Diszkrét matematika II.
8. előadás

Fancsali Szabolcs Levente
nudniq@inf.elte.hu

ELTE IK Komputeralgebra Tanszék

Mérai László diái alapján

A kommunikáció során információt hordozó adatokat viszünk át egy csatornán keresztül az információforrástól, az adótól az információ címzettjéhez, a vevőhöz.

$$\boxed{\text{Adó}} \rightarrow \boxed{\text{Csatorna}} \rightarrow \boxed{\text{Vevő}}$$

A kommunikáció vázlatos ábrája

Megjegyzés

Az információ átvitele térben és időben történik. Egyes esetekben az egyik, más esetekben a másik dimenzió a domináns (pl. telefonálás; információ rögzítése adathordozóra, majd későbbi visszaolvasása).

Definíció

Az információ új ismeret. Shannon nyomán az általa megszüntetett bizonytalansággal mérjük.

Definíció

Tegyük fel, hogy egy információforrás nagy számú, összesen n üzenetet bocsát ki. Az összes ténylegesen előforduló különböző üzenet legyen $a_1, a_2, \ldots, a_k$.
Ha az a_j üzenet m_j-szer fordul elő, akkor azt mondjuk, hogy a gyakorisága m_j, relatív gyakorisága pedig $p_j = \frac{m_j}{n} > 0$.
A $p_1, p_2, \ldots, p_k$ szám k-ast az üzenetek eloszlásának nevezzük ($\sum_{j=1}^{k} p_j = 1$).
Az a_j üzenet egyedi információtartalma $I_j = -\log_r p_j$, ahol r egy 1-nél nagyobb valós szám, ami az információ egységét határozza meg. Ha $r = 2$, akkor az információ egysége a bit.
Az üzenetforrás által kibocsátott üzenetek átlagos információtartalma, vagyis $H_r(p_1, p_2, \ldots, p_k) = -\sum_{j=1}^{k} p_j \log_r p_j$ a forrás entrópiája. Ez csak az üzenetek eloszlásától függ, a tartalmuktól nem.
Egy k tagú eloszlásnak olyan pozitív valós számokból álló $p_1, p_2, \ldots, p_k$ sorozatot nevezünk, amelyre $\sum_{j=1}^{k} p_j = 1$. Ennek az eloszlásnak az entrópiája $H_r(p_1, p_2, \ldots, p_k) = -\sum_{j=1}^{k} p_j \log_r p_j$.

Definíció

Legyen $I \subset \mathbb{R}$ egy intervallum. Az $f : I \to \mathbb{R}$ függvényt konvexnek nevezzük, ha bármely $x_1, x_2 \in I$ és $0 \leq t \leq 1$ esetén

$$f\,(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2).$$

f szigorúan konvex, ha egyenlőség csak $t = 0$ vagy $t = 1$ esetén lehetséges.

Lemma (Jensen-egyenlőtlenség, NB)

Legyen $p_1, p_2, \ldots, p_k$ egy eloszlás, $f : I \to \mathbb{R}$ pedig egy szigorúan konvex függvény az $I \subset \mathbb{R}$ intervallumon. Ekkor $q_1, q_2, \ldots, q_k \in I$ esetén

$$f\left(\sum_{j=1}^{k} p_j q_j\right) \leq \sum_{j=1}^{k} p_j f(q_j),$$

és egyenlőség pontosan akkor áll fenn, ha $q_1 = q_2 = \ldots = q_k$.

Tétel

Bármilyen eloszláshoz tartozó entrópiára

$$H_r(p_1, p_2, \ldots, p_k) \leq \log_r k,$$

és egyenlőség pontosan akkor teljesül, ha $p_1 = p_2 = \ldots = p_k = \frac{1}{k}$.

Bizonyítás

$r > 1$ esetén a $-\log_r(x)$ függvény szigorúan konvex, ezért használhatjuk a lemmát $q_j = \frac{1}{p_j}$ választással:

$$-H_r(p_1, p_2, \ldots, p_k) = \sum_{j=1}^{k} p_j \log_r p_j =$$

$$= \sum_{j=1}^{k} p_j \left(-\log_r \frac{1}{p_j}\right) \geq -\log_r \left(\sum_{j=1}^{k} p_j \frac{1}{p_j}\right) = -\log_r k.$$

Definíció

A kódolás alatt a legáltalánosabb értelemben az üzenetek halmazának egy másik halmazba való leképezését értjük.
Ha a leképezés injektív, akkor azt mondjuk, hogy a kódolás felbontható, egyértelműen dekódolható, vagy veszteségmentes, egyébként veszteségesnek nevezzük, mert információvesztéssel jár.

Betűnkénti kódolás

A betűnkénti kódolás során az üzenetet meghatározott módon egymáshoz átfedés nélkül csatlakozó részekre bontjuk, egy-egy ilyen részt egy szótár alapján kódolunk, és az így kapott kódokat az eredeti sorrendnek megfelelően egymáshoz láncoljuk.
Az általánosság csorbítása nélkül feltehetjük, hogy a szótár alapján kódolandó elemi üzenetek egy A ábécé (a kódolandó ábécé) betűi, és egy-egy ilyen betű kódja egy másik (az előbbitől nem feltétlenül különböző) B ábécé (kódoló ábécé vagy kódábécé) betűivel felírt szó, vagyis ezen ábécéből vett betűk véges sorozata, a sorozat elemeit egyszerűen egymás mellé írva. Az ábécékről feltesszük, hogy nem-üresek és végesek.

Definíció

Az A ábécé betűivel felírható összes (legalább egy betűt tartalmazó) szó halmazát A^+ jelöli, míg az egyetlen betűt sem tartalmazó üres szóval (jele: $\emptyset$ vagy λ) kibővített halmazt A^*.

Betűnkénti kódolás

Definíció

A betűnkénti kódolást egy $\varphi : A \to B^*$ leképezés határozza meg, amelyet természetes módon terjesztünk ki egy $\psi : A^* \to B^*$ leképezéssé:
$a_1 a_2 \dots a_n = \alpha \in A^*$ esetén $\psi(\alpha) = \varphi(a_1)\varphi(a_2)\dots\varphi(a_n)$.
$\operatorname{rng}(\psi)$-t **kódnak** nevezzük, elemei a **kódszavak**.

Megjegyzés

Ha φ nem injektív, vagy az üres szó benne van az értékkészletében, akkor a kapott ψ kódolás nem injektív (Miért?), tehát nem felbontható, ezért betűnkénti kódolásnál feltesszük, hogy φ injektív, és B^+-ba képez.

Betűnkénti kódolás

Definíció

Tekintsünk egy A ábécét, és legyen $\alpha, \beta, \gamma \in A^*$. Ekkor α prefixe (előtagja), míg γ szuffixe (utótagja) $\alpha\gamma$-nak, β pedig infixe (belső tagja) $\alpha\beta\gamma$-nak.

Definíció

Prefixmentes halmaznak nevezzük szavak egy halmazát, ha nincs benne két különböző szó, hogy egyik a másik prefixe.

Definíció

Az üres szó és α prefixe, szuffixe és infixe is α-nak, ezeket α triviális prefixeinek, triviális szuffixeinek és triviális infixeinek nevezzük.

Definíció

α egy prefixét, szuffixét, illetve infixét valódi prefixnek, valódi szuffixnek, illetve valódi infixnek nevezzük, ha nem egyezik meg α-val.

Betűnkénti kódolás

Definíció

Tekintsük az injektív $\varphi : A \to B^+$ leképezést, illetve az általa meghatározott ψ betűnkénti kódolást.
Ha $\mathrm{rng}(\varphi)$ prefixmentes halmaz, akkor prefix kódról beszélünk.
Ha $\mathrm{rng}(\varphi)$ elemei azonos hosszúságúak, akkor egyenletes kódról, fix hosszúságú kódról, esetleg blokk-kódról beszélünk.
Vesszős kódról beszélünk, ha van egy olyan $\vartheta \in B^+$ szó (a vessző), amely minden kódszónak szuffixe, de egyetlen kódszó sem áll elő $\alpha\vartheta\beta$ alakban nem üres β szóval.

Állítás

Prefix kód felbontható.

Bizonyítás

Konstruktív: nézzük az eddig beérkezett szimbólumokból összeálló szót. Amint ez kiadja a kódolandó ábécé valamely betűjének a kódját, azonnal dekódolhatunk a megfelelő betűre, mert a folytatásával kapott jelsorozat egyetlen betűnek sem lehet a kódja.

Betűnkénti kódolás

Állítás

Egyenletes kód prefix (így nyilván felbontható is).

Bizonyítás

Mivel a kódszavak hossza azonos, ezért csak úgy lehet egy kódszó prefixe egy másiknak, ha megegyeznek.

Állítás

Vesszős kód prefix (így nyilván felbontható is).

Bizonyítás

A vessző egyértelműen jelzi egy kódszó végét, hiszen ha folytatva kódszót kapnánk, abban a vessző tiltott módon szerepelne.

Betűnkénti kódolás

Példák

Legyen $A = \{\text{a,b,c}\}$, $B = \{0,1\}$, $\varphi : A \to B^+$ pedig az alábbi módon definiált.

	1.	2.	3.	4.	5.	6.
$\varphi(\text{a})$	01	1	01	0	00	01
$\varphi(\text{b})$	1101	01	011	10	10	001
$\varphi(\text{c})$	01	10	11	11	11	0001

1. $\varphi(\text{a}) = \varphi(\text{c}) \implies \varphi$ nem injektív
2. $\psi(\text{ab}) = 101 = \psi(\text{ca}) \implies$ nem felbontható
3. nem prefix, de felbontható
4. prefix
5. egyenletes
6. vesszős

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 $\ell_j = |\varphi(a_j)|$ jelöléssel

$$\sum_{j=1}^{n} \frac{1}{r^{\ell_j}} \leq 1.$$

Tétel (McMillan-egyenlőtlenség „megfordítása", NB)

Az előző tétel jelöléseit használva, ha $\ell_1, \ell_2, \ldots, \ell_n$ olyan pozitív egész számok, hogy $\sum_{j=1}^{n} r^{-\ell_j} \leq 1$, 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

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 $\ell_j = |\varphi(a_j)|$ jelöléssel

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

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

Az előző tétel jelöléseit használva, ha $\ell_1, \ell_2, \ldots, \ell_n$ olyan pozitív egész számok, hogy $\sum_{j=1}^{n} r^{-\ell_j} \leq 1$, 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 $\ell_j = |\varphi(a_j)|$.
Ekkor $\overline{\ell} = \sum_{j=1}^{n} p_j \ell_j$ 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 $p_j \ell_j > \ell$ esetén a kód nem lehet optimális (Miért?), ezért elég azokat a kódokat tekinteni, amelyekre $\ell_j \leq \frac{\ell}{p_j}$, 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 $\ell_j = |\varphi(a_j)|$.
Ha a φ által meghatározott betűnkénti kódolás felbontható, akkor $H_r(p_1, p_2, \ldots, p_n) \leq \overline{\ell}$.

Bizonyítás

$$\overline{\ell} - H_r(p_1, p_2, \ldots, p_n) = \sum_{j=1}^{n} p_j \ell_j + \sum_{j=1}^{n} p_j \log_r p_j =$$

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

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

Betűnkénti kódolás

Tétel (Shannon kód létezése)

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

Bizonyítás

Válasszunk olyan $\ell_1, \ell_2, \ldots, \ell_n$ természetes számokat, amelyekre $r^{-\ell_j} \leq p_j < r^{-\ell_j+1}$, ha $j = 1, 2, \ldots, n$ (Miért tudunk ilyeneket választani?). Ekkor $\sum_{j=1}^{n} r^{-\ell_j} \leq \sum_{j=1}^{n} p_j = 1$, így a McMillan-egyenlőtlenség megfordítása miatt létezik prefix kód az adott ℓ_j hosszakkal. Mivel $\ell_j < 1 - \log_r p_j$ (Miért?), ezért

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

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	
a	0,17	
h	0,17	
c	0,13	
j	0,09	
i	0,06	
b	0,02	
d	0,02	
g	0,02	$\Big\}$ 0, 03
e	0,01	

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

f	0,31	
a	0,17	
h	0,17	
c	0,13	
j	0,09	
((g,e),b,d)	0,07	$\Big\}$ 0, 22
i	0,06	

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

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

Példa Huffman-kódra folyt.

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

Kódolás:

$(a,h,c) \mapsto 0$ $\quad a \mapsto 00$

$h \mapsto 01$

$c \mapsto 02$

$f \mapsto 1$

$(j,((g,e),b,d),i) \mapsto 2$ $\quad j \mapsto 20$

$((g,e),b,d) \mapsto 21$ $\quad (g,e) \mapsto 210$ $\quad g \mapsto 2100$

$e \mapsto 2101$

$b \mapsto 211$

$d \mapsto 212$

$i \mapsto 22$

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

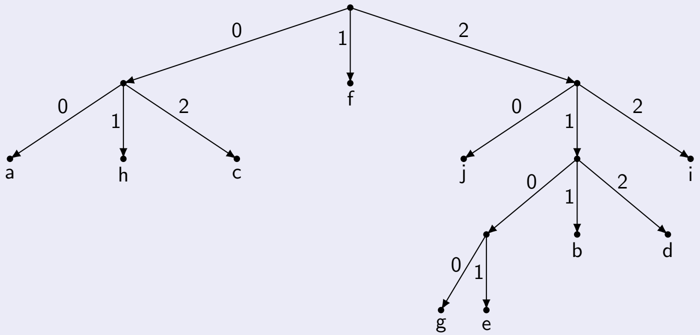
A betűnkénti kódolás szemléltethető egy címkézett irányított fával.
Legyen $\varphi : A \to B^*$ egy betűnkénti kódolás, és tekintsük $\mathrm{rng}(\varphi)$ prefixeinek halmazát. Ez a halmaz részbenrendezett a ,,prefixe” relációra. Vegyük ennek a Hasse-diagramját. Így egy irányított fát kapunk, aminek 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

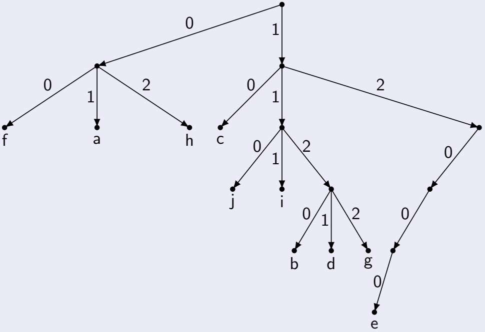
Példa



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:
$\{01, 0, \lambda, 1120, 112, 11, 1, 10, 1121, 12000, 1200, 120, 12, 00, 1122, 02, 111, 110\}$