Методы оценки малых вероятностей в задачах биоинформатики

Быков Кирилл Владимирович, гр. 422

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Вычислительная стохастика и статистические модели

Научный руководитель: к. ф.-м. н. Коробейников А.И.
Рецензент: Младший научный сотрудник Шлемов А. Ю.

Санкт-Петербург
2018г.

Довольно часто в прикладных задачах встает вопрос об оценке вероятности редких событий.

Примеры:

- **PSM**: задача о совпадении теоретического и экспериментального масс спектра пептидов [Abramova A. et al., 2017].
- Оценки вероятностей в задачах поляризация дрожжевых клеток и изучении ферментативного цикла [Bernie J. et al., 2011].
- Задачи, связанные с локальным выравниванием последовательностей [Wolfsheimer S. et al., 2007]

Известно, что метод Монте-Карло не эффективен в задачах оценки малых вероятностей. Актуальной задачей является уменьшение дисперсии оценки при фиксированном числе моделирований.

Постановка задачи

Пусть заданы:

- марковский процесс $\mathrm{S}(t) \in \mathbb{S}, t \in \mathbb{N}_0$ на множестве $\mathbb{S}$ со стационарным распределением $\mathcal{P}$,
- функция $\varphi(\mathrm{S})$: $\mathbb{S} \to \mathbb{R}$.

Требуется оценить вероятность «редкого» (например, порядка 10^{-7} или меньше) события

$$A = \{\mathrm{S} \in \mathbb{S} | \varphi(\mathrm{S}) \geqslant k\},$$

где k — некоторая заданная константа.

В работе рассматривается применение метода RESTART для данной задачи.

Задачи:

- формализовать метод для применения в прикладных задачах биоинформатики,
- определить способ для оценки дисперсии оценки, полученной методом RESTART,
- изучить закон распределения оценок,
- научиться строить доверительные интервалы.

Метод RESTART (**R**epetitive **S**imulation **T**rials **A**fter **R**eaching **T**hresholds) [Villen-Altamirano, M. and Villen-Altamirano, J., 1991.]

$\mathbb{Z}$ – область значений функции φ.

$A = \{S \in \mathbb{S} | \varphi(S) \geqslant k\}$, $\varphi(A)$ — образ множества A. Далее, определим на $\mathbb{Z}$ семейство вложенных подмножеств: $\{C_i\}_{i=0}^{M}$:

$$\mathbb{Z} = C_0 \supset C_1 \supset C_2 \supset ... \supset C_M \supset \varphi(A).$$

Определение

Уровень T_i:

$$T_i := \begin{cases} C_i \backslash C_{i+1} & i \in [0, 1, ..., M-1] \\ \varphi(A) & i = M \end{cases}$$

Таким образом:

$$\mathbb{Z} = \bigsqcup_{i=0}^{M} T_i.$$

Определение

Процесс $\mathrm{S}(t) \in \mathbb{S}, t \in \mathbb{N}_0$ будем называть **главной ветвью моделирования**.

Определение

$\Phi(\mathrm{S})$ — **функция состояния**, определяет номер уровня, которому принадлежит данное состояние.

$$\Phi(\mathrm{S}(t)) : \mathbb{S} \to \mathbb{I}, \mathbb{I} = [0, \dots M]$$

$$\Phi(\mathrm{S}(t_0)) = i, \text{ где } i : \varphi(\mathrm{S}(t_0)) \in T_i.$$

Определим события:

- B_i^t – переход на новый уровень: $\Phi(\mathrm{S}(t)) \geqslant i$ при $\Phi(\mathrm{S}(t-1)) < i$,
- D_i^t – угасание: $\Phi(\mathrm{S}(t)) < i$ при $\Phi(\mathrm{S}(t-1)) \geqslant i$.

Определение

Ветвью уровня $i, i > 0$ называется процесс $\{\mathrm{S}^i_{t_0}(t) \in \bar{\mathbb{S}} = \mathbb{S} \cup \Delta, t \geqslant t_0\}$, где $\bar{\mathbb{S}} = \mathbb{S} \cup \Delta$, Δ — поглощающее состояние.

Ветвь $\mathrm{S}^i_{t_0}$ определена аналогично главной ветви, но вероятность перейти из состояний принадлежащих $X_i = \mathbb{S} \backslash \varphi^{-1}(C_i)$ в поглощающее состояние Δ равна 1.

Расщепление процесса:
Если на ветви уровня $i-1$ происходит $B^{t_0}_i$, то процесс *расщепляется*

$$\mathrm{S}^{i-1}_{t'}(t), [\mathrm{S}^i_{1,t_0}(t), \mathrm{S}^i_{2,t_0}(t), \ldots, \mathrm{S}^i_{R_i-1,t_0}(t)],$$

то есть определяются $R_i - 1$ независимых *ветвей уровня* i параллельно с *ветвью уровня* $i-1$. Начальное состояние:

$$\mathrm{S}^i_{j,t_0}(t_0) = \mathrm{S}^{i-1}_{t'}(t_0), \forall j \in [1, R_i - 1].$$

Оценка вероятности A:

$$\hat{p} = \frac{N_A}{N\, r_M},$$

где

- N_A— суммарное время, которое все ветви *провели* в A

$$N_A = \sum_{t=0}^{N-1} \sum_{j \in \mathbb{H}_t} \mathbb{1}_A(\mathrm{S}_j(t)),$$

  где $\mathbb{H}_t$ — множество индексов *активных* ветвей моделирования в момент времени t,
- N – число моделирований в главной ветви, параметр,
- $r_M = \prod_{j=1}^{M} R_j$, параметр.

Утверждение [Villen-Altamirano, M. and Villen-Altamirano, J., 1991.]

$\hat{p}$ – несмещенная и состоятельная оценка вероятности $\mathcal{P}(A)$.

Численно на нескольких примерах была продемонстрирована сходимость:

$$\sqrt{N}(\hat{p} - p) \xrightarrow[N\to\infty]{D} \mathcal{N}(0, \sigma_p^2).$$

Для построения доверительных интервалов хотим оценить величину σ_p^2. Для этого воспользуемся методами Batch Means и Overlapping Batch Means.

Определим последовательность:

$$Y_t = \frac{\sum_{j\in\mathbb{H}_t} \mathbb{1}_A(\mathrm{S}_j(t))}{r_M}, t \geqslant 0$$

Тогда

$$\hat{p} = \bar{Y}_N = \frac{\sum_{i=0}^{N-1} Y_i}{N}$$

Методы Batch Means и Overlapping Batch Means [Flegal J. et al., 2010]:

- **Batch Means**: Поделим N на a частей длины b: $N = a \cdot b$.

$$\bar{Y_j} = \frac{1}{b} \sum_{i=jb}^{j(b+1)-1} Y_i, j \in [0, a-1].$$

Тогда оценка σ_p^2 по методу Batch Means:

$$\hat{\sigma}^2_{BM} = \frac{b}{a} \sum_{i=0}^{a-1} (\bar{Y_j} - \hat{p})^2.$$

- **Overlapping Batch Means**: Считаем среднее по всем «окнам» фиксированного размера b :

$$\bar{Y_j} = \frac{1}{b} \sum_{i=j}^{j+b-1} Y_i, j \in [0, N-b].$$

Тогда оценка σ_p^2 по методу Overlapping Batch Means:

$$\hat{\sigma}^2_{OBM} = \frac{Nb}{(N-b)(N-b+1)} \sum_{i=0}^{N-b} (\bar{Y_i} - \hat{p})^2.$$

Продемонстрировано, что $\hat{\sigma}^2_{OBM}, \hat{\sigma}^2_{BM}$ при $a = b = \sqrt{N}$ и оценки, полученные путем выборочной дисперсии оценок независимых реализаций метода (Multistart), показывают с достаточной точностью схожие результаты, из чего можно предположить, что

$$\hat{\sigma}^2_{BM} \xrightarrow[N\to\infty]{} \sigma^2_p, \hat{\sigma}^2_{OBM} \xrightarrow[N\to\infty]{} \sigma^2_p$$

с вероятностью 1.

Таким образом, в предположении о состоятельности и в предположении

$$\sqrt{N}(\hat{p} - p) \xrightarrow[N\to\infty]{D} \mathcal{N}(0, \sigma^2_p),$$

доверительные интервалы для оценки вероятности имеют вид:

$$P\left(\hat{p} - z_{1-\frac{\alpha}{2}} \frac{\sigma^2_{BM}}{\sqrt{N}} < p < \hat{p} + z_{\frac{\alpha}{2}} \frac{\sigma^2_{BM}}{\sqrt{N}}\right) \xrightarrow[N\to\infty]{} 1 - \alpha,$$

$$P\left(\hat{p} - z_{1-\frac{\alpha}{2}} \frac{\sigma^2_{OBM}}{\sqrt{N}} < p < \hat{p} + z_{\frac{\alpha}{2}} \frac{\sigma^2_{OBM}}{\sqrt{N}}\right) \xrightarrow[N\to\infty]{} 1 - \alpha.$$

Определение

$\mathbb{S}^{(n)}$ – множество строк длины n над конечным алфавитом $\mathfrak{A}$.

$$|\mathbb{S}^{(n)}| = |\mathfrak{A}|^n$$

Определение

Расстояние Хэмминга $d : d(\mathbb{S}^{(n)} \times \mathbb{S}^{(n)}) \to \mathbb{Z}$ — число позиций, в которых соответствующие символы двух строк одинаковой длины различны.

Пусть $\mathfrak{A} = \{\mathrm{A}, \mathrm{G}, \mathrm{T}, \mathrm{C}\}$, $\bar{\mathrm{S}} \in \mathbb{S}^{(50)}$ — некоторая строка, $\mathcal{P}$ – равномерное распределение на $\mathbb{S}^{(50)}$. Требуется оценить вероятность $\mathcal{P}(A)$ события $A = \{\mathrm{S} \in \mathbb{S}^{(50)} | d(\mathrm{S}, \bar{\mathrm{S}}) = 50\}$.

Аналитически известно:

$$\mathcal{P}(A) = \left(\frac{3}{4}\right)^{50} \approx 5.6632 \times 10^{-7}.$$

Зафиксируем C — число моделирований процесса S и сравним два метода.

Таблица 1: Таблица с результатами

C	$\hat{p}_{RESTART}$	$\hat{p}_{MC}$
1.7×10^4	4.17×10^{-7}	0
1.8×10^5	5.97×10^{-7}	0
1.8×10^6	5.76×10^{-7}	1.07×10^{-6}

Таблица 2: Сравнение доверительных интервалов

C	95% CI RESTART	95% CI Монте-Карло
1.7×10^4	$[0, 1.3 \times 10^{-6}]$	—
1.8×10^5	$[3.46 \times 10^{-7}, 8.48 \times 10^{-7}]$	—
1.8×10^6	$[5.06 \times 10^{-7}, 6.45 \times 10^{-7}]$	$[0, 2.94 \times 10^{-6}]$

Задача вычисления вероятности случайного совпадения экспериментального и теоретического масс-спектра пептидов (Peptide-Spectrum-Matches, PSM) [Abramova A. et al., 2017].

Таблица 3: Список пептидов, по которым производились вычисления

Имя	Масса	Пептид	Длина
Fr1.2116.2116.2	544.74634	GEEEPSQGQK	10
Fr1.2076.2076.2	414.68990	GPDGPEEK	8
Fr1.1059.1059.2	436.21185	PPAEDSQK	8
Fr1.1201.1201.2	531.24677	SSSGAGEGQGPK	12

Таблица 4: Сравнение методов RESTART и Монте-Карло: оценки вероятностей

Пептид	C	$\hat{p}_{RESTART}$	$\hat{p}_{MC}$
GEEEPSQGQK	2×10^6	3.28×10^{-7}	0
GPDGPEEK	1.3×10^6	3.41×10^{-5}	3.13×10^{-5}
PPAEDSQK	1.6×10^6	2.28×10^{-6}	2.44×10^{-6}
SSSGAGEGQGPK	1.9×10^6	4.57×10^{-12}	0

Таблица 5: Сравнение методов RESTART и Монте-Карло: доверительные интервалы

Пептид	95% CI RESTART	95% CI Монте-Карло
GEEEPSQGQK	$[2.98 \times 10^{-7}, 3.59 \times 10^{-7}]$	-
GPDGPEEK	$[3.17 \times 10^{-5}, 3.65 \times 10^{-5}]$	$[2.1 \times 10^{-5}, 4.08 \times 10^{-5}]$
PPAEDSQK	$[2.08 \times 10^{-6}, 2.47 \times 10^{-6}]$	$[4.88 \times 10^{-8}, 4.83 \times 10^{-6}]$
SSSGAGEGQGPK	$[0, 1.6 \times 10^{-11}]$	-

Заключение

Полученные в данной работе результаты открывают простор для дальнейшего применения метода RESTART в задачах оценки вероятностей редких событий в задачах биоинформатики.

В работе были получены следующие результаты:

- изучен метод RESTART, описание метода было формализовано для применения в прикладных задачах,
- предложен метод для оценки дисперсии,
- изучен закон распределения оценок по методу RESTART,
- предложен способ построения доверительных интервалов
- определена алгоритмизация метода, написана реализация на языке программирования «Python 3»,
- при помощи метода были получены результаты для нескольких биоинформатических задач.