1966

О СТАТИСТИЧЕСКОЙ ОЦЕНКЕ ЭНТРОПИИ В ОДНОРОДНОЙ ЦЕПИ МАРКОВА

Э. В. МИСЕВИЧЮС

В настоящей заметке обобщается один результат Г. П. Башарина [1] о статистической оценке энтропии.

Рассматривается последовательность случайных величин

$$\xi_1, \xi_2, \ldots, \xi_n, \ldots,$$

связанных в простую стационарную цепь Маркова с конечным числом возможных состояний E_i, причем

$$p_i = p(\xi_k = E_i) > 0,$$
$$p_{ij} = p(\xi_{k+1} = E_j / \xi_k = E_i),$$
$$p(ij) = p(\xi_k = E_i,\ \xi_{k+1} = E_j); \qquad i, j = 1, 2, \ldots, N;\ k = 1, 2, \ldots$$

Для оценки энтропии такой цепи Маркова (см. [5])

$$H = -\sum_{i,j=1}^{N} p_i p_{ij} \ln p_{ij} = \sum_{i=1}^{N} p_i \ln p_i - \sum_{p_{ij} \neq 0} p(ij) \ln p(ij) \tag{1}$$

на основании выборочных данных в формулу (1) вместо неизвестных априорных вероятностей p_i, $p(ij)$ подставляются соответствующие оценки $\hat{p}_i = \frac{m_i}{n}$, $\hat{p}(ij) = \frac{m_{ij}}{n}$, где m_i – число попаданий в состояние E_i, а m_{ij} – число попаданий в состояния (E_i, E_j) за первые n шагов при гипотезах $\xi_1 = E_r$, $\xi_n = E_s$; $i, j, r, s = 1, 2, \ldots, N$.

Имеем:

$$\begin{aligned} &\sum_{i=1}^{N} m_i = n, \\ &\sum_{i,j=1}^{N} m_{ij} = n - 1, \\ &\sum_{j=1}^{N} m_{ij} = m_i - \delta_{is}; \qquad i, s = 1, 2, \ldots, N, \end{aligned} \tag{2}$$

где δ_{is} – символы Кронекера.

Изучим выборочное распределение случайной величины

$$\hat{H} = \sum_{i=1}^{N} \hat{p}_i \ln \hat{p}_i - \sum_{p_{ij} \neq 0} \hat{p}(ij) \ln \hat{p}(ij). \tag{3}$$

Имеет место следующая

Теорема. *$\hat{H}$ является смещенной, состоятельной, асимптотически нормальной оценкой энтропии H с математическим ожиданием*

$$M\hat{H} = H - \frac{A}{n} + O\left(\frac{1}{n^2}\right) \tag{4}$$

и дисперсией

$$D\hat{H}=\frac{B}{n}+O\left(\frac{1}{n^2}\right). \tag{5}$$

Точные выражения A *и* B *даны в* (9) *и* (11).

Доказательство. Определим новую цепь Маркова C с возможными состояниями $E_{ij}=(E_i, E_j)$, причем такими, что $p_{ij}\neq 0$, вероятностями перехода за один шаг из состояния E_{ij} в состояние E_{kl}

$$p_{ij,\,kl}=\begin{cases} p_{kl}, & j=k,\\ 0, & j\neq k,\end{cases}$$

и стационарным распределением $p(ij)$. Цепь Маркова C будет простой стационарной с конечным числом возможных состояний.

Посредством разложения функции $y=x\ln x$ по формуле Тейлора получим

$$\hat{H}_1=-\sum_{i=1}^{N}\hat{p}_i\ln\hat{p}_i=H_1-\sum_{i=1}^{N}(\ln p_i+1)(\hat{p}_i-p_i)-\frac{1}{2}\sum_{i=1}^{N}\frac{(\hat{p}_i-p_i)^2}{p_i}+$$

$$+\frac{1}{6}\sum_{i=1}^{N}\frac{(\hat{p}_i-p_i)^3}{p_i^2}-\frac{1}{12}\sum_{i=1}^{N}\frac{(\hat{p}_i-p_i)^4}{\left(p_i+\Theta_1(\hat{p}_i-p_i)\right)^3}, \qquad 0<\Theta_1<1; \tag{6}$$

$$\hat{H}_2=-\sum_{p_{ij}\neq 0}\hat{p}(ij)\ln\hat{p}(ij)=H_2-\sum_{p_{ij}\neq 0}\left(\ln p(ij)+1\right)\left(\hat{p}(ij)-p(ij)\right)-$$

$$-\frac{1}{2}\sum_{p_{ij}\neq 0}\frac{\left(\hat{p}(ij)-p(ij)\right)^2}{p(ij)}+\frac{1}{6}\sum_{p_{ij}\neq 0}\frac{\left(\hat{p}(ij)-p(ij)\right)^3}{p^2(ij)}-$$

$$-\frac{1}{12}\sum_{p_{ij}\neq 0}\frac{\left(\hat{p}(ij)-p(ij)\right)^4}{\left(p(ij)+\Theta_2\left(\hat{p}(ij)-p(ij)\right)\right)^3}, \qquad 0<\Theta_2<1. \tag{7}$$

Очевидно, что при условии $0<p_i<1$, $0<p(ij)<1$ соответствующие разложения (6) и (7) будут иметь смысл.

Для того, чтобы с помощью разложений (6) и (7) определить моменты оценок $\hat{H}_1$ и $\hat{H}_2$, нужно вычислить моменты случайных величин $\hat{p}_i$, $\hat{p}(ij)$ $(i, j=1, 2, \ldots, N)$. Как известно (см. [2]), величины m_i и m_{ij} распределены асимптотически нормально с математическими ожиданиями np_i и $(n-1)p(ij)$ и ковариационными матрицами $\|nb_{ij}\|$ и $\|(n-1)r_{ij,\,kl}\|$ соответственно. Имеем (см. [4]), что

$$M(\hat{p}_i-p_i)^3=O\left(\frac{1}{n^2}\right), \quad M\frac{(\hat{p}_i-p_i)^4}{\left(p_i+\Theta_1(\hat{p}_i-p_i)\right)^3}\leqslant\frac{M(\hat{p}_i-p_i)^4}{(1-\Theta_1)^3p_i^3}\leqslant O\left(\frac{1}{n^2}\right),$$

$$M\left(\hat{p}(ij)-p(ij)\right)^3=O\left(\frac{1}{n^2}\right),$$

$$M\frac{\left(p(ij)-p(ij)\right)^4}{\left(p(ij)+\Theta_2\left(\hat{p}(ij)-p(ij)\right)\right)^3}\leqslant O\left(\frac{1}{n^2}\right); \qquad i, j=1, 2, \ldots, N. \tag{8}$$

Используя (6), (7) и (8), получаем:

$$M\hat{H} = H - \frac{A}{n} + O\left(\frac{1}{n^2}\right),$$

где

$$A = H_2 + \frac{1}{2}\sum_{p_{ij}\neq 0}\frac{r_{ij,\,ij}}{p(ij)} - \frac{1}{2}\sum_{i=1}^{N}\frac{b_{ii}}{p_i} - 1. \tag{9}$$

Для вычисления дисперсии $\hat{H}$ воспользуемся соотношением

$$D\hat{H} = M\hat{H}_1^2 + M\hat{H}_2^2 - 2M\hat{H}_1\hat{H}_2 - M^2\hat{H}. \tag{10}$$

Из (2), (6), (7), (8), (10) и неравенства Шварца следует, что

$$D\hat{H} = \frac{B}{n} + O\left(\frac{1}{n^2}\right),$$

где

$$B = \sum_{i,j=1}^{N}(\ln p_i + 1)(\ln p_j + 1)\,b_{ij} +$$

$$+ \sum_{p_{ij}\neq 0,\,p_{kl}\neq 0}\left(\ln p(kl) + 1\right)\left(\ln\frac{p_{ij}}{p_i} - 1\right)r_{ij,\,kl} + 2H_2 + 2H_1 - 2. \tag{11}$$

Ввиду громоздкости выражений, этих вычислений мы не приводим.

Для доказательства асимптотической нормальности величины $\hat{H}_1$, запишем разложение (6) в виде

$$\sqrt{n}\,(\hat{H}_1 - H_1) = -\sqrt{n}\sum_{i=1}^{N}(\ln p_i + 1)(\hat{p}_i - p_i) - \frac{\sqrt{n}}{2}\sum_{i=1}^{N}\frac{(\hat{p}_i - p_i)^2}{p_i(1-\Theta_3) + \Theta_3\hat{p}_i},\quad 0 < \Theta_3 < 1$$

и применим к случайной величине $\sqrt{n}\,(\hat{H}_1 - H_1)$ теорему 28.4 из [3]. Аналогично доказывается, что и величина $\hat{H}_2$ асимптотически нормальна. Так как $\hat{H} = \hat{H}_2 - \hat{H}_1$, то и $\hat{H}$ является асимптотически нормальной величиной.

В заключение выражаю благодарность научному руководителю доценту В. Статулявичюсу за постановку задачи.

Вильнюсский Государственный
университет им. В. Капсукаса

Поступило в редакцию
12.I.1966

ЛИТЕРАТУРА

1. Г. П. Башарин. О статистической оценке энтропии последовательности независимых случайных величин, Теор. вер. и ее прим., **IV**, 3 (1959).
2. А. И. Колмогоров. Локальная предельная теорема для классических цепей Маркова, Изв. АН СССР (серия математическая), **XIII**, 4 (1949).
3. Г. Крамер. Математические методы статистики, М., ИЛ, 1948.
4. С. Х. Сираждинов. Предельные теоремы для однородных цепей Маркова, Ташкент, 1955.
5. А. Я. Хинчин. Понятие энтропии в теории вероятностей, УМН, **VIII**, 3 (55).

APIE HOMOGENINĖS MARKOVO GRANDINĖS ENTROPIJOS STATISTIKINĮ ĮVERTINIMĄ

E. MISEVIČIUS

(Reziumė)

Šiame straipsnyje gautas entropijos statistikinis įvertinimas. Parodoma, kad šis įvertinimas yra asimptotiškai normalinis ir paslinktas.

ÜBER DIE STATISTISCHE ABSCHÄTZUNG DER ENTROPIE IN DEN HOMOGENEN MARKOFFSCHEN KETTEN

E. MISEVIČIUS

(Zusammenfassung)

In diesem Artikel ist die statistische Abschätzung der Entropie angeführt. Es wird bewiesen, dass diese Abschätzung asymptotisch normal und asymptotisch erwartungstreu ist.