Zadanie 3 z listy 3 - "Kompresja Danych"

Łukasz Klasiński

4 kwietnia 2020

Zadanie 3

Optymalną długością słowa kodowego kodów Tunstalla dla zadanych prawdopodobieństw symboli $p_1, \ldots, p_N$ nazywać będziemy taką liczbę n, że kody Tunstalla ze słowami kodowymi długości n dają najmniejszą średnią liczbę bitów przypadającą na jeden znak kodowanych tekstów o rozkładzie $p_1, \ldots p_N$.

Czy dla każdych $p_1, \ldots p_N$ istnieje optymalna długość słowa kodowego kodów Tunstalla?

Rozwiązanie

W ogólności na internecie można znaleźć dowód na to, że długość słowa kodowego kodów Tunstalla wraz ze wzrostem n przybliża się do Entropi (dla odpowiednio dużego słownika), ale ze względu na jego skomplikowanie, przedstawię prostszy przykład, dla którego rzeczywiście nie istnieje takie n, które daje optymalne długości słów kodowych.

Weźmy alfabet 2 literowy oraz prawdopodobieństwa $p_1 = \frac{1}{2}, p_2 = \frac{1}{2}$. Chce pokazać, że niezależnie jakie dobierzemy n, dla takich danych średnia długość zawsze wynosi 1 (więc nie ma optymalnego n). Dowód będzie indukcyjny po n.

D-d

- Podstawa indukcji:

  n $= 1$, wtedy oczywiście średnia długość wynosi $\frac{1}{2} + \frac{1}{2} = 1$.

- Krok indukcyjny:

  Weźmy $n + 1$. Wtedy wzór na średnią liczbę kodową wygląda następująco:

  $$\frac{n+1}{\sum_{i=1}^{2^{n+1}} |c_i| * p_i}$$

  Widać, że algorytm biorąc kolejnego kandydata z takich danych, dla których wszystkie prawdopodobieństwa są równe, usunie go oraz doda dwa nowe słowa, które mają prawdopodobieństwa mniejsze o 2. W takim razie algorytm będzie w taki sposób usuwać wartości z poprzedniej iteracji i zastępować 2 nowymi. Zatem algorytm produkuje drzewo z dokładnie 2^{n+1} liśćmi, dla których wszystkie mają słowa kodowe długości $n + 1$. Dlatego możemy zapisać wzór jako:

  $$\frac{n+1}{\sum_{i=1}^{2^{n+1}} (n+1) * p_i}$$

  Ale

  $$\sum_{i=1}^{2^{n+1}} (n+1) * p_i = (n+1) * \sum_{i=1}^{2^{n+1}} p_i = n + 1$$

  Zatem otrzymujemy $\frac{n+1}{n+1} = 1$

Zatem dla tego przykładu nie istnieje stała n, dla której Tunstall stworzy optymalne długości słów.

□