоторого $k = k_0 > 0$, то оно выполнено также для $0 < k \leq k_0$ и*

$$\phi(n) = \frac{H_n}{h} + C_0 + \mathcal{O}(n^{-a}), \tag{33}$$

где

$$C_0 = \sum_{s=1}^{\infty}\left(\Delta\phi(s) - \frac{1}{hs}\right).$$

Лемма 8. *Предположим, что существуют константы $h > 0$, $a > 0$ и $C > 0$ такие, что*

$$\Delta^{(k)}\phi(n) = \frac{(-1)^{k-1}}{hk}\binom{n}{k}^{-1}(1 + \psi_k(n)), \tag{34}$$

выполнено для некоторого $k = k_0 > 0$, где

$$|\psi_k(n)| \leq Cn^{-a}. \tag{35}$$

Тогда условия (34) и (35) верны для всех $0 < k \leq k_0$ *с некоторой константой* C *и справеддливы следующие равенства:*

$$\phi(n) = \frac{H_n}{h} + C_0 + \psi_0(n), \tag{36}$$

где $\psi_0(n)$ *определена в (35) и* C_0 *определена в лемме 7.*

Вычислительный эксперимент подтвердил эффективность предложенного подхода.

Глава 4.

В этой главе описан экономный (по трудоемкости) алгоритм, который позволяет наблюдать смещение оценки. При вычислении исследуемых оценок энтропии и построения оценок размерности фракталов в работе [13] находят следующую величину:

$$r_n^{(k)} = \frac{1}{n+1} \sum_{j=0}^{n} \phi \left(\min_{i:i \neq j}{}^{(k)} \rho(\boldsymbol{\xi_i}, \boldsymbol{\xi_j}) \right), \tag{37}$$

где ϕ — некоторая монотонная функция.

Предлагаемый в диссертации алгоритм позволяет находить величины $r_s^{(k)}$ для всех значений $s \leqslant n$ с такой же трудоемкостью (по порядку роста) как и величину $r_s^{(k)}$.

Трудоемкость алгоритма равна $\mathcal{O}(mn)$, где m — длина слов. Нахождение оценок для различных значений n позволяет визуально наблюдать смещение, которое может зависеть от n.

Постановка задачи.

Пусть заданы $n+1$ последовательностей $\boldsymbol{\xi_0} = (x_{01}, \ldots, x_{0m})$, $\ldots$, $\boldsymbol{\xi_n} = (x_{n1}, \ldots, x_{nm})$, где $x_{ij} \in \mathcal{A} = \{1, 2, \ldots, a\}$.

Будем предполагать, что применяемая метрика ρ на Ω она согласована с лексикографическим порядком, т.е.

$$\rho(\boldsymbol{x}, \boldsymbol{y}) \leq \rho(\boldsymbol{x}, \boldsymbol{z}) \quad \forall \boldsymbol{x} \preccurlyeq \boldsymbol{y} \preccurlyeq \boldsymbol{z}. \tag{38}$$

Алгоритм находит величины $r_s^{(l)}$, определенные по формуле (37), для всех $k < s \leq n$, $1 \leq l \leq k$, где k – заданное число.

Описание алгоритма.

Работа алгоритма состоит из трех шагов.

1. Выполним сортировку слов $\boldsymbol{\xi_0}, \ldots, \boldsymbol{\xi_n}$ в возрастающем порядке. Пусть

$$\boldsymbol{\eta_0} \preccurlyeq \ldots \preccurlyeq \boldsymbol{\eta_n}$$

– отсортированные слова, а σ – перестановка такая, что

$$\boldsymbol{\eta_{\sigma_i}} = \boldsymbol{\xi_i}, \quad i = 0, 1, \ldots, n.$$

Организуем слова $\boldsymbol{\eta_0}$, ..., $\boldsymbol{\eta_n}$ в список с двумя связями.

2. Для $j = 0, 1, \ldots, n$ и $l = 1, 2, \ldots, k$ находим

$$g_j^l = \min_{i \neq j : j-l \leq i \leq j+l}{}^{(k)}\rho(\boldsymbol{\eta_i}, \boldsymbol{\eta_j}). \tag{39}$$

Затем находим

$$r_n^{(l)} = \frac{1}{n+1} \sum_{j=0}^{n} \phi(g_j^l), \tag{40}$$

где $l = 1, 2, \ldots, k$.

3. Для $s = n, n-1, \ldots, k$ делаем следующее:

   (a) удаляем слова $\boldsymbol{\eta_{\sigma_s}} = \xi_s$ из списка,

   (b) для $l = 1, 2, \ldots, k$ и $j = \sigma_s - l, \ldots, \sigma_s + l$ находим величины g_j^l по формуле (39); нумерация по j идет по списку с двумя связями;

   (c) для $l = 1, 2, \ldots, k$ пересчитываем величины $r_{s-1}^{(l)}$, заменяя в формуле (40) величины g_j^l для $j = \sigma_s - l, \ldots, \sigma_s + l$.

Отметим, что величины g_j^l находятся последовательно для $l = 1, 2, \ldots, k$, поэтому для нахождения каждой величины g_j^l по формуле (39) достаточно находить минимум только из двух расстояний.

Легко видеть, что трудоемкость алгоритма равна $\mathcal{O}(k^2 nm)$. Таким образом, при небольших k (не зависящих от n) трудоемкость алгоритма равна $\mathcal{O}(nm)$.

Свойство согласованности (38) выполнено для метрик ρ_0 и ρ_1 (см. формулы (1) и (11) соответственно), однако подчеркнем, что оно не выполнено для метрики

$$\rho_2(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{m} \theta^{-k} |x_k - y_k|,$$

где $\theta > 1$.

Список литературы

1. A. Wyner and J. Ziv. Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. / A. Wyner and J. Ziv. - IEEE Trans. Inform. Theory, 1989. - No 35. - pp.1250–1258.

2. Grassberger, P. Estimating the information content of symbol sequences abd efficient codes. /P. Grassberger - IEE Trans. Inform. Theory, 1989. - No 35. - pp. 669-675.

3. I. Kontoyiannis and Yu. M. Suhov. Prefixes and the entropy rate for long-range sources / I. Kontoyiannis and Yu. M. Suhov. - Probability Statistics and Optimization; ed. F.P. Kelly. - Wiley, New York, 1994. - pp. 89-98.

4. I. Kontoyiannis, P. H. Algoet, Yu. M. Suhov and A. J. Wyner. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. /I. Kontoyiannis, P. H. Algoet, Yu. M. Suhov and A. J. Wyner. - IEEE Trans. Inform. Theory, 1998. - N 44. - pp. 1319-1327.

5. Dobrushin, R.L. The simplified method of experimental estimatation of the entropy of stationary sequence. / R.L. Dobrushin. - Theor. Prob. Appl., 1958, - No 3. - pp. 462-464.

6. V. A. Vatutin and V. G. Mikhaĭlov. Statistical estimation of the entropy of discrete random variables with a large number of outcomes. /V. A. Vatutin and V. G. Mikhaĭlov. - Uspekhi Mat. Nauk, 1995. - No 55. - pp. 121-134.

7. R. Badii and A. Politi. Hausdorff dimension and uniformity factor of strange attractors. /R. Badii and A. Politi. - Phys. Rev. Lett., 1984. - No 55. - pp. 1661-1664.

8. Billingsley, P. Ergodic Theory and Information. / P.Billingsley. - Wiley, New York, 1965. - pp. 120.

9. Shields, P.C.. Entropy and prefxes. /P.C. Shields. - Annals of Probability, 1992. - No 20. - pp. 403-409.

10. Башарин, Г.П. О статистическом оценивании энтропии последовательности независимых случайных величин. /Г.П. Башарин. Теория вероятности и ее применение, 1959. - Т. IV, No 3. - С.361-364.

11. Колмогоров, А.Н. Теория информации и теория алгоритмов./ А.Н. Колмогоров. - М.: Наука, 1987. - 300 с.

12. Орнстейн, Д. Эргодическая теория, случайность и динамические системы. / Д. Орнстейн. - М.: Мир, 1971. - 166 с.

13. В.В. Майоров, Е.А.Тимофеев. Статистическая оценка обобщенных размерностей. /В.В. Майоров, Е.А.Тимофеев. // Мат. заметки. - 2002. - Т. 71. № 5. - С.679 - 712.

14. Тимофеев, Е.А. Статистически оцениваемые инварианты мер. / Е.А. Тимофеев // Алгебра и анализ. - 2005. - Т.17. No 3. - С.204-236.

РАБОТЫ АВТОРА ПО ТЕМЕ ДИССЕРТАЦИИ

Статьи в журналах из перечня ВАК:

1. Тимофеева, Н.Е. Экономный алгоритм нахождения средних минимальных расстояний / Н.Е. Тимофеева // Моделирование и анализ информационных систем / Яросл. гос. ун-т. - Ярославль: ЯрГУ, 2007. - Т.14, No 3. - С.50-52.

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

3. A. Kaltchenko, N. Timofeeva, and E. Timofeev. Bias Reduction of the Nearest Neighbor Entropy Estimator /A. Kaltchenko, N. Timofeeva, and E. Timofeev // International Journal of Bifurcation and Chaos. - 2008. - pp.3781–3787.

4. Тимофеева, Н.Е. Линейная метрика для оценивания энтропии / Н.Е. Тимофеева // Моделирование и анализ информационных систем / Яросл. гос. ун-т. - Ярославль: ЯрГУ, 2009. - Т.16, No 1. - С.39-48.

5. A. Kaltchenko and N. Timofeeva. Rate of convergence of the Nearest Neighbor Entropy Estimator / A. Kaltchenko and N. Timofeeva // International Journal of Electronics and Communications. - 2010. - Vol.64. - pp.75–79.

Работы, опубликованные в других журналах:

6. A. Kaltchenko, E-H. Yang, and N. Timofeeva. Entropy Estimators with Almost Sure Convergence and an $O(n^{-1})$ Variance / A. Kaltchenko, E-H. Yang, and N. Timofeeva // Proceedings of the 2007 IEEE Information Theory Workshop / Lake Tahoe, California, USA, 2007. - pp.644 - 649.

7. A. Kaltchenko, I. Kotsireas, E-H. Yang, and N. Timofeeva. Entropy Rate Estimators with a Near-Optimal Upper Bound on Variance / A. Kaltchenko, I. Kotsireas, E-H. Yang, and N. Timofeeva // Proceedings of the XI International Symposium on Problems of Redundancy in Information and Control Systems / Saint Petersburg, Russia, 2007. - pp. 18 – 21.

8. A. Kaltchenko, E-H. Yang, and N. Timofeeva. Exact Analysis of the Bias of the Nearest Neighbor Entropy Estimator for I.I.D. Information Sources / A. Kaltchenko, E-H. Yang, and N. Timofeeva // Proceedings of the Forty-Fifth Annual Allerton Conference / University of Illinois at Urbana-Champaign, IL, USA, 2007. - pp. 165– 168.

9. A. Kaltchenko, E-H. Yang, and N. Timofeeva. Bias Reduction of the Nearest Neighbor Entropy Estimator / A. Kaltchenko, E-H. Yang, and N. Timofeeva // Proceedings of the 2008 IEEE International Conference on Computational Technologies in Electrical and Electronics Engineering / Novosibirsk, Russia, 2008. - pp. 261 – 265.

10. A. Kaltchenko, N. Timofeeva. Bias Reduction via Linear Combination of Nearest Neighbor Entropy Estimators/ A. Kaltchenko, N. Timofeeva // International Journal of Information and Coding Theory. - 2009. - pp.39–56.

Оригинал-макет подготовлен
в редакционно-издательском отделе ЯрГУ

Отпечатано на ризографе

Ярославский государственный университет
150000 Ярославль, ул. Советская, 14.