ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 24 **МАТЕМАТИКИ** **Выпуск 5**
2017

А. К. Мельников (Москва, НТЦ ЗАО «ИнформИнвестГрупп»). **Методика расчета распределений вероятностей значений статистик, близких к их точным распределениям.**

Вопросам вычислительной сложности расчета точных распределений вероятностей значений статистик $P_T\{S_n \geqslant c\}$ (точных распределений) при обработке потоков текстов посвящены работы [1–3]. Построение статистических критериев для обработки текстов длины $n > 50$ в алфавите мощности $N > 64$ требует использования точных распределений [4], так как использование предельных распределений приводит к увеличению числа ложно отобранных текстов.

1. Направление модернизации частотного метода расчета точных распределений. В [5] предложено направление модернизации частотного метода, позволяющее рассчитывать распределения статистики $P_\Delta\{S_n \geqslant c\}$, отличающиеся от их точных распределений $P_T\{S_n \geqslant c\}$ не более чем на заданную величину Δ

$$\bigl|P_T\{S_n \geqslant c\} - P_\Delta\{S_n \geqslant c\}\bigr| \leqslant \Delta.$$

Смысл модернизации частотного метода заключается в ограничении области перечисления решений уравнения, описывающего возможные значения частот h_i встречаемости знаков алфавита a_i в тексте,

$$h_1 + \cdots + h_N = n \tag{1}$$

в неотрицательных целых числах, определяемой как

$$\{h_i | i = \overline{1, N},\ h_i \in N,\ 0 \leqslant h_i \leqslant n\}$$

до области

$$\{h_i | i = \overline{1, N},\ h_i \in N,\ 0 \leqslant h_i \leqslant m,\ m < n\}, \tag{2}$$

где m значительно меньше n.

Эффект ограничения области перечисления (2) можно оценить, если известны вероятности $\mathbf{P}\{M_n < m\}$, где $M_n = \max_{i=1}^{N} h_i$. Для оценки вероятности $\mathbf{P}\{M_n < m\}$ воспользуемся рекуррентной формулой Б. И. Селиванова, предложенной им в 70-х годах XX века, и впервые опубликованной в трудах МГУ им. М. В. Ломоносова:

$$\mathbf{P}\{M_{n+1} < m\} = \sum_{\nu=0}^{n} \binom{n}{\nu} \mathbf{P}\{M_{n-\nu} < m\}\, d_{\nu+1}^{(m)} \frac{1}{N^\nu}, \tag{3}$$

с начальным условием $\mathbf{P}\{M_0 < m\} = 1$, где коэффициенты $d_{\nu+1}^{(m)}$ вычисляются по рекуррентной формуле

$$d_{n+1}^{(m)} = -\sum_{\nu=1}^{m-1} \binom{n}{\nu} \times d_{n+1-\nu}^{(m)}$$

с начальными условиями

$$d_1^{(m)} = 1, \quad d_2^{(m)} = d_3^{(m)} = \cdots = d_{m-1}^{(m)} = 0, \quad d_m^{(m)} = -1, \quad d_{m+1}^{(m)} = m.$$

2. Разработка методики расчета распределений, близких к точным. Определим μ_ν как число таких решений уравнения (1), у которых $h_i = \nu$. Тогда для расчета вероятностей $P_\Delta\{S_n \geqslant c\}$ от перебора решений уравнения (1) можно перейти к перебору решений системы уравнений

$$\begin{cases} \mu_0 + \mu_1 + \cdots + \mu_n = N, \\ 1\mu_1 + 2\mu_2 + \cdots + n\mu_n = n. \end{cases} \tag{4}$$

Учитывая (2), получаем, что при принятых ограничениях

$$\mu_{m+1} + \mu_{m+2} = \cdots = \mu_n = 0$$

можно от перебора решений системы уравнений (4) перейти к перебору усеченной системы уравнений $(m < n)$

$$\begin{cases} \mu_0 + \mu_1 + \cdots + \mu_m = N \\ 1\mu_1 + 2\mu_2 + \cdots + m\mu_m = n. \end{cases} \tag{5}$$

Выделяя независимые переменные и применяя метод их последовательного задания с определением зависимых переменных, получаем все решения системы (5)

$$\{\mu^{(i)} \,|\, (\mu_0^{(i)}, \mu_1^{(i)}, \ldots, \mu_m^{(i)}),\ i = \overline{1, Z}\} \tag{6}$$

Так, при $n = 50$, $N = 26$ и $\Delta = 10^{-5}$ по формуле (3) рассчитывается вероятность $\mathbf{P}\{M_{50} < 12\} = 0{,}9999992$ и соответственно $m = 12$. Вычисления показывают, что в этом случае $Z = 92154$, что намного меньше, чем сложность частотного метода, равная числу сочетаний с повторениями из N элементов по n и оцениваемая как 5×10^{19}.

Заметим, что с каждым решением системы (5) связано

$$K^{(i)} = \frac{N!}{\mu_0^{(i)}!\mu_1^{(i)}! \cdots \mu_m^{(i)}!} \tag{7}$$

решений уравнения (1). Тогда вероятность $P^{(i)}$ того, что решение уравнения (4)

$$\mu_0^{(i)} + \mu_1^{(i)} + \cdots + \mu_n^{(i)} = N$$

примет значение $\mu^{(i)}$ из (6) равна

$$P^{(i)} = K^{(i)} \frac{n!}{(2!)^{\mu_2^{(i)}} \times (3!)^{\mu_3^{(i)}} \times \cdots \times (m!)^{\mu_m^{(i)}} \times N^n}. \tag{8}$$

Теперь для каждого $\{\mu^{(i)} | i = \overline{1, Z}\}$ мы можем вычислить $P^{(i)}$ и значение статистики $S_n^{(i)}$, например $\chi_n^{(i)}$, где

$$\chi_n^{(i)} = \chi_n(\mu^{(i)}) = \frac{N}{n} \sum_{\nu=0}^{m} \mu_\nu^{(i)} \Big(\nu - \frac{n}{N}\Big)^2. \tag{9}$$

Имея вычисленные тройки $\{\mu^{(i)}, P^{(i)}, S_n^{(i)} | i = \overline{1, Z}\}$, можно перейти непосредственно к вычислению вероятностей $\mathbf{P}\{S_n \geqslant c\}$. Для этого для всех

$$\{c_j | j = 1, 2, \ldots, \max_{i=1,\ldots,Z} S_n^{(i)}\}$$

вычисляем

$$\mathbf{P}\{S_n \geqslant c_j\} = \sum_{i=1}^{Z} P^{(i)} I(S_n^{(i)}, c_j),$$

где

$$I(S_n^{(i)}, c_j) = \begin{cases} 1 & \text{при} \quad S_n^{(i)} \geqslant c_j \\ 0 & \text{при} \quad S_n^{(i)} < c_j. \end{cases}$$

Полученная последовательность

$$\left\{\mathbf{P}\left\{S_n \geqslant c_j\right\} \big| \, j = 1, 2, \ldots, \max_{i=1,\ldots,Z} S_n^{(i)}\right\}$$

может служить оценкой дискретного распределения статистики S_n, отличающейся от точного распределения не более чем на заданную величину Δ.

3. Методика расчета распределений вероятностей значений статистик, близких к точным. Пусть n — длина выборки (текста) и N — мощность алфавита текста. Точное распределение вероятностей $P_T\{S_n \geqslant c\}$ рассчитать не представляется возможным, поэтому рассчитывается ее Δ-точное распределение $P_\Delta\{S_n \geqslant c\}$, отличающееся от точного не более чем на заданную величину Δ

$$|P_T\{S_n \geqslant c\} - P_\Delta\{S_n \geqslant c\}| \leqslant \Delta.$$

Шаг 1. По заданным (n, N) и выбранной точности Δ (например, 10^{-5}) для ограничения области перебора решений уравнения (1) и нахождения m по формулам (3) последовательно вычисляем вероятности

$$\begin{array}{c} \mathbf{P}\{M_1 < 2\},\ \mathbf{P}\{M_2 < 2\}, \ldots, \mathbf{P}\{M_n < 2\}; \\ \mathbf{P}\{M_1 < 3\},\ \mathbf{P}\{M_2 < 3\}, \ldots, \mathbf{P}\{M_n < 3\}; \\ \ldots\ldots\ldots\ldots \\ \mathbf{P}\{M_1 < m\},\ \mathbf{P}\{M_2 < m\}, \ldots, \mathbf{P}\{M_n < m\}, \end{array}$$

пока не выполнится условие $\mathbf{P}\{M_n < m\} > 1 - \Delta$. Таким образом определяем m.

Шаг 2. Для перечисления всех решений системы уравнений

$$\begin{cases} \mu_0 + \mu_1 + \cdots + \mu_m = N \\ 1\mu_1 + 2\mu_2 + \cdots + m\mu_m = n \end{cases}$$

выделяем независимые переменные и применяем метод их последовательного задания с определением зависимых переменных. Получаем все решения, количество которых обозначим Z:

$$\left\{\mu^{(i)} | (\mu_0^{(i)}, \mu_1^{(i)}, \ldots, \mu_m^{(i)}),\ i = \overline{1, Z}\right\}.$$

Шаг 3. Для каждого решения

$$\left\{\mu^{(i)} | (\mu_0^{(i)}, \mu_1^{(i)}, \ldots, \mu_m^{(i)}),\ i = \overline{1, Z}\right\}.$$

вычисляем по формуле (9) значение статистики $S_n^{(i)}$, а по формулам (7) и (8) — вероятность его появления $P^{(i)}$. Таким образом получаем набор $\{\mu^{(i)}, S_n^{(i)}, P^{(i)} |\ i = \overline{1, Z}\}$. Отметим, что формулы расчета значений одной и той же статистики от значений частот встречаемости знаков h_i и от значений так называемых «вторых» маркировок μ_ν отличаются друг от друга, что необходимо учитывать при проведении расчетов.

Шаг 4. Для получения распределения вероятностей значений $P_\Delta\{S_n \geqslant c_j\}$ маркируем значения статистики $\{S_n^{(i)} | i = \overline{1, Z}\}$ в интервалах $\{c_j \,|\, j = 1, 2, \ldots, \max_{i=1}^{Z} S_n^{(i)}\}$ с одновременным суммированием соответствующих вероятностей $P^{(i)}$.

З а к л ю ч е н и е. Модернизация частотного метода расчета точных распределений позволяет увеличить значения длин и мощностей алфавитов текстов, для которых могут быть получены распределения, близкие к точным. Приведена методика расчета Δ точных распределений, отличающихся от точных не более чем на заданную величину Δ. Приводятся результаты расчета Δ-точных распределений для конкретных значений параметров выборки.

Для повышения эффективности статистических процедур обработки потоков текстов представляется перспективным построение методов расчета Δ-точных распределений в тех областях значений параметров, для которых точности предельных распределений недостаточно.

СПИСОК ЛИТЕРАТУРЫ

1. *Мельников А. К.* Исследование путей модернизации реконфигурируемых вычислительных систем в интересах решения вычислительно трудоемких задач. — Вестник компьютерных и информационных технологий, 2016, №2(140), с. 52–59.

2. *Зелюкин Н. Б.*, *Мельников А. К.* Сложность расчета точных распределений вероятности значений статистик и область применения предельных распределений. В сб.: Электронные средства и системы управления: Материалы докладов XIII Международной научно-практической конференции. (29 ноября – 1 декабря 2017 г.) Томск: В-Спектр, 2017, Ч. 2, с. 84–90.

3. *Мельников А. К.* Сложность расчета точных распределений вероятности симметричных аддитивно разделяемых статистик и область применения предельных распределений. — Доклады ТУСУРа. 2017, т. 20, №4, с. 126–130.

4. *Мельников А. К.*, *Ронжин А. Ф.* Обобщенный статистический метод анализа текстов, основанный на расчете распределений вероятности значений статистик. — Информатика и ее применения, 2016, т. 10, в. 4, с. 89–95,

5. *Мельников А. К.* Направление модернизации частотного метода расчета точных распределений вероятностей значений статистик. — Обозрение прикл. и промышл. матем., 2017, т. 24, в. 5 (http://tvp.ru/conferen/vsppmXVIII/kisso073.pdf)