

УДК 519.987

Экономный алгоритм нахождения средних минимальных расстояний

Тимофеева Н.Е.

Ярославский государственный университет

150 000, Ярославль, Советская, 14

E-mail : ninaveta@gmail.com

получена 15 августа 2007

Аннотация

Пусть заданы $n + 1$ строк $\xi_0, \dots, \xi_n$ с символами из некоторого конечного алфавита. В работе предлагается алгоритм нахождения величин среднего значения k -го минимального расстояния между строками $\xi_0, \dots, \xi_s$ для всех значений $s \leq n$. Трудоемкость алгоритма равна $\mathcal{O}(nm)$, где m – длина строк.

В работе [1] для нахождения оценок размерностей предлагается строить следующую величину:

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

где $\xi_0, \dots, \xi_n$ – заданные точки метрического пространства Ω , ρ – заданная метрика, ϕ – заданная монотонная функция, а $\min^{(k)} \{x_1, \dots, x_N\} = x_k$, если $x_1 \leq x_2 \leq \dots \leq x_k \leq \dots \leq x_N$.

В настоящей работе предлагается алгоритм нахождения величин $r_s^{(k)}$, для всех значений $s \leq n$ в пространстве последовательностей Ω . Трудоемкость алгоритма равна $\mathcal{O}(nm)$, где m – длина строк. Здесь параметр k предполагается небольшим (константа, не зависящая от n).

Задача нахождения величин $r_s^{(k)}$ для всех значений $s \leq n$ представляет интерес в связи с тем, что дисперсия случайной величины $r_n^{(k)}$ быстро убывает с ростом n [1], если $\xi_0, \dots, \xi_n$ – независимые одинаково распределенные случайные величины, а смещение оценки $r_n^{(k)}$ не известно. Поэтому вычисление величин $r_s^{(k)}$ для $s \leq n$ позволяет заметить существенное смещение. Применение предлагаемого алгоритма для оценивания энтропии ($\phi(x) = -\log x$) [2] показало отсутствие значимого смещения в рассмотренных примерах.

1. Постановка задачи

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

Будем обозначать лексикографический порядок на пространстве $\Omega = \mathcal{A}^m$ через $\mathbf{x} \preceq \mathbf{y}$.

Пусть на Ω задана метрика ρ , про которую будем предполагать, что она согласована с порядком, т.е.

$$\rho(\mathbf{x}, \mathbf{y}) \leq \rho(\mathbf{x}, \mathbf{z}) \quad \forall \mathbf{x} \preceq \mathbf{y} \preceq \mathbf{z}. \quad (2)$$

Требуется найти величины $r_s^{(l)}$, заданные по формуле (1), для всех $k < s \leq n$, $1 \leq l \leq k$, где k – заданное число.

2. Описание алгоритма

1) Выполним сортировку строк $\xi_0, \dots, \xi_n$ в возрастающем порядке. Пусть

$$\eta_0 \preceq \dots \preceq \eta_n$$

– отсортированные строки, а σ – перестановка такая, что

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

Организуем строки $\eta_0, \dots, \eta_n$ в двусвязанный список.

2) Для $j = 0, 1, \dots, n$ и $l = 1, 2, \dots, k$ находим величины

$$g_j^l = \min_{i \neq j: j-l \leq i \leq j+l} {}^{(k)}\rho(\boldsymbol{\eta}_i, \boldsymbol{\eta}_j). \quad (3)$$

Для $l = 1, 2, \dots, k$ полагаем

$$r_n^{(l)} = \frac{1}{n+1} \sum_{j=0}^n \phi(g_j^l). \quad (4)$$

3) Для $s = n, n-1, \dots, k$ делаем следующее:

- (а) удаляем строку $\boldsymbol{\eta}_{\sigma_s}$ из двусвязанного списка;
- (б) для $l = 1, 2, \dots, k$ и $j = \sigma_s - l, \dots, \sigma_s + l$ находим величины g_j^l по формуле (3);
- (с) для $l = 1, 2, \dots, k$ пересчитываем величины $r_{s-1}^{(l)}$, заменяя в формуле (4) величины g_j^l для $j = \sigma_s - l, \dots, \sigma_s + l$.

Отметим, что нумерация по j на шаге 3б идет по двусвязанному списку.

Отметим также, что поскольку величины g_j^l находятся последовательно для $l = 1, 2, \dots, k$, то для нахождения каждой величины g_j^l по формуле (3) достаточно находить минимум только из двух расстояний.

3. Обоснование алгоритма

Корректность алгоритма следует из того, что по свойству (2)

$$\min_{i: i \neq j} {}^{(k)}\rho(\boldsymbol{\eta}_i, \boldsymbol{\eta}_j) = \min_{i \neq j: j-k \leq i \leq j+k} {}^{(k)}\rho(\boldsymbol{\eta}_i, \boldsymbol{\eta}_j).$$

Поэтому величины $r_n^{(k)}$, найденные по формулам (4) и (1), совпадают.

4. Трудоемкость алгоритма

Трудоемкость шага 1 равна $\mathcal{O}(nm)$.

На остальных шагах наиболее трудоемкой операцией является нахождение расстояния между двумя строками. Будем считать, что ее трудоемкость равна $\mathcal{O}(m)$.

Трудоемкость шага 2 равна $\mathcal{O}(k^2 nm)$.

Трудоемкость шага 3 равна $\mathcal{O}(k^2 nm)$.

Итак, трудоемкость алгоритма равна $\mathcal{O}(k^2 nm)$.

Подчеркнем, что при небольших k (не зависящих от n) трудоемкость алгоритма равна $\mathcal{O}(nm)$.

5. Свойство согласованности

Нетрудно видеть, что свойство согласованности (2) выполняется для метрики

$$\rho_0(\mathbf{x}, \mathbf{y}) = \frac{1}{\min \{k : x_k \neq y_k\}}, \quad (5)$$

где $\mathbf{x} = (x_1, x_2, \dots)$, и $\mathbf{y} = (y_1, y_2, \dots)$.

Для второй метрики

$$\rho_1(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^m \theta^{-k} |x_k - y_k|, \quad (6)$$

где $\theta > 1$, свойство (2) выполняется только при $\theta \geq a = |\mathcal{A}|$.

Список литературы

1. Майоров, В.В. Статистическая оценка обобщенных размерностей /В.В.Майоров, Е.А.Тимофеев //Мат. заметки. – 2002. –Т.71. №5. – С. 697 – 712.
2. Kaltchenko, A., Entropy Estimators with Almost Sure Convergence and an $O(n^{-1})$ Variance /A.Kaltchenko, En-hui Yang, N.Timofeeva //Information Theory Workshop. 2007. ITW '07. IEEE 2-6 Sept. 2007. – P. 644 – 649.

Fast algorithm for finding mean minimum distances

Timofeeva N.E.

Let $\xi_0, \dots, \xi_n$ be strings drawn from some finite alphabet. In this paper we describe an algorithm for finding mean minimum distances between strings $\xi_0, \dots, \xi_s$ for all $s \leq n$. The complexity of the algorithm is $\mathcal{O}(nm)$, where m is the length of strings.