Diszkrét matematika 2.C szakirány

9. előadás

Nagy Gábor
nagygabr@gmail.com
nagy@compalg.inf.elte.hu
compalg.inf.elte.hu/∼nagy

Komputeralgebra Tanszék

2017. tavasz

Betűnkénti kódolás

Tétel (McMillan-egyenlőtlenség, NB)

Legyen $A = \{a_1, a_2, \ldots, a_n\}$ és B két ábécé, B elemeinek száma $r \geq 2$, és $\varphi : A \to B^+$ injektív leképezés.
Ha a φ által meghatározott betűnkénti kódolás felbontható, akkor $l_j = |\varphi(a_j)|$ jelöléssel

$$\sum_{j=1}^{n} r^{-l_j} \leq 1.$$

Tétel (McMillan-egyenlőtlenség ,,megfordítása", NB)

Az előző tétel jelöléseit használva, ha $l_1, l_2, \ldots, l_n$ olyan pozitív egész számok, hogy $\sum_{j=1}^{n} r^{-l_j} \leq 1$, akkor van az A-nak a B elemeivel való olyan felbontható (sőt prefix) kódolása, hogy az a_j betű kódjának hossza l_j.

Betűnkénti kódolás

Definíció

Legyen $A = \{a_1, a_2, \ldots, a_n\}$ a kódolandó ábécé, $p_1, p_2, \ldots, p_n$ a betűk eloszlása, $\varphi : A \to B^+$ injektív leképezés, továbbá $l_j = |\varphi(a_j)|$.
Ekkor $\bar{l} = \sum_{j=1}^{n} p_j l_j$ a kód átlagos szóhossza.
Ha adott elemszámú ábécével és eloszlással egy felbontható betűnkénti kód átlagos szóhosszúsága minimális, akkor optimális kódnak nevezzük.

Megjegyzés

Az átlagos kódhossz valós szám, és valós számok halmazában nem feltétlenül van minimális elem (ld. $\{\frac{1}{n} | n \in \mathbb{N}\}$), ezért optimális kód létezése nem triviális.

Betűnkénti kódolás

Állítás

Adott ábécé és eloszlás esetén létezik optimális kód.

Bizonyítás

Válasszunk egy tetszőleges felbontható kódot (Miért van ilyen?), ennek átlagos szóhosszúsága legyen l. Mivel $p_j l_j > l$ esetén a kód nem lehet optimális (Miért?), ezért elég azokat a kódokat tekinteni, amelyekre $l_j \leq \frac{l}{p_j}$, ha $j = 1, 2, \ldots, n$. Ilyen kód csak véges sok van, így van köztük minimális átlagos hosszúságú.

Betűnkénti kódolás

Tétel (Shannon tétele zajmentes csatornára)

Legyen $A = \{a_1, a_2, \ldots, a_n\}$ a kódolandó ábécé, $p_1, p_2, \ldots, p_n$ a betűk eloszlása, $\varphi : A \to B^+$ injektív leképezés, B elemeinek a száma $r \geq 2$, továbbá $l_j = |\varphi(a_j)|$.
Ha a φ által meghatározott betűnkénti kódolás felbontható, akkor $H_r(p_1, p_2, \ldots, p_n) \leq \bar{l}$.

Bizonyítás

$$\bar{l} - H_r(p_1, p_2, \ldots, p_n) = \sum_{j=1}^{n} p_j l_j + \sum_{j=1}^{n} p_j \log_r p_j =$$

$$= \sum_{j=1}^{n} p_j \cdot \left(-\log_r(r^{-l_j})\right) + \sum_{j=1}^{n} p_j \cdot \left(-\log_r \frac{1}{p_j}\right) = \sum_{j=1}^{n} p_j \cdot \left(-\log_r \frac{r^{-l_j}}{p_j}\right) \geq$$

$$\geq -\log_r \left(\sum_{j=1}^{n} r^{-l_j}\right) \geq -\log_r 1 = 0$$

Betűnkénti kódolás

Tétel (Shannon kód létezése)

Az előző tétel jelöléseivel, ha $n > 1$, akkor van olyan prefix kód, amire $\bar{l} < H_r(p_1, p_2, \ldots, p_n) + 1$.

Bizonyítás

Válasszunk olyan $l_1, l_2, \ldots, l_n$ természetes számokat, amelyekre $r^{-l_j} \leq p_j < r^{-l_j+1}$, ha $j = 1, 2, \ldots, n$ (Miért tudunk ilyeneket választani?). Ekkor $\sum_{j=1}^{n} r^{-l_j} \leq \sum_{j=1}^{n} p_j = 1$, így a McMillan-egyenlőtlenség megfordítása miatt létezik prefix kód az adott l_j hosszakkal. Mivel $l_j < 1 - \log_r p_j$ (Miért?), ezért

$$\bar{l} = \sum_{j=1}^{n} p_j l_j < \sum_{j=1}^{n} p_j (1 - \log_r p_j) = 1 + H_r(p_1, p_2, \ldots, p_n).$$

Optimális kódkonstrukció: Huffman-kód

Legyen $\{a_1, a_2, \ldots, a_n\}$ az üzenetek halmaza, a hozzájuk tartozó eloszlás pedig $\{p_1, p_2, \ldots, p_n\}$, a kódábécé elemszáma r.
Rendezzük relatív gyakoriság szerint csökkenő sorrendbe a betűket.
Osszuk el maradékosan $n-2$-t $r-1$-gyel:
$n-2 = q(r-1)+m \quad 0 \leq m < r-1$, és legyen $t = m+2$.
Helyettesítsük az utolsó t betűt egy új betűvel, amihez az elhagyott betűk relatív gyakoriságainak összegét rendeljük, és az így kapott gyakoriságoknak megfelelően helyezzük el az új betűt a sorozatban.
Ezek után ismételjük meg az előző redukciót, de most már minden lépésben r betűvel csökkentve a kódolandó halmazt, mígnem már csak r betű marad.
Most a redukált ábécé legfeljebb r betűt tartalmaz, és ha volt redukció, akkor pontosan r-et.
Ezeket a kódoló ábécé elemeivel kódoljuk, majd a redukciónak megfelelően visszafelé haladva, az összevont betűk kódját az összevonásként kapott betű már meglévő kódjának a kódoló ábécé különböző betűivel való kiegészítésével kapjuk.

Példa Huffman-kódra

Legyen $A = \{a, b, \ldots, j\}$, a relatív gyakoriságok $0,17; 0,02; 0,13; 0,02; 0,01; 0,31; 0,02; 0,17; 0,06; 0,09$, a kódoló ábécé pedig $\{0, 1, 2\}$. $10 - 2 = 4 \cdot (3 - 1) + 0$, így $t = 0 + 2 = 2$.

f	0,31		f	0,31		f	0,31		
a	0,17		a	0,17		a	0,17		
h	0,17		h	0,17		h	0,17		
c	0,13		c	0,13		c	0,13		
j	0,09		j	0,09		j	0,09		
i	0,06		i	0,06		((g,e),b,d)	0,07	}	0, 22
b	0,02		(g,e)	0,03		i	0,06		
d	0,02		b	0,02	} 0, 07				
g	0,02	} 0, 03	d	0,02					
e	0,01								

f	0,31		(a,h,c)	0,47	
(j,((g,e),b,d),i)	0,22		f	0,31	
a	0,17		(j,((g,e),b,d),i)	0,22	
h	0,17	} 0, 47			
c	0,13				

Példa Huffman-kódra folyt.

(a,h,c)	0,47
f	0,31
(j,((g,e),b,d),i)	0,22

Kódolás:

$(a,h,c) \mapsto 0$

$a \mapsto 00$

$h \mapsto 01$

$c \mapsto 02$

$f \mapsto 1$

$(j,((g,e),b,d),i) \mapsto 2$

$j \mapsto 20$

$((g,e),b,d) \mapsto 21$

$(g,e) \mapsto 210$

$g \mapsto 2100$

$e \mapsto 2101$

$b \mapsto 211$

$d \mapsto 212$

$i \mapsto 22$

Entrópia: $\approx 1,73$.

Átlagos szóhossz: $1,79$.

Betűnkénti kódolás

Tétel (NB)

A Huffman-kód optimális.

Példa Shannon-kódra

Az előző példában használt ábécét és eloszlást fogjuk használni. Rendezzük sorba az ábécét relatív gyakoriságok szerinti csökkenő sorrendben:

f	0,31
a	0,17
h	0,17
c	0,13
j	0,09
i	0,06
b	0,02
d	0,02
g	0,02
e	0,01

Példa Shannon-kódra folyt.

Határozzuk meg a szükséges szóhosszúságokat:
$\frac{1}{9} \leq 0,31; 0,17; 0,13 < \frac{1}{3}$, ezért f, a, h és c kódhossza 2.
$\frac{1}{27} \leq 0,09; 0,06 < \frac{1}{9}$, ezért j és i kódhossza 3.
$\frac{1}{81} \leq 0,02 < \frac{1}{27}$, ezért b, d és g kódhossza 4.
$\frac{1}{243} \leq 0,01 < \frac{1}{81}$, ezért e kódhossza 5.
Az f kódja 00, az a kódja 01, a h kódja 02, és ez utóbbihoz 1-et adva hármas alapú számrendszerben kapjuk c kódját, ami 10. Ehhez 1-et adva 11-et kapunk, de j kódjának hossza 3, ezért ezt még ki kell egészíteni jobbról egy 0-val, tehát j kódja 110. Hasonlóan folytatva megkapjuk a teljes kódot:

f	00
a	01
h	02
c	10
j	110
i	111
b	1120
d	1121
g	1122
e	12000

Átlagos szóhossz: $2,3 < 1,73 + 1$.

Betűnkénti kódolás

Kódfa

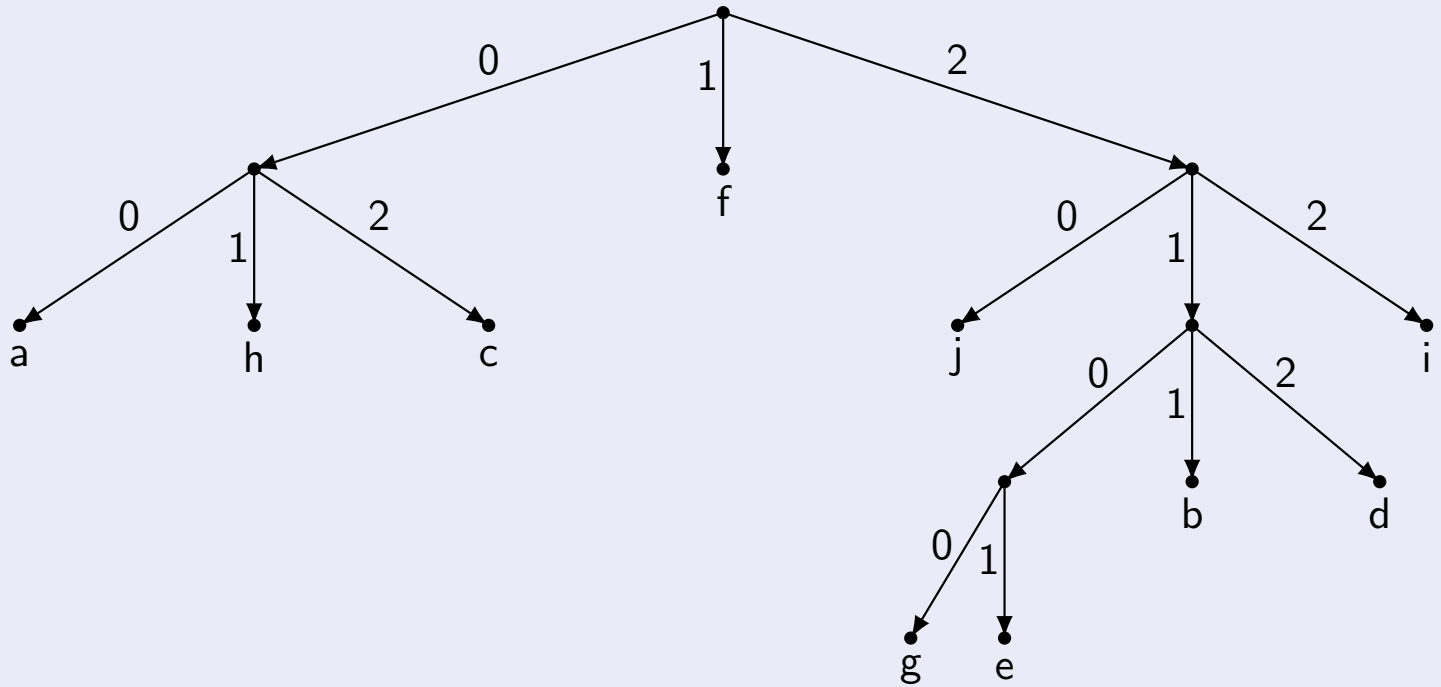
A betűnkénti kódolás szemléltethető egy címkézett irányított fával.
Legyen $\varphi : A \to B^*$ egy betűnkénti kódolás, és tekintsük $\mathrm{rng}(\varphi)$ prefixeinek halmazát. Ez a halmaz részbenrendezett a ,,prefixe" relációra. Vegyük ennek a Hasse-diagramját. Így egy irányított fát kapunk, aminek a gyökere az üres szó, és minden szó a hosszának megfelelő szinten van.
A fa éleit címkézzük úgy B elemeivel, hogy ha $\beta = \alpha b$ valamely $b \in B$-re, akkor az α-ból β-ba vezető él címkéje legyen b.
A kódfa csúcsait is megcímkézhetjük: az $a \in A$ kódjának megfelelő csúcs címkéje legyen $a \in A$; azon csúcs címkéje, amely nincsen $\mathrm{rng}(\varphi)$-ben, legyen ,,üres".

Megjegyzés

Az előbbi konstrukció meg is fordítható. Tekintsünk egy véges, élcímkézett irányított fát, ahol az élcímkék halmaza B, az egy csúcsból kiinduló élek mind különböző címkéjűek, továbbá az A véges ábécének a csúcsokra való leképezését, amelynél minden levél előáll képként.
Az $a \in A$ betű kódja legyen az a szó, amelyet úgy kapunk, hogy a gyökértől az a-nak megfelelő csúcsig haladó irányított út mentén összeolvassuk az élek címkéit.

Kódfa

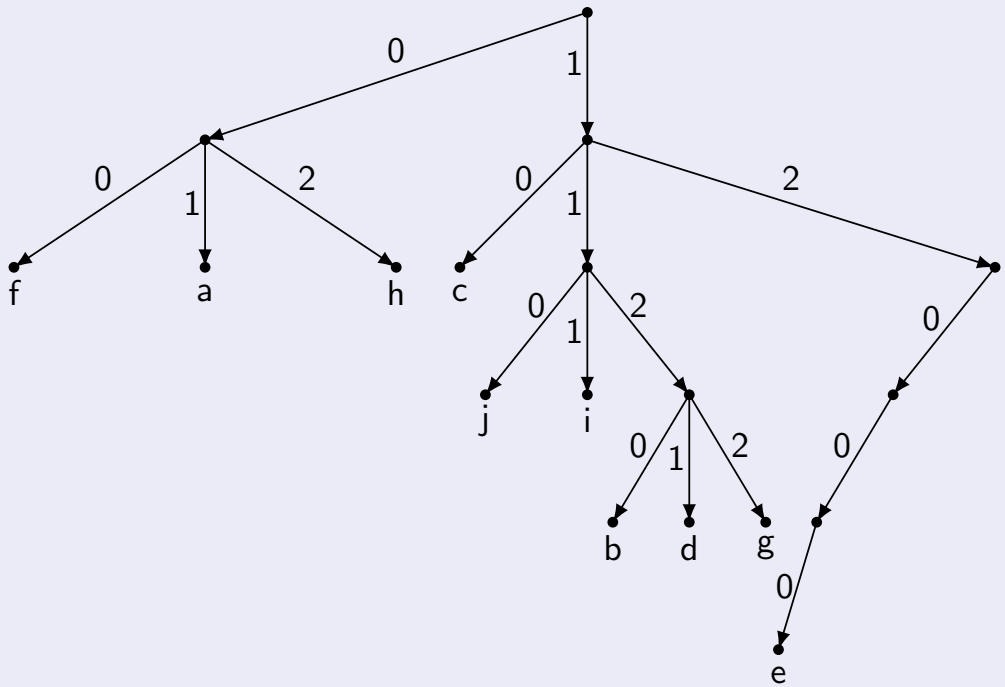
Példa



A Huffman-kódos példában szereplő kódhoz tartozó kódfa.
$\varphi(a) =$00, $\varphi(b) =$211, $\varphi(c) =$02, $\varphi(d) =$212, $\varphi(e) =$2101, $\varphi(f) =$1, $\varphi(g) =$2100, $\varphi(h) =$01, $\varphi(i) =$22, $\varphi(j) =$20.
A kódszavak prefixeinek halmaza:
$\{\lambda, 1, 00, 0, 01, 02, 20, 2, 22, 211, 21, 212, 2100, 210, 2101\}$

Kódfa

Példa



A Shannon-kódos példában szereplő kódhoz tartozó kódfa.
$\varphi(a) =$01, $\varphi(b) =$1120, $\varphi(c) =$10, $\varphi(d) =$1121, $\varphi(e) =$12000,
$\varphi(f) =$00, $\varphi(g) =$1122, $\varphi(h) =$02, $\varphi(i) =$111, $\varphi(j) =$110.
A kódszavak prefixeinek halmaza:
$\{01, 0, \lambda, 1120, 112, 11, 1, 10, 1121, 12000, 1200, 120, 12, 00, 1122, 02, 111, 110\}$

Hibakorlátozó kódolás

Példa (ISBN (International Standard Book Number) kódolása)

Legyen $d_1, d_2, \ldots, d_n$ decimális számjegyek egy sorozata ($n \leq 10$). Egészítsük ki a sorozatot egy $n+1$-edik számjeggyel, amelynek értéke

$$d_{n+1} = \sum_{j=1}^{n} j \cdot d_j \mod 11,$$

ha az nem 10, különben d_{n+1} legyen X.
Ha valamelyik számjegyet elírjuk, akkor az összefüggés nem teljesülhet: d_{n+1} elírása esetén ez nyilvánvaló, $j \leq n$-re d_j helyett d'_j-t írva pedig az összeg $j(d'_j - d_j)$-vel nőtt, ami nem lehet 11-gyel osztható (Miért?).
Azt is észrevesszük, ha $j < n$ esetén d_j-t és d_{j+1}-et felcseréljük:
az összeg $jd_{j+1} + (j+1)d_j - jd_j - (j+1)d_{j+1} = d_j - d_{j+1}$-gyel nő, ami csak akkor lehet 11-gyel osztható, ha $d_j = d_{j+1}$.

Megjegyzés

2007 óta 13 jegyű.
A személyi számnál is használják.

Hibakorlátozó kódolás

Példa (Paritásbites kód)

Egy n hosszú 0-1 sorozatot egészítsünk ki egy $n+1$-edik bittel, ami legyen 1, ha a sorozatban páratlan sok 1-es van, különben pedig legyen 0. Ha egy bit megváltozik, akkor észleljük a hibát.

Példa (Kétdimenziós paritásellenőrzés)

$$\begin{array}{ccccc|c} b_{0,0} & \cdots & b_{0,j} & \cdots & b_{0,n-1} & b_{0,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ b_{i,0} & \cdots & b_{i,j} & \cdots & b_{i,n-1} & b_{i,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ b_{m-1,0} & \cdots & b_{m-1,j} & \cdots & b_{m-1,n-1} & b_{m-1,n} \\ \hline b_{m,0} & \cdots & b_{m,j} & \cdots & b_{m,n-1} & b_{m,n} \end{array}$$

Oszlopok és sorok végén paritásbit. Ha megváltozik egy bit, akkor a sor és az oszlop végén jelez az ellenőrző bit, ez alapján tudjuk javítani a hibát. Ha két bit változik meg, akkor észleljük a hibát, de nem tudjuk javítani.

Hibakorlátozó kódolás

Definíció

Egy kód t-hibajelző, ha minden olyan esetben jelez, ha az elküldött és megkapott szó legfeljebb t helyen tér el.
Egy kód pontosan t-hibajelző, ha t-hibajelző, de van olyan $t+1$-hiba, amit nem jelez.

Példa

- ISBN - 1-hibajelző
- paritásbites kód - 1-hibajelző
- kétdimenziós paritásellenőrzés - 2-hibajelző

Hiba javításának módjai

ARQ (Automatic Retransmission Request) - újraküldés,
FEC (Forward Error Correction) - javítható, pl.: kétdimenziós paritásell.

Hibakorlátozó kódolás

Definíció

Legyen A véges ábécé, továbbá $u, v \in A^n$. Ekkor u és v Hamming-távolsága alatt az azonos pozícióban lévő különböző betűk számát értjük:

$$d(u, v) = |\{i : 1 \leq i \leq n \wedge u_i \neq v_i\}|.$$

Példa

0	1	1	1	0
1	0	1	0	1
$\neq$	$\neq$	$=$	$\neq$	$\neq$

$d(01110,10101)=4$

A	L	M	A
A	N	N	A
$=$	$\neq$	$\neq$	$=$

$d(ALMA,ANNA)=2$