Złożoność Kołmogorowa a entropia

Maria Fronczak

15. czerwca 2004

1 Przypomnienie notacji

Na początku przypomnijmy notację z poprzedniego wykładu:

- $\bar{x}$, czyli „delimitowany" x:

  $$\bar{x} = a_k 0 a_{k-1} 0 \ldots a_1 0 a_0 1 x$$

  gdzie $a_k a_{k-1} \ldots a_1 a_0 = bin\ (|x|)$

- Parę określamy następująco: $\langle x, y \rangle = 00\bar{x}y$
  Zauważmy, że $|\langle x, y \rangle| = 2 \cdot log\ (|x|)\ +\ \mathcal{O}(1)\ +\ |x|\ +\ |y|$

- Złożoność względna: $C_\phi\ (w|u) = min\ \{|x| :\ \phi\ (\langle u, x \rangle) = w\}$
  dla dowolnej funkcji częściowo obliczalnej ϕ

- Złożoność bezwzględna: $C_\phi\ (w) = C_\phi\ (w|\epsilon)$

- Maszyna uniwersalna: $U\ (\langle u, \langle z, x \rangle \rangle) = M_z\ (\langle u, x \rangle)$
  (M_z jest maszyną o kodzie z)

- $\phi_0\ (w) = U\ (w)$

Złożoność Kołmogorowa pozwala nam określić, jak bardzo „typowy" jest dany ciąg.
Złożoność Kołmogorowa (C) definiowana jest od razu relatywistycznie (złożoność bezwzględna równa jest złożoności względnej względem słowa pustego). W definicji na funkcję częściowo obliczalną ϕ możemy patrzeć jak na język programowania; złożoność to długość najkrótszego programu, który wygeneruje nam słowo w – przy czym przy generowaniu może być wykorzystywane dodatkowo u.

Na poprzednim wykładzie udowodnione także zostało następujące twierdzenie o niezmienniczości:

Twierdzenie 1 *Dla każdej funkcji częściowo obliczalnej ϕ istnieje c_ϕ takie, że*

$$C_{\phi_0}(w|u) \leq C_\phi(w|u) + c_\phi$$

Twierdzenie o niezmienniczości mówi, że złożoność względem ϕ_0 (funkcji obliczalnej przez maszynę uniwersalną U) jest złożonością optymalną z dokładnością do stałej. Inaczej mówiąc, programy w innych niż ϕ_0 językach programowania nie będą dowolnie krótsze niż w ϕ_0.

2 Ograniczenia złożoności

Zdefiniowanie złożoności Kołmogorowa przez funkcję K było stosowne, gdy mogliśmy założyć, że języki rozpoznawane przez maszynę Turinga są bezprefiksowe (żadne słowo z języka nie jest prefiksem innego słowa z języka – poza sobą samym). Bez tego założenia występowałyby bowiem problemy z odwracaniem konkatenacji. Przy definicji C nie ma takich problemów dzięki zastosowaniu opisanego powyżej kodowania pary – z kodu pary $\langle x, y\rangle$ potrafimy już efektywnie odkodowywać x i y.

Korzystając z definicji złożoności Kołmogorowa K nie można było udowodnić, że złożoność względna jest mniejsza niż bezwzględna (co wyraża intuicję, że zazwyczaj łatwiej – a w każdym razie na pewno nie trudniej – jest wygenerować program mając dodatkowe informacje u niż ich nie mając). Przy definicji C natomiast możemy już tę zależność pokazać.

Fakt 1 *Zachodzą następujące nierówności:*

(a) $C_{\phi_0} \leq |w| + c$

(b) $C_{\phi_0}(w|u) \leq C_{\phi_0}(w) + c$

Dowód:

Ad. (a)
Będziemy chcieli znaleźć ograniczenie na złożoność C_ϕ dla pewnego ϕ, a następnie skorzystać z twierdzenia o niezmienniczości, by oszacować złożoność C_{ϕ_0}.
Gdybyśmy chcieli użyć ϕ generującego w w jak najprostszy sposób, to najlepsza jako funkcja ϕ byłaby funkcja „identycznościowa". (Cudzysłowy sygnalizują, że ze względu na użyte kodowanie pary nie będzie to mogła być po

prostu identyczność.) Z definicji

$$C_\phi(w) = C_\phi(w|\epsilon) = min\ \{|x| : \phi(\langle \epsilon, x\rangle) = w\}$$

Chcemy, by x było równe po prostu w, mamy więc parę $\langle \epsilon, w\rangle = 0001w$. Dokładniej więc potrzebujemy nie tyle funkcji identycznościowej, ile funkcji dającej w na parze $\langle \epsilon, w\rangle$, czyli funkcji odcinającej 4 pierwsze cyfry. Bierzemy

$$\phi\ (y) = [(0+1)^4]^{-1} y$$

Dla takiego ϕ mamy oczywiście $C_\phi\ (w) \leq |w|$. Korzystając z tego i z twierdzenia o niezmienniczości mamy ostatecznie, że:

$$C_{\phi_0}\ (w) \leq |w| + c$$

gdzie c jest stałą, o której istnieniu mówi twierdzenie o niezmienniczości.

Ad. (b)

Intuicyjnie: chcemy pokazać, że dla programu generującego w przy użyciu ϵ (czyli bez dodatkowych informacji), możemy uzyskać niedłuższy program generujący w z wykorzystaniem u (przy czym programy te mogą być równej długości np. w przypadku, gdy nie będziemy w ogóle korzystać z dodatkowej informacji zawartej w u).

Przypuśćmy, że $C_{\phi_0}\ (w) = |x_0|$, gdzie $\phi_0\ (\langle \epsilon, x_0\rangle) = w$, tzn. $U\ (\langle \epsilon, x_0\rangle) = w$. Możemy łatwo utworzyć maszynę M_{z_1} taką, że $M_{z_1}\ (\langle u, x\rangle) = \phi_0\ (\langle \epsilon, x\rangle)$ dla dowolnych u, x. (Czyli maszyna ta dla każdego u, x robi to, co robiłaby uniwersalna z wejściem ϵ, inaczej mówiąc u jest ignorowane). Wówczas

$$U\ (\langle u, \langle z_1, x_0\rangle\rangle) = M_{z_1}(\langle u, x_0\rangle) = \phi_0\ (\langle \epsilon, x_0\rangle) = w$$

Z definicji złożoności względnej $C_{\phi_0}\ (w|u) = min\ \{|x| : \phi_0\ (\langle u, x\rangle) = w\}$, mamy też, że $\phi_0\ (\langle u, \langle z_1, x_0\rangle\rangle) = U\ (\langle u, \langle z_1, x_0\rangle\rangle)$ i $U\ (\langle u, \langle z_1, x_0\rangle\rangle) = w$, zatem $C_{\phi_0}\ (w|u) \leq |\langle z_1, x_0\rangle|$. Uwzględniając szacowanie wartości $|\langle z_1, x_0\rangle|$ dostajemy

$$C_{\phi_0}\ (w|u) \leq \underbrace{2 \cdot log|z_1| + |z_1| + \mathcal{O}(1)}_{c} + \underbrace{|x_0|}_{C_{\phi_0}(w)}$$

□

C spełnia zatem (z dokładnością do stałych):

- długość najkrótszego programu generującego jakiś ciąg jest nie większa niż długość tego ciągu
- mając dodatkowe informacje można co najwyżej skrócić program.

3 Połączenie złożoności Kołmogorowa z entropią

Prosta obserwacja
Przyjmijmy $C \equiv C_{\phi_0}$, zachodzi $C(w|w) = 0$ (z dokładnością do stałej: jeśli wybierzemy z_0 jako kod maszyny liczącej „identyczność" – taką, jak było to opisane wcześniej, czyli:

$$M_{z_0}(x) = x \ bez \ 4 \ pierwszych \ cyfr$$

to mamy $U\ (\langle y, \langle z_0, \epsilon\rangle\rangle) = y$, a zatem także $C\ (w|w) \leq |\langle z_0, \epsilon\rangle|)$.
Pozwala to zastanawiać się nad związkiem między złożonością Kołmogorowa a entropią, gdyż entropia „gdy wszystko jest wiadomo" również wynosi 0.

Twierdzenie 2 *Niech $u = y_1 y_2 \ldots y_m \in (0+1)^*$, gdzie $|y_i| = n$.*
(Nie każdy ciąg da się oczywiście podzielić na bloki równej długości, ale zaniedbujemy to.) Określmy dla $w \in (0+1)^n$

$$p(w) = \frac{\#\{i : y_i = w\}}{m}$$

(czyli częstość wystąpień słowa w w ciągu u – liczby te są jak prawdopodobieństwa w przestrzeni losowej ze słowami długości n, $\sum_{w\in(0+1)^n} p(w) = 1$) Wówczas

$$C(u) \leq m \cdot \left(\overbrace{\sum_{w\in(0+1)^n} p(w) \cdot (-log_2\ p(w))}^{entropia\ \ przestrzeni\ \ probabilistycznej} + \epsilon\ (m) \right) + c$$

gdzie c jest stałą, a $\epsilon(m) \to 0$ przy ustalonym n i $m \to \infty$

W istocie moglibyśmy oczekiwać takiego rezultatu. W najbardziej chaotycznym przypadku, gdy wszystkie $p(w)$ są sobie równe (i przy dużym m), entropia równa będzie logarytmowi z rozmiaru przestrzeni – czyli n; z powyższego twierdzenia mamy wówczas $C\ (u) \leq m \cdot n = |u|$ (z dokładnością do stałej), czyli jak przy wcześniejszym szacowaniu złożoności bezwzględnej.

Dowód:
Ciąg u możemy odtworzyć znając ciąg liczb $s(w) = \#\{i : y_i = w\} = p(w) \cdot m$. Nie wystarczy jednak tylko ta informacja (gdyż może być wiele ciągów o identycznym rozkładzie częstości podciągów), dodatkowo potrzebujemy znać numer (w porządku leksykograficznym np.) ciągu u wśród wszystkich ciągów spełniających to samo ograniczenie (rozkład częstości).

Pozostaje tylko oszacować rozmiar potrzebnych informacji – jak się okaże, sporą wagę ma właśnie ów numer ciągu, reszta wejdzie w składnik sumy opisany jako $\epsilon(m)$.

Jako „program" generujący u przyjmijmy parę $\langle\langle n, X_n\rangle, j_n\rangle$, gdzie j_n jest numerem ciągu u, $X_n = \tilde{s}(w_1)\tilde{s}(w_2)\ldots\tilde{s}w_{2^n}$ i $\{0,1\}^n = \{w_1, w_2, \ldots, w_{2^n}\}$. $\tilde{s}(w)$ to binarne przedstawienie liczby $s(w)$, ewentualnie dopełnione z przodu zerami, by wszystkie ciągi $\tilde{s}$ były tej samej długości: $|\tilde{s}(w)| = \lfloor log_2\ m \rfloor + 1$.

Dzięki równej długości tych liczb możemy łatwo oszacować długość $\langle\langle n, X_n\rangle, j_n\rangle$. Szacujemy rozmiar pierwszego elementu tej pary (jako że on sam jest parą, korzystamy ze wcześniejszych szacowań rozmiaru kodowania pary):

$$|\langle n, X_n\rangle| = (\lfloor log_2\ n \rfloor + 1) + 2^n \cdot (\lfloor log_2\ m \rfloor + 1) + 2 \cdot log(\lfloor log_2\ n \rfloor + 1) + \mathcal{O}(1)$$

Każdy składnik powyższej sumy będzie jeszcze dodatkowo dzielony przez m, toteż w wyprowadzanym ograniczeniu na $C(u)$ trafi on do składnika ϵ.

By oszacować j należy się zastanowić, ile ciągów spełnia podane ograniczenie na u (rozkład częstości podciągów w u). j jest liczbą pomiędzy 1 i $\binom{m}{s_1 \ldots s_N}$, gdzie $s(w_i) = s_i$, $2^n = N$, $\binom{m}{s_1 \ldots s_N} = \frac{m!}{s_1! \ldots s_N!}$
Musimy więc oszacować $log\frac{m!}{s_1! \ldots s_N!}$, zrobimy to korzystając ze wzoru Stirlinga:

$$k! \approx \sqrt{2\pi k} \cdot \left(\frac{k}{e}\right)^k$$

skąd $log\ k! \approx \frac{1}{2}(log\ 2\pi k) + k \cdot (log\ k - log\ e)$.
Szacując j mamy zatem:

$$\frac{1}{2} \cdot [log\ 2\pi m - log\ 2\pi s_1 - \ldots - log\ 2\pi s_N] - (log_2\ e)[\overbrace{m - s_1 - \ldots - s_N}^{0}] +$$
$$+ m \cdot log\ m - s_1 \cdot log\ s_1 - \ldots - s_N \cdot log\ s_N = -s_1 log\frac{s_1}{m} - s_2 log\frac{s_2}{m} - \ldots +$$
$$- s_N log\frac{s_N}{m} + \overbrace{\frac{1}{2} \cdot [log\ 2\pi m - log\ 2\pi s_1 - \ldots - log\ 2\pi s_N]}^{zbiega\ do\ 0\ po\ podzieleniu\ przez\ m}$$

(gdzie skorzystaliśmy z faktu, że $m = s_1 + \ldots + s_N$)
Suma ta dzielona będzie jeszcze przez m (by wyciągnąć m przed nawias) i jako składnik w szacowaniu $C(u)$ pojawi się nam wówczas:

$$\underbrace{m \cdot (-\frac{s_1}{m} log \frac{s_1}{m} - \ldots - \frac{s_N}{m} log \frac{s_N}{m})}_{m \cdot \sum_{w \in (0+1)^n} p(w)(-log\ p(w))}$$

□

4 Uniwersalny test Martina-Löfa

Definicja 1 *Testem nazywamy funkcję całkowitą* $\delta : (0+1)^* \to \omega$, *niekoniecznie obliczalną, ale taką, że zbiór* $\{\langle m, x\rangle : \delta(x) \geq m\}$ *jest częściowo obliczalny.* [1]
Ponadto wymagamy, by dla każdych m, n

$$\frac{\#\{w \in (0+1)^n : \ \delta(w) \geq m\}}{2^n} \leq \frac{1}{2^m}$$

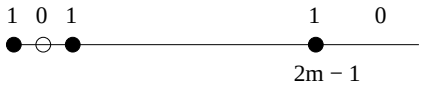
Intuicyjnie przez ciąg typowy rozumiemy taki, który należy do każdej rozsądnej większości; który ma niewiele „dziwactw". Test δ wskazuje właśnie na poziom „dziwactwa". Dodatkowo w podanym warunku n odnosi się do długości ciągów, zaś m jest progiem, pewną miarą „dziwactwa". Sam warunek mówi, że liczba ciągów długości n dziwacznych na poziomie co najmniej m jest nie większa niż $\frac{1}{2^m}$.

Ciekawym faktem jest, że istnieje taki uniwersalny test „dziwactwa" ciągu, tzn. zupełnie niezależny od własności, której dziwactwo badamy.

Przykład
Niech ciągami dziwacznymi będą takie ciągi, mające same jedynki na początkowych nieparzystych pozycjach. Test dla tej własności ma postać:

$$\delta(x) = max\ \{i : x_1 = x_3 = \ldots = x_{2i-1} = 1\}$$

Np. $\delta(0111) = 0, \quad \delta(1110100) = 3$
To jest test: jeśli $\delta(w) \geq m$, to musi być $|w| = n \geq 2m - 1$:



(gdyby ciąg w był krótszy, nie miałby dość nieparzystych pozycji, by test mógł zwrócić m).

Należy sprawdzić, czy spełniony jest warunek proporcji ciągów (czyli że „dziwacznych" na poziomie co najmniej m jest nie więcej niż $\frac{1}{2^m}$). W tym celu, należy się zastanowić, ile ciągów długości n ma jedynki na pierwszych m nieparzystych pozycjach. W takim ciągu trzeba jeszcze ustalić:

[1] Przykładem takiej funkcji jest $\delta = \begin{cases} 1 & gdy\ M_x(x)\downarrow \\ 0 & wpp. \end{cases}$

- co stoi na pierwszych $m-1$ pozycjach o numerach parzystych (między jedynkami - to nie muszą być zera!), co daje 2^{m-1} możliwości

- w jaki sposób uzupełnione są pozycje za ostatnią jedynką z podciągu, na pozycji $2m-1$. Za tą jedynką pozostaje $n-(2m-1)$ pozycji dających jednak tylko 2^{n-2m} możliwości uzupełnienia (jeśli ciąg jest dłuższy niż $2m$, a test zwrócił dla niego m, to znaczy że na pozycji $2m+1$ musi być zero. Jedynie gdy ciąg jest długości $2m$ do uzupełnienia pozostaje $n-2m+1$, czyli jedna, ostatnia pozycja; gdy zaś jest długości $2m-1$ nie ma już w ogóle dalszych pozycji do uzupełnienia)

Rozpatrzmy teraz osobno przypadki:

- gdy ciąg ma długość $2m-1$:
  wówczas ciągów dziwacznych na poziomie co najmniej m jest tyle, co ciągów dziwacznych na poziomie m (bo dysponujemy tylko m pozycjami o nieparzystych numerach) i proporcja liczby ciągów dziwacznych na poziomie co najmniej m do wszystkich ciągów wynosi $\frac{2^{m-1}\cdot 2^0}{2^{2m-1}} \leq \frac{1}{2^m}$ (tu zachodzi równość)

- gdy ciąg ma długość $2m$:
  wówczas ciągów dziwacznych na poziomie co najmniej m jest tyle, co ciągów dziwacznych na poziomie m (gdyż w takich ciągach jest m pozycji nieparzystych) i proporcja ciągów dziwacznych na poziomie co najmniej m do wszystkich wynosi $\frac{2^{m-1}\cdot 2^1}{2^{2m}} \leq \frac{1}{2^m}$ (tu też zachodzi równość)

- gdy ciąg jest dłuższy niż $2m$:
  razem ciągów, o których wiemy, że maksymalny podciąg jedynek na nieparzystych pozycjach kończy się na pozycji $2m-1$ (czyli dziwacznych na poziomie dokładnie m) jest wówczas $2^{m-1}\cdot 2^{n-2m} = 2^{n-m-1}$ i proporcja ciągów dziwacznych na poziomie co najmniej m (czyli na poziomach $m,\ m+1,\ldots,\lceil\frac{n}{2}\rceil$) do wszystkich długości n wynosi

$$\frac{2^{n-m-1}+2^{n-m-2}+\ldots+2^{n-\lceil\frac{n}{2}\rceil}}{2^n} \leq \frac{2^{n-m}}{2^n} = \frac{1}{2^m}$$

zatem rzeczywiście tak zdefiniowana δ jest testem.

Moglibyśmy się spodziewać, że dla różnych własności testy będą różne. Tymczasem istnieje test uniwersalny.

Lemat 1 *Istnieje rekurencyjna numeracja wszystkich testów $\delta_1,\ \delta_2, \ldots$ (istnieje maszyna Turinga, która dla każdego testu potrafi wygenerować indeks zbioru opisanego w teście).*

Dowód powyższego lematu zamieszczony jest w pracy [1], str. 130 (lemat 2.4.1).

Twierdzenie 3 *Niech $\delta_1, \delta_2, \ldots$ będzie numeracją wszystkich testów. Wtedy $\delta^U(x) = max\ \{\delta_n(x) - n : n \geq 1\}$ jest testem uniwersalnym w tym sensie, że*

$$\delta^U(x) \geq \delta(x) - c_\delta$$

dla pewnej stałej c_δ.

W powyższej definicji testu uniwersalnego bierzemy co prawda maksimum po nieskończonym zbiorze, będzie jednak zawsze istniało, bo dla odpowiednio dużych n liczba $\delta_n\ (x) - n$ będzie liczbą ujemną.

Dowód:
Dość łatwo jest pokazać, że tak zdefiniowany test uniwersalny spełnia warunek, że liczba ciągów długości n „dziwacznych" na poziomie m do wszystkich ciągów długości n jest co najwyżej $\frac{1}{2^m}$.
Z definicji liczba ciągów dziwacznych na poziomie m:

$$\#\{w \in (0{+}1)^n : \delta^U(w) \geq m\} = \#\{w \in (0+1)^n : max\{\delta_k(w) - k : k \geq 1\} \geq m\}$$

a oczywiście

$$\#\{w \in (0+1)^n : max\ \{\delta_k\ (w) - k : k \geq 1\} \geq m\} \leq$$

$$\sum_{k=1}^{\infty} \#\{w \in (0+1)^n : \delta_k\ (w) - k \geq m\} =$$

$$\sum_{k=1}^{\infty} \underbrace{\#\{w \in (0+1)^n : \delta_k(w) \geq m+k\}}_{\leq\ 2^{n-(m+k)}} \leq$$

$$2^{n-m} \cdot \sum_{k=1}^{\infty} \frac{1}{2^k} = 2^{n-m}$$

gdzie skorzystaliśmy z tego, że δ_k (dla $k = 1, 2, \ldots$) jest testem, zachodzi więc $\#\{w \in (0+1)^n : \delta_k(w) \geq m\} \leq 2^{n-m}$.
A zatem proporcja ciągów dziwacznych (według δ^U) na poziomie m do wszystkich ciągów długości n wynosi co najwyżej $\frac{2^{n-m}}{2^n} = \frac{1}{2^m}$.

□

Literatura

[1] Ming Li and Paul Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Verlag, 1997.