



UNIUNEA EUROPEANĂ



GUVERNUL ROMÂNIEI



Instrumente Structurale
2007-2013



Platformă de e-learning și curriculum e-content pentru învățământul superior tehnic

Transmisia datelor multimedia in rețele de calculatoare

10. Cvasi-optimalitatea codarii aritmetice

Cvasi-optimalitatea codarii aritmetice

- Dupa teorema codarii a lui Shannon, numarul lungimea medie a cuvintelor de cod trebuie sa indeplineasca relatia

$$\bar{l} = \frac{H(S)}{H_{\max}(X)} = \frac{H(S)}{\log(D)} = H(S)$$

- Fie un mesaj ce contine M simboluri. Dupa statistica sursei, simbolul apare de ori
- Latimea intervalului final, dupa considerarea tuturor simbolurilor, este

$$\begin{aligned} R_f = H_f - L_f &= \prod_{j=1}^M P(s_j) = \underbrace{P(s_1) \cdot \dots \cdot P(s_1)}_{\text{de } M \cdot P(s_1) \text{ ori}} \cdot \dots \cdot \underbrace{P(s_N) \cdot \dots \cdot P(s_N)}_{\text{de } M \cdot P(s_N) \text{ ori}} = \\ &= \prod_{i=1}^N P(s_i)^{M \cdot P(s_i)} \end{aligned}$$

Cvasi-optimalitatea codarii aritmetice

- Numarul de simboluri binare pentru codarea latimii intervalului final se calculeaza cu relatia

$$k = -\log R_f = -\log \prod_{i=1}^N P(s_i)^{M \cdot P(s_i)} = -\sum_{i=1}^N M \cdot P(s_i) \log P(s_i) = M \cdot H(S)$$

de unde rezulta ca numarul mediu de simboluri binare pentru reprezentarea mesajului de lungime M este

$$\bar{k} = \frac{k}{M} = H(S)$$

adica limita Shannon

Exemplu

- Fie, de exemplu, o sursa cu doua simboluri: X si Y , cu probabilitatile $P(X) = 2/3$ si $P(Y) = 1/3$
- Daca se doreste codarea mesajelor de lungime 2 atunci se poate organiza urmatoarea imagine pentru codarea tuturor mesajelor in intervalul $[0,1)$



Exemplu

- Secventa XX este codata cu intervalul $[0, 4/9)$ si secventa YY este codata cu intervalul $[8/9, 1)$
- Pentru codarea intervalului poate fi transmis orice numar din intervalul considerat, evident pe un numar de biti corespunzator, ce este dat de marimea intervalului

| Message | | 0 | Codeword | |
|---------|----|---|----------|-------|
| X | XX | ← | 1/4 | .01 |
| | XY | ← | 2/4 | .10 |
| Y | YX | ← | 3/4 | .110 |
| | YY | ← | 15/16 | .1111 |
| | | | 1 | |

Exemplu

- Pentru codarea secvențelor de lungime 3 se poate construi imaginea / tabelul de codare de mai alaturat:

| | | | | | |
|---|----|-----|---|-------|--------|
| | | | 0 | | |
| X | XX | XXX | ← | 1/4 | .01 |
| | | XXY | ← | 3/8 | .011 |
| | XY | XYX | ← | 4/8 | .100 |
| | | XYY | ← | 10/16 | .1010 |
| Y | YX | YXX | ← | 6/8 | .110 |
| | | YXY | ← | 14/16 | .1110 |
| | YY | YYX | ← | 15/16 | .1111 |
| | | YYY | ← | 31/32 | .11111 |

Exemplu

$$n_{XX} \geq -\log\left(\frac{2}{3} \cdot \frac{2}{3}\right) = -\log\left(\frac{4}{9}\right) = 1,16 \rightarrow n_{XX} = 2$$

$$n_{XY} \geq -\log\left(\frac{2}{3} \cdot \frac{1}{3}\right) = -\log\left(\frac{2}{9}\right) = 2,16 \rightarrow n_{XY} = 3$$

$$n_{YX} \geq -\log\left(\frac{1}{3} \cdot \frac{2}{3}\right) = -\log\left(\frac{2}{9}\right) = 2,16 \rightarrow n_{YX} = 3$$

$$n_{YY} \geq -\log\left(\frac{1}{3} \cdot \frac{1}{3}\right) = -\log\left(\frac{1}{9}\right) = 3,16 \rightarrow n_{YY} = 4$$

- **XX**

$$\begin{array}{l} 0,25 \cdot 2 = 0,5 \rightarrow a_{-1} = 0 \\ 0,5 \cdot 2 = 1,0 \rightarrow a_{-2} = 1 \end{array} \left| \rightarrow code = (01)_2$$
- **XY**

$$\begin{array}{l} 0,5 \cdot 2 = 1,0 \rightarrow a_{-1} = 1 \\ 0,0 \cdot 2 = 0,0 \rightarrow a_{-2} = 0 \\ 0,0 \cdot 2 = 0,0 \rightarrow a_{-3} = 0 \end{array} \left| \rightarrow code = (100)_2$$
- **YX**

$$\begin{array}{l} 0,75 \cdot 2 = 1,5 \rightarrow a_{-1} = 1 \\ 0,5 \cdot 2 = 1,0 \rightarrow a_{-2} = 1 \\ 0,0 \cdot 2 = 0,0 \rightarrow a_{-3} = 0 \end{array} \left| \rightarrow code = (110)_2$$
- **YY**

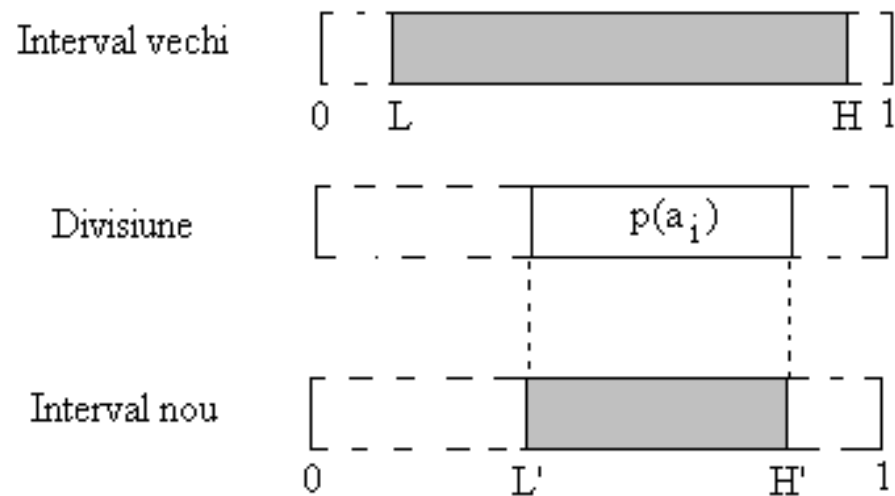
$$\begin{array}{l} 0,9375 \cdot 2 = 1,875 \rightarrow a_{-1} = 1 \\ 0,875 \cdot 2 = 1,75 \rightarrow a_{-2} = 1 \\ 0,75 \cdot 2 = 1,5 \rightarrow a_{-3} = 1 \\ 0,5 \cdot 2 = 1,0 \rightarrow a_{-4} = 1 \end{array} \left| \rightarrow code = (1111)_2$$

Algoritmul de baza

1. Se considera intervalul curent $[L,H)$ initializat cu $[0,1)$
2. Pentru fiecare simbol, a_i , din fisier, se parcurg doi pasi:
 1. Se imparte intervalul curent in subintervale, cate unul pentru fiecare simbol din alfabetul sursei; marimea subintervalului unui simbol din alfabet este proportional cu probabilitatea estimata, considerata in cadrul modelului
 2. Se selecteaza subintervalul corespunzator simbolului care apare in fisierul de comprimat, si se considera ca fiind noul interval curent
3. Se scriu in fisierul comprimat suficiente simboluri binare pentru a putea identifica (reconstitui) intervalul curent final din multimea tuturor intervalelor finale, posibil a fi considerate

Algoritmul de baza

- Diviziunea intervalului curent este bazata pe probabilitatea simbolului de intrare a_i , care apare in fisierul de comprimat:

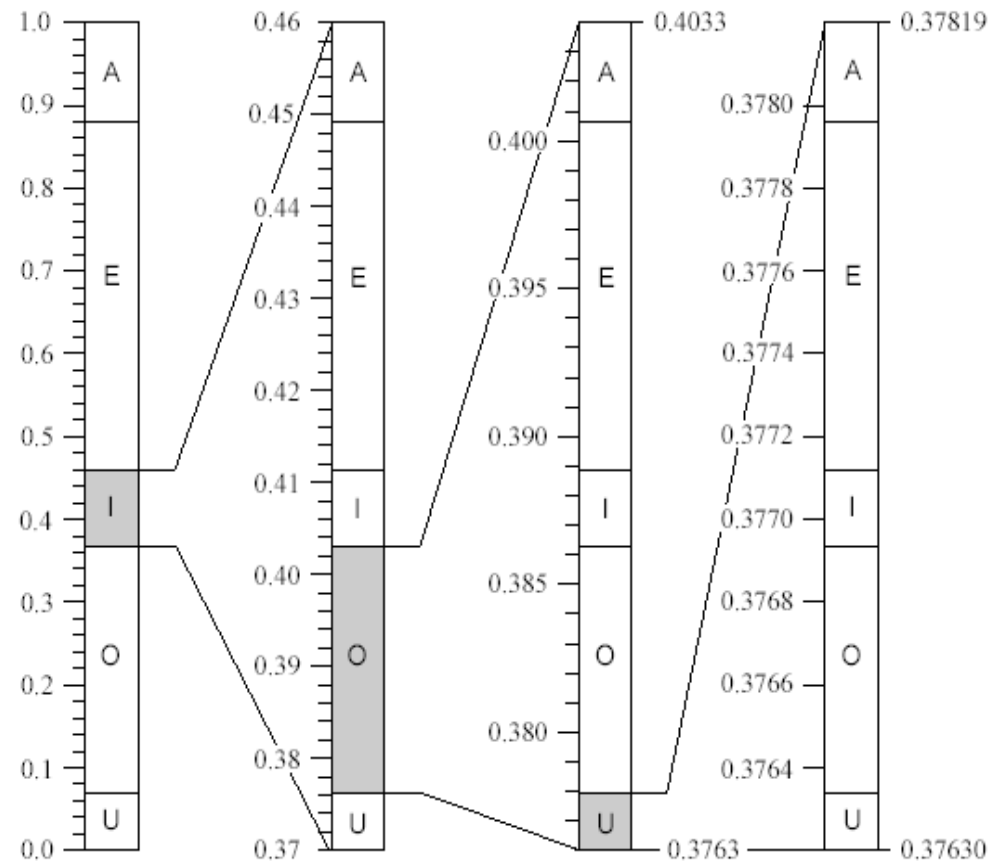


Exemplu

- Fie sursa data prin distributia

$$S: \begin{pmatrix} a & e & i & o & u \\ 0.12 & 0.42 & 0.09 & 0.3 & 0.07 \end{pmatrix}$$

- Figura alaturata arata modul de realizare a partiilor pentru mesajul „iou”



Observatii

- Lungimea intervalului final este egala cu produsul probabilitatilor simbolurilor individuale ce apar in fisierul/secventa de comprimat

$$(H - L)_{final} = p = \prod_{k=1}^{N_s} p(s_k)$$

unde N_s este numarul de simboluri din fisierul de intrare (cel care trebuie comprimat).

- Se calculeaza numarul de biti care trebuie memorati in fisierul comprimat (cel de iesire) dupa relatia

$$n = \lceil -\log_2 p \rceil$$

unde paranteza are semnificatia de parte intreaga, obtinuta prin rotunjire

- In acest fel, numarul de biti generati de codarea aritmetica este egal cu entropia, H (evident, daca nu se considera rotunjirea)
- In acest fel, codarea aritmetica atinge compresia care este indicata de entropia sursei, de este optima

Observatii

- Trebuie introdus un simbol suplimentar in alfabetul sursei, pentru a indica sfarsitul de fisier.
- La pasul 2 al algoritmului se calculeaza in mod efectiv numai subintervalul corespunzator simbolului care apare in fisier, de exemplu a_i

- Sunt necesare probabilitatile cumulative

$$P_c(i) = \sum_{k=1}^i p_k$$

- Fie k indicele alfabetului si i indicele simbolurilor mesajului/secventei
- Relatiile de calcul ale sub-intervalului nou sunt

$$L' = L + P_c(i-1) \cdot (H - L)$$

$$H' = L + P_c(i) \cdot (H - L)$$

unde L' si H' sunt limitele intervalului curent. Pentru primul pas se considera $L=0$ si $H=1$;

Observatii

- Intersectia intervalelor obtinute este vida pentru simboluri diferite, deci

$$I(s_i) \cap I(s_j) = \emptyset \quad \text{daca} \quad s_i \neq s_j$$

- In mod curent, codarea aritmetica poate fi implementata sub forma a doua module, care separa modelul sursei de procesul de codare
 - Modelul sursei este un modul de program si este consultat atat de codor cat si de decodor, la fiecare pas al procesarii
 - Posibilitatea separarii modelului in doua sub-modele, unul pentru estimarea probabilitatilor si unul pentru codare, permite cuplarea codarii aritmetice cu orice metoda statica sau dinamica de estimare a probabilitatilor (sau frecventelor) mesajelor sursei

Observatii

- Pentru realizarea decodarii trebuie sa se cunoasca modelul sursei utilizate de codor (mesajele si intervalele asociate) si un singur numar din intervalul determinat de codor
- Decodarea consta intr-o serie de comparatii a numarului i cu valorile reprezentand mesajele sursei

Exemplu

- Se considera o sursa cu doua simboluri, cu distributia de mai jos, in care se recunoaste introducerea unui simbol suplimentar EOF si normalizarea probabilitatilor la 1

$$S : \begin{pmatrix} a & b & EOF \\ 0.4 & 0.5 & 0.1 \end{pmatrix}$$

- Se va exemplifica codificarea/comprimarea mesajului „**bbbEOF**”

Exemplu

- Probabilitatea mesajului este: $0.5 \times 0.5 \times 0.5 = 0.125$
- Numarul de biti = $-\log_2 (0.125) = 6.3219 \rightarrow 7$ biti
- Intervalul final = $[0.8125000000000000 \quad 0.8250000000000000]$
- Mijlocul intervalului

$$x = (0.8125 + 0.8250) / 2 = 0.8187$$

$$(0.8187)_{10} = (1 \quad 1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0)_2$$

- Codul mesajului = **1 1 0 1 0 0 0**
- Marimea initiala a mesajului = $3 * 8 = 24$ biti = 3 octeti
- Marimea finala = 7 biti = 1 octet.
- Raport de compresie > 3