Analiza Algorytmów - Zadanie 21

Janusz Witkowski 254663

25 marca 2023

1 Zadanie 21

1.1 Treść

Rozważ następujący algorytm, z którego wywodzi się idea algorytmu HyperLogLog.

1: Probabilistic Counter

Initialization: $C \leftarrow 1$
Upon event: if $random() <= 2^{-C}$ **then**
    $C \leftarrow C + 1$
end if

Innymi słowy, przy wystąpieniu zdarzenia rzucamy monetą C razy i jeśli za każdym razem otrzymujemy reszkę zwiększamy licznik C o jeden. W przeciwnym razie nie robimy nic.

Niech C_n oznacza wartość przechowywaną w liczniku C po zaobserwowaniu n zdarzeń. Pokaż, że $\mathbb{E}(2^{C_n}) = n+2$ oraz $\mathbb{V}\mathrm{ar}(2^{C_n}) = \frac{1}{2}n(n+1)$. W oparciu o C_n zdefiniuj nieobciążony estymator wartości n i policz jego wariancję.

1.2 Rozwiązanie

1.2.1 Wartość oczekiwana

Pokażemy, że $\mathbb{E}(2^{C_n}) = n + 2$, za pomocą indukcji po n. Dla $n = 0$, czyli przed zaobserwowaniem jakichkolwiek zjawisk, wartość licznika jest równa $C_n = C_0 = 1$, a więc wartość oczekiwana licznika wynosi

$$\mathbb{E}(2^{C_n}) = \mathbb{E}(2^{C_0}) = \mathbb{E}(2^1) = 2^1 = 2 = 0 + 2 = n + 2$$

Teraz załóżmy, że $\mathbb{E}(2^{C_n}) = n + 2$. Chcemy pokazać, że $\mathbb{E}(2^{C_{n+1}}) = (n + 1) + 2 = n + 3$. Możemy rozpisać tę wartość oczekiwaną:

$$\mathbb{E}(2^{C_{n+1}}) = \sum_{\mathrm{k}\geq 0} \mathbb{E}(2^{C_{n+1}}|C_n = k) \cdot Pr[C_n = k] = \sum_{\mathrm{k}\geq 0} (\frac{1}{2^k} \cdot 2^{k+1} + (1 - \frac{1}{2^k}) \cdot 2^k) \cdot Pr[C_n = k] =$$

$$= \sum_{k\geq 0}(2+2^k-1)\cdot Pr[C_n = k] = \sum_{k\geq 0}(1+2^k)\cdot Pr[C_n = k] = \sum_{k\geq 0} Pr[C_n = k] + \sum_{k\geq 0} 2^k\cdot Pr[C_n = k]$$

Łatwo stwierdzić że $\sum_{k\geq 0} Pr[C_n = k] = 1$, natomiast z definicji wartości oczekiwanej $\mathbb{E}(2^{C_n}) = \sum_{k\geq 0} 2^k \cdot Pr[C_n = k]$. Stąd możemy podstawić:

$$\mathbb{E}(2^{C_n}) = 1 + \mathbb{E}(2^{C_n}) = 1 + (n+2) = n+3 = (n+1)+2$$

co kończy dowód indukcyjny.

1.2.2 Wariancja

Obliczymy wariancję z następującego wzoru: $\mathbb{V}\mathrm{ar}(2^{C_n}) = \mathbb{E}[(2^{C_n})^2] - (\mathbb{E}[2^{C_n}])^2$. Do policzenia wartości wariancji potrzeba nam wiedzieć ile wynosi $\mathbb{E}[(2^{C_n})^2] = \mathbb{E}[4^{C_n}]$.

Udowodnimy indykcyjnie po n, że $\mathbb{E}(4^{C_n}) = \frac{3}{2}(n+1)(n+2)+1$. Jasnym jest, że dla $n = 0$ mamy

$$\mathbb{E}(4^{C_n}) = \mathbb{E}(4^{C_0}) = \mathbb{E}(4^1) = 4^1 = 4 = 3+1 = \frac{3}{2}\cdot 1\cdot 2 + 1 = \frac{3}{2}(n+1)(n+2)+1$$

Teraz wprowadźmy założenie indykcyjne, ustalmy że chcemy dojść do postaci $\mathbb{E}(4^{C_{n+1}}) = \frac{3}{2}(n+2)(n+3)+1$ i zacznijmy rachować:

$$\mathbb{E}(4^{C_{n+1}}) = \sum_{k\geq 0} 4^k \cdot Pr[C_{n+1} = k] =$$

Zauważmy, że możemy podzielić prawdopodobieństwo wewnątrz sumy na dwie sytuacje - albo licznik miał tę wartość wcześniej, albo właśnie ją nabył:

$$= \sum_{k\geq 0} 4^k\cdot Pr[C_{n+1} = k|C_n = k]\cdot Pr[C_n = k] + \sum_{k\geq 0} 4^k\cdot Pr[C_{n+1} = k|C_n = k-1]\cdot Pr[C_n = k-1] =$$

$$= \sum_{k\geq 0} 4^k(1 - \frac{1}{2^k})Pr[C_n = k] + \sum_{k\geq 0} 4^k\frac{1}{2^{k-1}}Pr[C_n = k-1] =$$

$$= \sum_{k\geq 0}(4^k - 2^k)Pr[C_n = k] + \sum_{k\geq 0} 2\cdot 2^k \cdot Pr[C_n = k-1] =$$

$$= \sum_{k\geq 0} 4^k Pr[C_n = k] - \sum_{k\geq 0} 2^k Pr[C_n = k] + 2\cdot 2\sum_{k\geq 0} 2^{k-1} Pr[C_n = k-1] =$$

$$= \mathbb{E}(4^{C_n}) - \mathbb{E}(2^{C_n}) + 4\,\mathbb{E}(2^{C_n}) = \mathbb{E}(4^{C_n}) + 3\,\mathbb{E}(2^{C_n}) =$$

$$= \frac{3}{2}(n+1)(n+2) + 3(n+2) = \frac{3}{2}(n+2)(n+3)+1$$

Mając wartość $\mathbb{E}(4^{C_{n+1}})$ dowiedzioną indukcyjnie możemy obliczyć wariancję:

$$\mathbb{V}\mathrm{ar}(2^{C_n}) = \frac{3}{2}(n+1)(n+2) - (n+2)^2 = \frac{1}{2}n(n+1)$$

1.2.3 Nieobciążony estymator

Zdefiniujmy następujący estymator wartości n:

$$\hat{n} = 2^{C_n} - 2$$

Estymator ten jest **nieobciążony**, ponieważ jego wartość oczekiwana jest dokładnie równa szacowanej wartości n:

$$\mathbb{E}(\hat{n}) = \mathbb{E}(2^{C_n} - 2) = \mathbb{E}(2^{C_n}) - 2 = (n+2) - 2 = n$$

Do obliczenia wariancji tego estymatora możemy znów wykorzystać wzór $\mathbb{V}\mathrm{ar}(\hat{n}) = \mathbb{E}[(\hat{n})^2] - (\mathbb{E}[\hat{n}])^2$. Wartość oczekiwana estymatora jest nam znana, zatem znamy też wartość drugiej składowej. Policzmy pierwszą składową:

$$\mathbb{E}[(\hat{n})^2] = \mathbb{E}[(2^{C_n} - 2)^2] = \mathbb{E}[4^{C_n} - 2 \cdot 2 \cdot 2^{C_n} + 4] =$$

$$= \mathbb{E}[4^{C_n}] - 4\,\mathbb{E}[2^{C_n}] + \mathbb{E}[4] = \frac{3}{2}(n+1)(n+2) - 4(n+2) + 4 = \frac{3}{2}n^2 + \frac{1}{2}n$$

Podstawiamy do wzoru na wariancję:

$$\mathbb{V}\mathrm{ar}[\hat{n}] = \frac{3}{2}n^2 + \frac{1}{2}n - n^2 = \frac{1}{2}n(n+1)$$