ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 25 МАТЕМАТИКИ Выпуск 4
2018

А. К. Мельников (Москва, НТЦ ЗАО «ИнформИнвестГрупп»). **Сравнение эффективности обработки текстов при применении в статистических критериях точных и предельных приближений базовых распределений вероятностей значений тестовых статистик.**

При решении задач, связанных с обработкой текстов [1], некоторые из них могут быть сведены к процедуре отбора текстов с равновероятным распределением входящих в них знаков, основанной на применении критериев согласия с равновероятным распределением [2]. В работе проводится сравнение эффективности обработки, построенной на использовании в применяемом для отбора текстов критерии различных приближений базового распределения вероятностей значения тестовой статистики [3].

1. Построение процедуры обработки текстов и ее эффективность. Пусть на первом этапе из массива, состоящего из M текстов

$$T_{n,N}(j) = \{t_1(j), \ldots, t_n(j)\}, \qquad j = 1, 2, \ldots, M,$$

длины n, содержащих знаки алфавита $A_N = \{a_1, \ldots, a_N\}$ мощности N, нам необходимо отобрать тексты, являющиеся реализациями случайных выборок длины n из равновероятного распределения на алфавите мощности N. Используя для отбора текстов критерий согласия со статистикой S_n и ее равновероятным распределением $\mathbf{P}\{S_n \geqslant x\}$, отбираем из M текстов, поступающих на вход процедуры, $\overline{M}$ текстов в качестве равновероятных. Количество ложно отобранных как равновероятные текстов $\overline{M''}$, входящих в количество $\overline{M}$, определяется размером критерия α, связанным с распределением статистики $S_n - \mathbf{P}\{S_n \geqslant x\}$ равенством $\mathbf{P}\{S_n \geqslant c\} = \alpha$, посредством соотношений $\overline{M''} \cong \alpha\,\overline{M}$ и $\overline{M} = \overline{M'} + \overline{M''}$, где $\overline{M'}$ — количество правильно отобранных текстов. Под эффективностью $\omega(\alpha, \mathbf{P}\{S_n \geqslant x\})$ процедуры обработки, аналогично [4], будем понимать

$$\omega(\alpha, \mathbf{P}\{S_n \geqslant x\}) = \frac{\overline{M'}}{\overline{M}} = \frac{\overline{M} - \alpha\,\overline{M}}{\overline{M}} = \frac{(1-\alpha)\,\overline{M}}{\overline{M}} = 1 - \alpha. \tag{1}$$

2. Эффективность обработки при применении различных приближений базовых распределений тестовых статистик. Пусть, исходя из общего числа отобранных текстов $\overline{M}$ и ограничения на величину ложно отобранных текстов $\overline{M''}$, выбран размер критерия α. Тогда, используя точное приближение распределения тестовой статистики $S_n - P_\Delta\{S_n \geqslant x\}$ [5] из условия

$$P_\Delta\{S_n \geqslant c\} = \alpha,$$

получаем разделяющую константу $c = P_\Delta^{-1}\{\alpha\}$ критерия, основанного на точном приближении распределения статистики S_n.

Если бы в качеств базового распределения для решения данной задачи мы использовали предельное приближение распределения тестовой статистики $S_n - P_{\lim}\{S_n \geqslant x\}$ [6], то значением разделяющей константы c_1 критерия, основанного на предельном приближении распределения статистики S_n, служило бы значение

$$c_1 = P_{\lim}^{-1}\{\alpha\}. \tag{2}$$

Из свойств неотрицательности, невозрастания и монотонности функций распределения $P_\Delta\{S_n \geqslant x\}$, $P_{\lim}\{S_n \geqslant x\}$ и согласно данным расчетов и сравнений точных и предельных приближений распределений тестовых статистик [7]

$$P_\Delta\{S_n \geqslant x\} \geqslant P_{\lim}\{S_n \geqslant x\} \geqslant 0 \quad \text{для любых} \quad x \geqslant 0$$

и из $P_\Delta\{S_n \geqslant c\} = P_{\lim}\{S_n \geqslant c_1\} \geqslant 0 = \alpha$ следует, что $c_1 \leqslant c$.

Применение решающего правила с разделяющей константой критерия c_1 к выборке, имеющей точное распределение $P_T\{S_n \geqslant x\}$, приведет к изменению размера критерия до $\alpha_2 = P_T\{S_n \geqslant c_1\}$. Точное распределение $P_T\{S_n \geqslant x\}$ нам не известно из-за большой вычислительной сложности его расчета [8, 9], но мы можем вычислять такие Δ-точные распределения $P_\Delta\{S_n \geqslant x\}$, что

$$|P_\Delta\{S_n \geqslant x\} - P_T\{S_n \geqslant x\}| \leqslant \Delta \quad \text{для любых} \quad x \geqslant 0.$$

Но применение разделяющей константы критерия c_1 при использовании в качестве базового распределения критерия точного приближения $P_\Delta\{S_n \geqslant x\}$ определяет размер критерия из условия $P_\Delta\{S_n \geqslant c_1\} = \alpha_1$, которое с учетом (2) может быть представлено в виде

$$P_\Delta\{S_n \geqslant P^{-1}_{\lim}\{\alpha\}\} = \alpha_1. \tag{3}$$

Повторно используя свойства неотрицательности, невозрастания и монотонности функций распределения $P_\Delta\{S_n \geqslant x\}$, $P_{\lim}\{S_n \geqslant x\}$, имеем согласно данным из [7]

$$P_\Delta\{S_n \geqslant x\} \geqslant P_{\lim}\{S_n \geqslant x\} \geqslant 0 \quad \text{для любых} \quad x \geqslant 0,$$

и тогда из неравенства $c_1 \leqslant c$ следует, что

$$\alpha_1 \geqslant \alpha. \tag{4}$$

Из свойств точного приближения, в качестве которого используется Δ-точное распределение, следует, что

$$\left|P_\Delta\{S_n \geqslant x\} - P_T\{S_n \geqslant x\}\right| \leqslant \Delta \quad \text{для любых} \quad x \geqslant 0$$

и, следовательно, $|\alpha_1 - \alpha_2| \leqslant \Delta$.

Проиллюстрируем приведенные выше рассуждения графически.

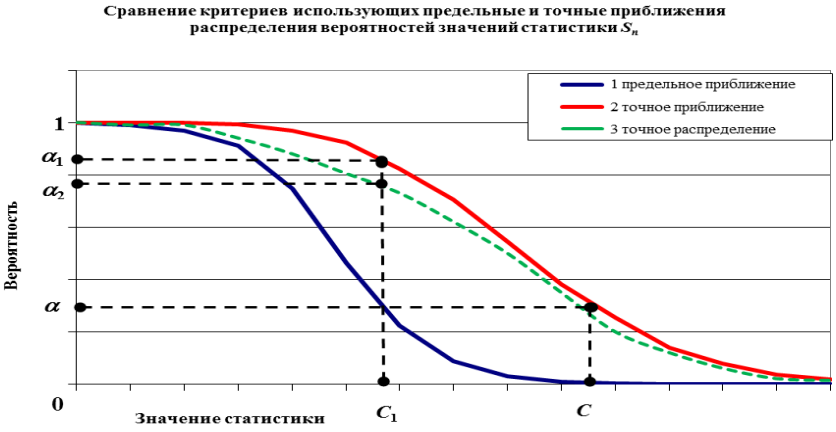

Рис. Иллюстрация результатов применения в критериях обработки текстов точных и предельных приближений распределений используемой статистики. Нижняя линия соответствует предельному приближению распределения, верхняя линия — точному приближению распределения, штриховая линия между ними — точному распределению.

Применение выбранной при использовании предельного приближения распределения разделяющей константы критерия c_1 приведет к увеличению размера критерия c α до α_2, и к увеличению числа ложно отобранных текстов в α_2/α раз, что на практике при ограничениях на $\Delta \approx 10^{-5}$ и выполнении условия

$$\left|P_\Delta\{S_n \geqslant x\} - P_{\lim}\{S_n \geqslant x\}\right| \gg \Delta \quad \text{для любых} \quad x \geqslant 0$$

может оцениваться как

$$\frac{\alpha_2}{\alpha} \geqslant \alpha_1 - \frac{\Delta}{\alpha} \cong \frac{\alpha_1}{\alpha}.$$

Для статистики хи-квадрат [6] увеличение числа ложно отобранных текстов будет до 4 раз [10] (см. табл. 3 и рис. 4), для статистики максимального правдоподобия [6] до 2,5 раз [10]. Использование предельного приближения распределения для статистики Матуситы [6] приводит к недостоверному результату, так как число ложно отобранных текстов увеличивается многократно [10].

Возвращаемся к вопросу эффективности обработки $\omega(\alpha, P\{S_n \geqslant x\})$ при применении в критерии согласия тестовой статистики S_n и ее базового распределения $\mathbf{P}\{S_n \geqslant x\}$. Задавая размер критерия α; $c = P_\Delta^{-1}\{\alpha\}$, с Δ-точного распределения $P_\Delta\{S_n \geqslant x\}$ (точного приближения), тогда согласно (1) эффективность обработки при использовании точного приближения равна

$$\omega(\alpha, P_\Delta\{S_n \geqslant x\}) = 1 - P_\Delta\{S_n \geqslant c\} = 1 - \alpha.$$

Использование в качестве приближения базового распределения предельного распределения $P_{\lim}\{S_n \geqslant x\}$ (предельного приближения) определяет разделяющую константу как

$$c_1 = P_{\lim}^{-1}\{\alpha\}.$$

Используя предыдущие рассуждения и (3), можем утверждать, что

$$\omega\big(\alpha, P_{\lim}\{S_n \geqslant x\}\big) = 1 - \big(P_\Delta\{S_n \geqslant P_{\lim}^{-1}\{\alpha\}\}\big) = 1 - P_\Delta\{S_n \geqslant c_1\} = 1 - \alpha_1.$$

Для сравнения эффективности обработки при применении разных приближений базового распределений тестовой статистики S_n исследуем их разность

$$\omega\big(\alpha, P_\Delta\{S_n \geqslant x\}\big) - \omega\big(\alpha, P_{\lim}\{S_n \geqslant x\}\big) = 1 - \alpha - 1 + \alpha_1 = \alpha_1 - \alpha \geqslant 0. \qquad (5)$$

Выражение (5) согласно (4) всегда неотрицательно и, следовательно, эффективность обработки текстов при применении точных приближений распределений используемых статистик будет не хуже, чем при применении предельных приближений, а как показывает практика, и в несколько раз лучше.

Заключение. Рассмотрена эффективность процедур отбора текстов с равновероятным распределением входящих в них знаков, построенных на основе критериев согласия и использующих в качестве базового распределения тестовой статистики их точные и предельные приближения. Показано, что использование в критерии согласия точного приближения не уменьшает по сравнению с использованием предельного приближения, а во многих случаях и увеличивает эффективность обработки в смысле уменьшения доли ложно отобранных текстов.

СПИСОК ЛИТЕРАТУРЫ

1. *Чеповский А. М.* Информационные модели в задачах обработки текстов на естественных языках. М.: Национальный открытый университет «ИНТУИТ», 2015, 228 с.
2. *Крамер Г.* Математические методы статистики./ Пер. с англ. А. С. Монина, А. А. Петрова под ред. А. Н. Колмогорова. М.: Мир, 1975, 648 с.

3. *Ивченко Г. И.*, *Медведев Ю. И.* Введение в математическую статистику. М.: ЛЕНАРД, 2017, 608 с.
4. *Мельников А. К.* Применение точных распределений в процедуре двухэтапной обработки текстов. — Обозрение прикл. и промышл. матем., 2018, т. 25, в. 2, с. 175–178.
5. *Мельников А. К.* Методика расчета распределения вероятностей значений симметричных аддитивно разделяемых статистик, приближенных к их точному распределению. — Научный вестник НГТУ. 2018, № 1(70), с. 153–166.
6. *Мельников А. К.*, *Ронжин А. Ф.* Обобщенный статистический метод анализа текстов, основанный на расчете распределений вероятности значений статистик. — Информатика и ее примен., 2016, т. 10, в. 4, с. 89–95.
7. *Мельников А. К.* Анализ точных и предельных приближений распределений вероятностей значений статистик. — Суперкомпьютерные технологии (СКТ-2018): материалы 5-й Всероссийской научно-технической конференции: в 2 т. Ростов-на-Дону – Таганрог: Изд-во Южного федерального ун-та, 2018, с. 100–104.
8. *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.
9. *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.) Ч. 2. Томск: В-Спектр, 2017, с. 84–90.
10. *Мельников А. К.* Применение точных и предельных приближений распределений вероятностей значений статистик при решении задачи обработке текстов. — Изв. ЮФУ. Сер. Техн. науки. Тематический выпуск. Суперкомпьютерные технологии. декабрь 2018, № 8 (202). (В печати.)