Komprese dat

Jan Outrata

KATEDRA INFORMATIKY
UNIVERZITA PALACKÉHO V OLOMOUCI

přednášky

Literatura

- Sayood K.: *Introduction to Data Compression*, Fourth Edition. Morgan Kaufmann, 2012. ISBN 978-0124157965
- Salomon D., Motta G.: *Handbook of Data Compression*, 5th edition. Springer, 2010. ISBN 978-1848829022
- Salomon D.: *Data Compression: The complete Reference*, 4th edition. Springer, 2006. ISBN 978-1846286025
- Hankerson D. C., Harris G. A., Johnson P. D.: *Introduction to Information Theory and Data Compression*, Second Edition (Applied Mathematics). Chapman and Hall/CRC, 2003. ISBN 978-1584883135
- Sayood K.: *Lossless compression handbook*. Academic Press, 2003. ISBN 0126208611

Úvod

Komprese dat

= zmenšení velikosti reprezentace obsahu/dat – jeden z účelů kódování dat, (experimentální) vědní obor

Dvě fáze:

1. identifikování a modelování struktury dat s vynecháním redundancí
   - struktura např. opakování vzorů, statistická ≈ frekvence/četnost vzorů, korelace mezi vzory, vzory elementární symboly nebo skupiny symbolů, také např. daná zdrojem dat → modelování zdroje a syntéza dat (zvuk)
   - také různé modely pro různé části dat
2. kódování dat podle modelu
   - plus případně kódování (části) modelu
   - také predikce hodnoty dle modelu a kódování rozdílu (residua)
   - typicky binární kód

Komprese dat

Příklad

$x_1, x_2, \ldots, x_{12} = 2, 2, 4, 6, 7, 7, 7, 10, 10, 11, 11, 14$

Komprese dat

Příklad

$x_1, x_2, \ldots, x_{12} = 2, 2, 4, 6, 7, 7, 7, 10, 10, 11, 11, 14$

1. číslo ve dvojkové soustavě $\Rightarrow 4\,$b/číslo $= 48\,$b

Komprese dat

Příklad

$x_1, x_2, \ldots, x_{12} = 2, 2, 4, 6, 7, 7, 7, 10, 10, 11, 11, 14$

1. číslo ve dvojkové soustavě $\Rightarrow 4\,\mathrm{b}/\text{číslo} = 48\,\mathrm{b}$
2. 7 různých čísel ve dvojkové soustavě $\Rightarrow 3\,\mathrm{b}/\text{číslo} = 36\,\mathrm{b}$

Komprese dat

Příklad

$x_1, x_2, \ldots, x_{12} = 2, 2, 4, 6, 7, 7, 7, 10, 10, 11, 11, 14$

1. číslo ve dvojkové soustavě $\Rightarrow$ 4 b/číslo $=$ 48 b
2. 7 různých čísel ve dvojkové soustavě $\Rightarrow$ 3 b/číslo $=$ 36 b
3. častější číslo kratší kód $\rightarrow 2 \times 2,\ 1 \times 4,\ 1 \times 6,\ 3 \times 7,\ 2 \times 10,\ 2 \times 11,\ 1 \times 14 \rightarrow$ $\mathbf{0I}$ pro 7, $\mathbf{III}$ pro 11, $\mathbf{II0}$ pro 10, $\mathbf{I0I}$ pro 2, $\mathbf{I00}$ pro 14, $\mathbf{000}$ pro 4 a $\mathbf{00I}$ pro 6 $\Rightarrow$ 33 b $=$ 2.75 b/číslo

Komprese dat

Příklad

$x_1, x_2, \dots, x_{12} = 2, 2, 4, 6, 7, 7, 7, 10, 10, 11, 11, 14$

1. číslo ve dvojkové soustavě $\Rightarrow 4\,\mathrm{b}/\text{číslo} = 48\,\mathrm{b}$
2. 7 různých čísel ve dvojkové soustavě $\Rightarrow 3\,\mathrm{b}/\text{číslo} = 36\,\mathrm{b}$
3. častější číslo kratší kód $\rightarrow 2 \times 2,\ 1 \times 4,\ 1 \times 6,\ 3 \times 7,\ 2 \times 10,\ 2 \times 11,\ 1 \times 14 \rightarrow$ $\mathbf{0I}$ pro 7, $\mathbf{III}$ pro 11, $\mathbf{II0}$ pro 10, $\mathbf{I0I}$ pro 2, $\mathbf{I00}$ pro 14, $\mathbf{000}$ pro 4 a $\mathbf{00I}$ pro 6 $\Rightarrow 33\,\mathrm{b} = 2.75\,\mathrm{b}/\text{číslo}$
4. kódování opakování čísla $\rightarrow$ $\mathbf{0}$ pro žádné, $\mathbf{I0}$ pro jedno a $\mathbf{II}$ pro dvě $\Rightarrow$ $7 \times 3 + 11 = 32\,\mathrm{b} = 2.\overline{6}\,\mathrm{b}/\text{číslo}$

Komprese dat

Příklad

$x_1, x_2, \dots, x_{12} = 2, 2, 4, 6, 7, 7, 7, 10, 10, 11, 11, 14$

1. číslo ve dvojkové soustavě $\Rightarrow 4\,\text{b/číslo} = 48\,\text{b}$
2. 7 různých čísel ve dvojkové soustavě $\Rightarrow 3\,\text{b/číslo} = 36\,\text{b}$
3. častější číslo kratší kód $\rightarrow 2 \times 2,\ 1 \times 4,\ 1 \times 6,\ 3 \times 7,\ 2 \times 10,\ 2 \times 11,\ 1 \times 14 \rightarrow$ $\mathbf{0I}$ pro 7, $\mathbf{III}$ pro 11, $\mathbf{II0}$ pro 10, $\mathbf{I0I}$ pro 2, $\mathbf{I00}$ pro 14, $\mathbf{000}$ pro 4 a $\mathbf{00I}$ pro 6 $\Rightarrow 33\,\text{b} = 2.75\,\text{b/číslo}$
4. kódování opakování čísla $\rightarrow$ $\mathbf{0}$ pro žádné, $\mathbf{I0}$ pro jedno a $\mathbf{II}$ pro dvě $\Rightarrow$ $7 \times 3 + 11 = 32\,\text{b} = 2.\overline{6}\,\text{b/číslo}$
5. malé rozdíly mezi sousedními čísly $\rightsquigarrow$ predikce $\rightarrow d_1 = x_1 = 2$, $d_i = x_i - x_{i-1} = 0, 2, 2, 1, 0, 0, 3, 0, 1, 0, 3 \rightarrow 4\,\text{b} + 2\,\text{b/číslo} \Rightarrow 26\,\text{b} = 2.1\overline{6}\,\text{b/číslo}$

Komprese dat

Příklad

$x_1, x_2, \ldots, x_{12} = 2, 2, 4, 6, 7, 7, 7, 10, 10, 11, 11, 14$

1. číslo ve dvojkové soustavě $\Rightarrow 4\,\mathrm{b}/$číslo $= 48\,\mathrm{b}$
2. 7 různých čísel ve dvojkové soustavě $\Rightarrow 3\,\mathrm{b}/$číslo $= 36\,\mathrm{b}$
3. častější číslo kratší kód $\rightarrow 2 \times 2,\ 1 \times 4,\ 1 \times 6,\ 3 \times 7,\ 2 \times 10,\ 2 \times 11,\ 1 \times 14 \rightarrow$ **0I** pro 7, **III** pro 11, **II0** pro 10, **I0I** pro 2, **I00** pro 14, **000** pro 4 a **00I** pro $6 \Rightarrow 33\,\mathrm{b} = 2.75\,\mathrm{b}/$číslo
4. kódování opakování čísla $\rightarrow$ **0** pro žádné, **I0** pro jedno a **II** pro dvě $\Rightarrow$ $7 \times 3 + 11 = 32\,\mathrm{b} = 2.\overline{6}\,\mathrm{b}/$číslo
5. malé rozdíly mezi sousedními čísly $\rightsquigarrow$ predikce $\rightarrow d_1 = x_1 = 2$, $d_i = x_i - x_{i-1} = 0, 2, 2, 1, 0, 0, 3, 0, 1, 0, 3 \rightarrow 4\,\mathrm{b} + 2\,\mathrm{b}/$číslo $\Rightarrow 26\,\mathrm{b} = 2.1\overline{6}\,\mathrm{b}/$číslo
6. vztah mezi čísly $\rightsquigarrow$ predikce $\rightarrow \hat{x}_i = i + 1 \rightarrow$ $d_i = x_i - \hat{x}_i = 0, -1, 0, 1, 1, 0, -1, 1, 0, 0, -1, 1 \rightarrow$ **0** pro 0, **I0** pro -1 a **II** pro $1 \Rightarrow$ $19\,\mathrm{b} = 1.58\overline{3}\,\mathrm{b}/$číslo

Komprese dat

- využití („zneužití“) omezení reprodukční techniky a příjemce obsahu (člověka) pro vynechání nevyužitelných informací (obraz, video, zvuk)
- data ...znaky textu, vzorky obrazu (body) a videa (body v čase), zvuku (úrovně v čase), aj., digitální (digitalizovaná) forma, narůstající objem – např. obraz foto $10\,$Mpx $24\,$bpp $\sim 30\,$MB, video HDTV 1920×1080 $12\,$bpp, $25\,$fps $\sim 590\,$Mb/s, zvuk CD $44.1\,$kHz, $16\,$bps, stereo $\sim 1.3\,$Mb/s
- vývoj úložných a přenosových technologií nestačí, navíc (fyzikální) omezení
- umožnění tzv. multimediální revoluce – komprese textu, obrazu, videa, zvuku při uložení a přenosu

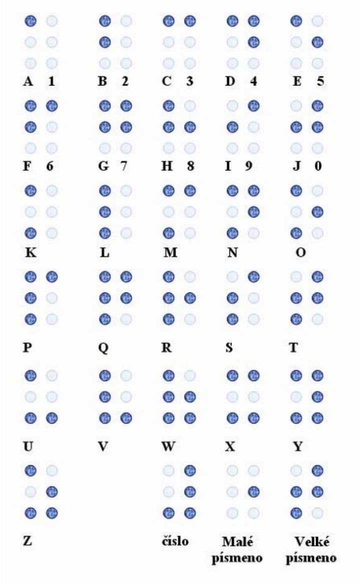
$\to$ všudypřítomná – počítače, spotřební elektronika, komunikační a distribuční sítě, ...

Příklady z minulosti

- morseovka: písmena (a číslice a interpunkce) kódována do posloupností teček a čárek, častější (e, t) kratšími pro zmenšení průměrné délky textu
- Braillovo písmo: do 2×3 matice teček kódována písmena (a číslice, interpunkce aj., Grade 1) a častá slova (a jejich zkratky, Grade 2)

Obrázek: Morseovka a Braillovo písmo

Komprese dat



Kompresní techniky a metody

- dva algoritmy: kompresní pro kompresi originálních dat na komprimovaná a dekompresní (rekonstrukční) pro dekompresi komprimovaných dat na dekomprimovaná (rekonstruovaná)
- standardy: ISO, ITU-T aj.

Bezeztrátové (lossless)

- dekomprimovaná data stejná jako data originální = žádná ztráta informace v datech
- např. pro text, programové (binární) soubory, citlivé záznamy (bankovní, zdravotní), nereprodukovatelná data (snímky v čase) aj.
- statistické: Huffmanovo a aritmetické kódování
- kontextové: PPM
- slovníkové: LZ*
- jiné: BWT, ACB, obrazové (JPEG-LS, JBIG)

Kompresní techniky a metody

Ztrátové(lossy)

= při kompresi vynechání nějaké informace v originálních datech → dekomprimovaná data (obecně) odlišná od originálních dat = ztráta informace z originálních dat – zkreslení dat

- vyšší míra komprese než u bezeztrátových za cenu vyšší míry zkreslení dat
- např. pro obraz, video, zvuk (hudba, řeč) – zkreslení dat vede k artefaktům při reprodukci obsahu
- vzorkování a kvantizace: skalární a vektorová
- diferenční kódování: DPCM, delta modulace
- transformační a podpásmové kódování: Fourierova, Z a kosinová transformace, wavelety
- aplikace: obraz – JPEG, fraktály, video – H.*, MPEG, zvuk – MDCT, G.*, MPEG, LPC, CELP

Míry kompresních algoritmů

- asymptotická časová a paměťová složitost algoritmů komprese a dekomprese
- experimentální časová a paměťová náročnost algoritmů – jejich implementací na referenčních datech
- míry komprese
  - kompresní poměr (compression ratio) = poměr velikosti originálních a komprimovaných dat, také jako procento velikosti komprimovaných dat z velikosti originálních dat
  - compression rate = průměrná velikost komprimovaných dat na vzorek originálních dat, např. pixel u obrazu – bitů/pixel, sekunda u videa a zvuku – bitů/s
  - na referenčních datech
- míry zkreslení (distortion) – rozdíl mezi originálními a dekomprimovanými daty, více způsobů měření „přesnosti (fidelity)“ a „kvality“ obsahu, viz dále, na referenčních datech

Modely dat

Fyzický

= popis zdroje dat – např. měřených, popis měřidla

- u ztrátové komprese zvuku (řeči) – popis syntezátoru a syntéza dat
- obecně příliš složitý nebo nemožný

Pravděpodobnostní model

= empiricky zjištěný statistický popis zdroje dat

- pro statistické a kontextové bezeztrátové kompresní metody
- ignorantní: výskyt každé hodnoty na výstupu zdroje dat je nezávislý na výskytu ostatních hodnot a je se stejnou pravděpodobností – nejjednodušší
- dostupná pravděpodobnost výskytu nezávisle se vyskytujících hodnot
- pravděpodobnost:
  - frekvence/četnost výskytu výsledku experimentu (hodnot na výstupu zdroje dat) – n opakování experimentu, n_i výskytů výsledku $\omega_i \in \Omega, i \in \{1, 2, \ldots, N\}$ (Ω ... prostor výsledků (sample space)) $\rightarrow$ frekvence/četnost výskytu výsledku ω_i: $f(\omega_i) = f_i = \frac{n_i}{n} =$ přibližná hodnota/odhad pro pravděpodobnost výskytu výsledku ω_i: $P(\omega_i) = p_i = \lim_{n\to\infty} \frac{n_i}{n}$, událost (event) $A \subseteq \Omega$, výskyt události = výskyt kteréhokoliv výsledku události, $f(A) \geq 0 \Rightarrow P(A) \geq 0$ (1), $P(\Omega) = 1$ (2), $B \subseteq \Omega, A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$ (3), $\sum_i P(\omega_i) = 1$

Modely dat

Pravděpodobnostní model

- pravděpodobnost:
  - míra víry (belief) v událost – a priori pravděpodobnost $P(A)$ události A před výskytem události (získání informace) B, a posteriori pravděpodobnost $P(A|B)$ po/za předpokladu, sdružená (joint) pravděpodobnost $P(A, B)$ výskytu obou událostí A, B, Bayesovo pravidlo $P(A|B) = \frac{P(A,B)}{P(B)}$, (statisticky) nezávislé události $\ldots P(A, B) = P(A)P(B)$, tj. při $P(A|B) = P(A)$, pro případy, kdy experiment není možné provést
  - míra („velikost“) události (jako množiny) – jako jiné míry (1) a (3), normalizace (2) = axiomy, z nich např. $P(\bar{A}) = 1 - P(A)$, $P(\emptyset) = 0$, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ aj., pro nediskrétní prostor výsledků

Modely dat

Pravděpodobnostní model

- náhodná proměnná/veličina: měřitelné $X : \Omega \mapsto \mathbb{R}$ ($\mathbb{R}$ . . . obor reálných čísel), realizace $X(\omega) = x$, např. $P(X(\omega) \leq x) = P(X \leq x)$, diskrétní a spojitá
- rozdělení pravděpodobnosti: distribuční funkce/kumulovaná pravděpodobnost (cumulative distribution function) $F_X(x) = P(X \leq x)$, $x_1 \geq x_2 \Rightarrow F_X(x_1) \geq F_X(x_2)$, $P(X = x) = F_X(x) - F_X(x^-)$ pro $F_X(x^-) = P(X < x)$, rozdělení/distribuce/hustota pravděpodobnosti (probability distribution/density function) $f_X(x)$ . . . diference/derivace $F_X(x)$ pro diskrétní/spojitou X, pro diskrétní typicky $f_X(x) = P(X = x)$, např. binomické, Poissonovo, uniformní, normální (Gaussovo), aj.
- sdružená (joint) distribuční funkce $F_{X_1X_2\ldots X_n}(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$, sdružené rozložení pravděpodobnosti $f_{X_1X_2\ldots X_n}(x_1, x_2, \ldots, x_n)$, marginální pro jednotlivé X_i, X_1, X_2 nezávislé, jestliže $F_{X_1X_2}(x_1, x_2) = F_{X_1}(x_1)F_{X_2}(x_2)$ (a tedy i $f_{X_1X_2}(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2)$)

Modely dat

Pravděpodobnostní model

- střední hodnota (expected value) náhodné proměnné X: $E[X] = \sum_i x_i P(X = x_i)$ pro diskrétní X, $E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$ pro spojitou X, statistický průměr (mean, statistical average) $\mu_X = E[X]$, rozptyl (variance) $\sigma_X^2 = E[(X - \mu_X)^2] = E[X^2] - \mu_X^2$, standardní odchylka (standard deviation) $\sigma_X = \sqrt{\sigma_X^2}$, X_1, X_2 nekorelované, jestliže $E[(X_1 - \mu_1)(X_2 - \mu_2)] = 0$
- náhodný/stochastický proces: měřitelné $X : \Omega \mapsto \mathcal{F}, \mathcal{F} : \mathbb{R} \mapsto \mathbb{R}$, realizace $X(\omega) = x(t), -\infty < t < \infty$ funkce času, ensemble $X(t) = \{x_\omega(t)\}$, střední hodnota ensemble, vzorek (sample) $X(t_0)$ ensemble = náhodná proměnná
- problém nulové pravděpodobnosti/frekvence (zero probability/frequency problem): kompresní metody předpokládají u modelu všechny uvažované pravděpodobnosti/frekvence nenulové $\rightarrow$ místo nulových nastavení velice malých

Modely dat

Markovův model (Andrei A. Markov)

- výskyt hodnoty x_j na výstupu zdroje dat je závislý na výskytu (některých, ne nutně bezprostředně) předchozích hodnot $x_i, i < j$
- vychází z pravděpodobnostního modelu
- v bezeztrátové kompresi Markovův řetěz s diskrétním časem: posloupnost hodnot x_j (náhodné proměnné X_j) následuje Markovův model/proces k-tého řádu, jestliže $P(x_j|x_{i_1}, x_{i_2}, \ldots, x_{i_k}) = P(x_j|x_{j-1}, x_{j-2}, \ldots), i_1, i_2, \ldots, i_k < j$ (znalost některých předchozích k hodnot je stejná jako znalost všech předchozích hodnot), posloupnosti s_j hodnot $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$ = stavy modelu/procesu/řetězu, $P(x_j|s_j)$ = pravděpodobnosti přechodu mezi stavy
- nejběžnější model 1. řádu: $P(x_j|x_i = P(x_j|x_{j-1}, x_{j-2}, \ldots), i < j$
- s_j, $P(s_j)$, $P(x_j|s_j)$ …stavový diagram
- různé modely podle formy závislosti, se zvyšujícím se k vyšší míra komprese než s nezávislými výskyty hodnot
- v kompresi textu Markovův model k-tého řádu = model konečného kontextu (finite context model) – kontext = stav modelu

Modely dat

Markovův model

Příklad

$x_1x_2 \ldots x_{10} = aababbabaa$

stavy modelu 1. řádu = posloupnosti (bezprostředně) předchozích symbolů délky 1 pro všechny symboly: a, b
$P(a) = \frac{6}{10}$, $P(b) = \frac{4}{10}$,
$P(a|a) = \frac{2}{5}$, $P(b|a) = \frac{3}{5}$, $P(a|b) = \frac{3}{4}$, $P(b|b) = \frac{1}{4}$

stavy modelu 2. řádu = posloupnosti (bezprostředně) předchozích symbolů délky 2 pro všechny symboly: aa, ab, ba, bb
$P(aa) = \frac{2}{9}$, $P(ab) = \frac{3}{9}$, $P(ba) = \frac{3}{9}$, $P(bb) = \frac{1}{9}$,
$P(a|aa) \to 0$, $P(b|aa) \to 1$, $P(a|ab) = \frac{2}{3}$, $P(b|ab) = \frac{1}{3}$, $P(a|ba) = \frac{1}{3}$, $P(b|ba) = \frac{2}{3}$,
$P(a|bb) \to 1$, $P(b|bb) \to 0$

Modely dat

Typy

- statický – neměnný pro různá originální data a během kódování, známý algoritmu dekomprese
- semi-adaptivní – vytvořený pro originální data (1. průchod daty při kompresi), během kódování neměnný (2. průchod) a předaný algoritmu dekomprese (např. s komprimovanými daty)
- adaptivní – dynamicky vytvářený/modifikovaný podle doposud zakódovaných originálních a dekomprimovaných dat

Klasická Shannonova

- rámec pro bezeztrátové kompresní metody, vychází z pravděpodobnostního modelu dat
- Claude E. Shannon: A Mathematical Theory of Communication. *Bell System Technical Journal* **27**, pp. 379–423, 623–656, 1948.
- „míra průměrné informace (asociované s) experimentu(-em)" – požadavky, pro nezávislé jevy $A_i, i = 1, \ldots, m, \bigcup A_i = \Omega$:
  1. spojitá funkce $H(p_i), p_i = P(A_i)$
  2. monotónně rostoucí vzhledem k počtu m stejně pravděpodobných jevů A_i $(p_i = \frac{1}{m})$
  3. stejná při rozdělení experimentu na k podexperimentů (s disjunktními podmnožinami množiny jevů A_i), výsledek experimentu = podmnožina s jevem, výsledek podexperimentu = jev v podmnožině:
     $H(p_i) = H(q_1, q_2, \ldots, q_k) + q_1 H(\frac{p_1}{q_1}, \frac{p_2}{q_1}, \ldots, \frac{p_{j_1}}{q_1}) + q_2 H(\frac{p_{j_1+1}}{q_2}, \frac{p_{j_1+2}}{q_2}, \ldots, \frac{p_{j_2}}{q_2}) + \ldots + q_k H(\frac{p_{j_{k-1}+1}}{q_k}, \frac{p_{j_{k-1}+2}}{q_k}, \ldots, \frac{p_{j_k}}{q_k}), q_1 = \sum_{i=1}^{j_1} p_i, q_2 = \sum_{i=j_1+1}^{j_2} p_i, \ldots, q_k = \sum_{i=j_{k-1}+1}^{j_k=m} p_i$
- jediné možné řešení požadavků (Shannon): $H(p_i) = -K \sum_i p_i \log p_i$, K kladná konstanta

Teorie informace

Klasická Shannonova

- informace (self-information) (asociovaná s výskytem) jevu A: $i(A) = \log_b \frac{1}{P(A)} = -\log_b P(A)$ – $\log(1) = 0$ a roste s klesající $P(A) \neq 0$, pro nezávislé A, B $i(AB) = i(A) + i(B)$
- jednotka i: bit (shannon) pro $b = 2$, nat pro $b = e$, hartley pro $b = 10$
- entropie (asociovaná s) experimentu(-em): průměr $H(A_i) = \sum_i P(A_i) i(A_i) = -\sum_i P(A_i) \log P(A_i)$ informací nezávislých jevů $A_i, \bigcup A_i = \Omega$ (jako náhodných proměnných), $0 \log 0 := 0$
- Shannon: experiment = zdroj Z posloupnosti $X_1, X_2, \ldots, X_n$ symbolů z množiny $\{a_1, a_2, \ldots, a_m\}$ jako náhodných proměnných $X_j(a_i) = i$, pak entropie zdroje = průměrný počet binárních symbolů (bitů) potřebných pro zakódování každého symbolu posloupnosti = $H(Z) = \lim_{n\to\infty} \frac{1}{n} G_n, G_n = -\sum_{i_1=1}^{m} \sum_{i_2=1}^{m} \ldots \sum_{i_n=1}^{m} P(X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n) \log P(X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n)$ – limit pro bezeztrátovou kompresi

Klasická Shannonova

- jestliže je výskyt každého symbolu X_j (jako náhodné proměnné) nezávislý a stejně pravděpodobnostně rozložený, pak $X_j = X$, $G_n = -n \sum_{i=1}^{m} P(X = i) \log P(X = i)$ a ${}^{1}H(Z) = -\sum_{i=1}^{m} P(a_i) \log P(a_i)$ = entropie 1. řádu
- podmíněná entropie (pro náhodné proměnné) X_1 v závislosti na X_2: průměr $H(Z) = H(X_1|X_2) = \sum_{i_2=1}^{m} P(a_{i_2}) H(X_1|X_2 = i_2) = -\sum_{i_2=1}^{m} P(a_{i_2}) \sum_{i_1=1}^{m} P(a_{i_1}|a_{i_2}) \log P(a_{i_1}|a_{i_2})$ podmíněných entropií X_1 v závislosti na $X_2 = i_2$
- entropie Markovova modelu 1. řádu se stavy $S = \{s_j\}$: $H(X|S)$
- entropie (obecně) nezjistitelná $\Rightarrow$ odhad závislý na modelu struktury dat!

Teorie informace

Klasická Shannonova

Příklad

$x_1, x_2, \ldots, x_{12} = 2, 2, 4, 6, 7, 7, 7, 10, 10, 11, 11, 14$

- $\{a_1, a_2, \ldots, a_7\} = \{2, 4, 6, 7, 10, 11, 14\}$
  $P(a_i) = p_i \approx f(a_i) = f_i$: $f_1 = f_5 = f_6 = \frac{2}{12}, f_2 = f_3 = f_7 = \frac{1}{12}, f_4 = \frac{3}{12}$
  $^1H(a_i) = -\sum_{i=1}^{7} p_i \log_2 p_i \doteq 2.689$ bitů/číslo

Teorie informace

Klasická Shannonova

Příklad

$x_1, x_2, \ldots, x_{12} = 2, 2, 4, 6, 7, 7, 7, 10, 10, 11, 11, 14$

- $\{a_1, a_2, \ldots, a_7\} = \{2, 4, 6, 7, 10, 11, 14\}$
  $P(a_i) = p_i \approx f(a_i) = f_i$: $f_1 = f_5 = f_6 = \frac{2}{12}, f_2 = f_3 = f_7 = \frac{1}{12}, f_4 = \frac{3}{12}$
  ${}^1H(a_i) = -\sum_{i=1}^{7} p_i \log_2 p_i \doteq 2.689$ bitů/číslo
- Sousední čísla nejsou nezávislá $\rightarrow$ odstranění závislosti (korelace):
  $d_2, d_3, \ldots, d_{12} = 0, 2, 2, 1, 0, 0, 3, 0, 1, 0, 3,\ \{a_1, a_2, a_3, a_4\} = \{0, 1, 2, 3\}$
  $f_1 = \frac{5}{11}, f_2 = f_3 = f_4 = \frac{2}{11}$
  ${}^1H(a_i) = -\sum_{i=1}^{4} p_i \log_2 p_i \doteq 1.859$ bitů/číslo

Teorie informace

Klasická Shannonova

Příklad

$x_1, x_2, \dots, x_{12} = 2, 2, 4, 6, 7, 7, 7, 10, 10, 11, 11, 14$

- $\{a_1, a_2, \dots, a_7\} = \{2, 4, 6, 7, 10, 11, 14\}$
  $P(a_i) = p_i \approx f(a_i) = f_i$: $f_1 = f_5 = f_6 = \frac{2}{12}, f_2 = f_3 = f_7 = \frac{1}{12}, f_4 = \frac{3}{12}$
  ${}^1H(a_i) = -\sum_{i=1}^{7} p_i \log_2 p_i \doteq 2.689$ bitů/číslo
- Sousední čísla nejsou nezávislá $\rightarrow$ odstranění závislosti (korelace):
  $d_2, d_3, \dots, d_{12} = 0, 2, 2, 1, 0, 0, 3, 0, 1, 0, 3,\ \{a_1, a_2, a_3, a_4\} = \{0, 1, 2, 3\}$
  $f_1 = \frac{5}{11}, f_2 = f_3 = f_4 = \frac{2}{11}$
  ${}^1H(a_i) = -\sum_{i=1}^{4} p_i \log_2 p_i \doteq 1.859$ bitů/číslo
- Všechna čísla jsou mezi sebou závislá $\rightarrow$ odstranění závislosti (korelace):
  $d_1, d_2, \dots, d_{12} = 0, -1, 0, 1, 1, 0, -1, 1, 0, 0, -1, 1,\ \{a_1, a_2, a_3\} = \{0, -1, 1\}$
  $f_1 = \frac{5}{12}, f_2 = \frac{3}{12}, f_3 = \frac{4}{12}$
  ${}^1H(a_i) = -\sum_{i=1}^{3} p_i \log_2 p_i \doteq 1.555$ bitů/číslo

Teorie informace

Klasická Shannonova

Příklad

$x_1 x_2 \dots x_{10} = aababbabaa$

výskyt a a b nezávislý: $P(a) = \frac{6}{10}$, $P(b) = \frac{4}{10}$
$H = -P(a) \log_2 P(a) - P(b) \log_2 P(b) \doteq 0.971$ b/symbol

Teorie informace

Klasická Shannonova

Příklad

$x_1 x_2 \ldots x_{10} = aababbabaa$

výskyt a a b nezávislý: $P(a) = \frac{6}{10}$, $P(b) = \frac{4}{10}$
$H = -P(a) \log_2 P(a) - P(b) \log_2 P(b) \doteq 0.971$ b/symbol

Markovův model 1. řádu: $P(a) = \frac{5}{9}$, $P(b) = \frac{4}{9}$,
$P(a|a) = \frac{2}{5}$, $P(b|a) = \frac{3}{5}$, $P(a|b) = \frac{3}{4}$, $P(b|b) = \frac{1}{4}$
$H = P(a)H(X|a) + P(b)H(X|b) = P(a)(-P(a|a) \log_2 P(a|a) - P(b|a) \log_2 P(b|a)) + P(b)(-P(a|b) \log_2 P(a|b) - P(b|b) \log_2 P(b|b)) \doteq 0.9$ b/symbol

Klasická Shannonova

Příklad

$x_1x_2\dots x_{10} = aababbabaa$

výskyt a a b nezávislý: $P(a) = \frac{6}{10}$, $P(b) = \frac{4}{10}$
$H = -P(a)\log_2 P(a) - P(b)\log_2 P(b) \doteq 0.971$ b/symbol

Markovův model 1. řádu: $P(a) = \frac{5}{9}$, $P(b) = \frac{4}{9}$,
$P(a|a) = \frac{2}{5}$, $P(b|a) = \frac{3}{5}$, $P(a|b) = \frac{3}{4}$, $P(b|b) = \frac{1}{4}$
$H = P(a)H(X|a) + P(b)H(X|b) = P(a)(-P(a|a)\log_2 P(a|a) - P(b|a)\log_2 P(b|a)) +$
$P(b)(-P(a|b)\log_2 P(a|b) - P(b|b)\log_2 P(b|b)) \doteq 0.9$ b/symbol

Markovův model 2. řádu: $P(aa) = \frac{1}{8}$, $P(ab) = \frac{3}{8}$, $P(ba) = \frac{3}{8}$, $P(bb) = \frac{1}{8}$,
$P(a|aa) \to 0$, $P(b|aa) \to 1$, $P(a|ab) = \frac{2}{3}$, $P(b|ab) = \frac{1}{3}$, $P(a|ba) = \frac{1}{3}$, $P(b|ba) = \frac{2}{3}$,
$P(a|bb) \to 1$, $P(b|bb) \to 0$
$H = \sum_{x_1x_2=aa}^{bb} P(x_1x_2)H(X|x_1x_2) =$
$\sum_{x_1x_2=aa}^{bb} P(x_1x_2) - \sum_{y=a,b} P(y|x_1x_2)\log_2 P(y|x_1x_2) \doteq 0.689$ b/symbol

Teorie informace

Klasická Shannonova

Uvažováním závislosti mezi symboly dat (posloupnosti) v modelu struktury dat nesnižujeme „entropii dat“. Entropie je vlastnost (hypotetického) zdroje dat a je stejná pro všechna data ze zdroje. Snižujeme odhad této entropie uvažováním delších n-tic symbolů dat a závilostí mezi nimi v modelu (až do $n \to \infty$)!

Algoritmická

- Kolmogorov/descriptive complexity / algoritmická entropie dat (Andrey N. Kolmogorov): délka nejmenšího/nejkratšího počítačového programu (včetně vstupu, v jakémkoliv programovacím jazyce), jehož jsou data výstupem – způsob modelování struktury dat
- není znám žádný systematický způsob výpočtu nebo libovolně blízkého odhadu
- Minimum Description Length (MDL) princip (J. Risannen): $MDL(x) = min_j(D_{M_j} + R_{M_j}(x))$, D_{M_j} délka popisu možného modelu M_j struktury dat x, $R_{M_j}(x)$ délka reprezentace x podle modelu M_j
- např. M_j polynomy j-tého řádu: pro vyšší j kratší $R_{M_j}(x)$ (přesnější model), ale delší D_{M_j} (složitější model), a naopak $\Rightarrow$ kompromis

Kódování

- abeceda $A = \{a_1, a_2, \dots, a_n\}$, a_i = symboly
- = (kód) ze zdrojové abecedy A do kódové abecedy B: injektivní $C : A \mapsto B^+$, $B^+ =$ množina konečných neprázdných posloupností (= slov) symbolů z B – často $B = \{\mathbf{0}, \mathbf{I}\} \rightarrow$ binární kódování (kód)
- $C(a_i) \in B^+$ … kódové slovo (kód) pro symbol a_i, $C(A) = \{C(a_i), a_i \in A\} \subseteq B^+$ … kód (pro zdrojovou abecedu A), $l(a_i)$ … délka $C(a_i)$, pro $B = \{\mathbf{0}, \mathbf{I}\}$ v bitech
- dekódování: $D : C(A) \mapsto A$
- např. $\{\mathbf{0}, \mathbf{I}, \mathbf{00}, \mathbf{II}\}$, ne $\{\mathbf{0}, \mathbf{0}, \mathbf{I}, \mathbf{II}\}$
- blokový kód (kód pevné délky, fixed-length code) = všechna kódová slova (pro všechny symboly) mají stejnou délku, např. ASCII

Kódování

Jednoznačně dekódovatelný kód

= každá (neprázdná) posloupnost symbolů z kódové abecedy je zřetězením nejvýše jedné posloupnosti kódových slov

= $C^+ : A^+ \mapsto B^+, C^+(a_{i_1} a_{i_2} \dots a_{i_j}) = C(a_{i_1})C(a_{i_2}) \dots C(a_{i_j})$ injektivní

- dekódování: $D^+ : C^+(A^+) \mapsto A^+$
- např. každý blokový, $\{\mathbf{0}, \mathbf{0I}, \mathbf{0II}, \mathbf{III}\}$, ne $\{\mathbf{0}, \mathbf{0I}, \mathbf{I0}, \mathbf{II}\}$
- test: $S \leftarrow C(A)$ a opakuj $S \leftarrow S \cup \{s \in B^+; ps \in S \wedge p \in S\}$ dokud některé $s \in C(A)$ nebo S zůstane stejná, při $s \in C(A)$ kód $C(A)$ není jednoznačně dekódovatelný

Prefixový (prefix, instantaneous) kód

= žádné kódové slovo není prefixem jiného kódového slova

- např. každý blokový, $\{\mathbf{0}, \mathbf{I0}, \mathbf{II0}, \mathbf{III}\}$
- jednoznačně dekódovatelný

Kódování

Věta (Kraftova)

Prefixový kód s k kódovými slovy délek $l_1, l_2, \ldots, l_k$ nad kódovou abecedou velikosti m existuje právě když

$$\sum_{i=1}^{k} m^{-l_i} \leq 1 \quad \ldots \quad \textit{Kraftova nerovnost.}$$

□

Kódování

Věta (Kraftova)

Prefixový kód s k kódovými slovy délek $l_1, l_2, \ldots, l_k$ nad kódovou abecedou velikosti m existuje právě když

$$\sum_{i=1}^{k} m^{-l_i} \leq 1 \quad \ldots \quad \textit{Kraftova nerovnost.}$$

□

Věta (McMillanova)

Jednoznačně dekódovatelný kód s k kódovými slovy délek $l_1, l_2, \ldots, l_k$ nad kódovou abecedou velikosti m existuje právě když

$$\sum_{i=1}^{k} m^{-l_i} \leq 1 \quad (\ldots \quad \textit{Kraft-McMillanova nerovnost}).$$

□

Kódování

Optimální kód

- pro pravděpodobnostní model dat (výskytu symbolů), prefixový kód
- průměrná délka kódu (na symbol, code rate): průměr $\bar{l}(C(A)) = \sum_{i=1}^{n} P(a_i)l(a_i)$ délek $l(a_i)$ pro všechny $a_i \in A$, $P(a_i) \neq 0 =$ pravděpodobnost výskytu symbolu a_i

$=$ s minimální $\bar{l}(C(A))$ (v rámci třídy kódů, např. prefixové)

Kódování

Optimální kód

- pro pravděpodobnostní model dat (výskytu symbolů), prefixový kód
- průměrná délka kódu (na symbol, code rate): průměr $\bar{l}(C(A)) = \sum_{i=1}^{n} P(a_i)l(a_i)$ délek $l(a_i)$ pro všechny $a_i \in A$, $P(a_i) \neq 0$ = pravděpodobnost výskytu symbolu a_i

$=$ s minimální $\bar{l}(C(A))$ (v rámci třídy kódů, např. prefixové)

Věta (Shannon noisless coding theorem)

Pro optimální jednoznačně dekódovatelný kód ze zdrojové abecedy A do kódové abecedy B platí

$$\frac{H(A)}{\log_b m} \leq \bar{l}(C(A)) < \frac{H(A)}{\log_b m} + 1$$

kde $H(A)$ je entropie zdroje symbolů z A, b je stejné jako v H a m je velikost B.
$\bar{l}(C(A)) = \frac{H(A)}{\log_b m}$ *právě když* $P(a_i) = m^{-l(a_i)}$ *pro všechny* $a_i \in A$. □

- redundance kódu: $\bar{l}(C(A)) - \frac{H(A)}{\log_b m}$, také v % z $\frac{H(A)}{\log_b m}$, $\bar{l}(C(A)) = \frac{H(A)}{\log_b m}$ … absolutně optimální kód

Kódování

Optimální kód

- změnou zdrojové abecedy na k-tice (nezávislých) symbolů z původní abecedy A (rozšíření zdrojové abecedy, source extension) se lze $\bar{l}(C(A)) = \frac{H(A)}{\log_b m}$ libovolně přiblížit (až do $k \to \infty$):

$$\frac{H(A^k)}{\log_b m} \leq \bar{l}(C(A^k)) < \frac{H(A^k)}{\log_b m} + 1$$
$$\frac{kH(A)}{\log_b m} \leq k\bar{l}(C(A)) < \frac{kH(A)}{\log_b m} + 1$$
$$\frac{H(A)}{\log_b m} \leq \bar{l}(C(A)) < \frac{H(A)}{\log_b m} + \frac{1}{k}$$

Kódování

Optimální kód

- změnou zdrojové abecedy na k-tice (nezávislých) symbolů z původní abecedy A (rozšíření zdrojové abecedy, source extension) se lze $\bar{l}(C(A)) = \frac{H(A)}{\log_b m}$ libovolně přiblížit (až do $k \to \infty$):

$$\frac{H(A^k)}{\log_b m} \leq \bar{l}(C(A^k)) < \frac{H(A^k)}{\log_b m} + 1$$
$$\frac{kH(A)}{\log_b m} \leq k\bar{l}(C(A)) < \frac{kH(A)}{\log_b m} + 1$$
$$\frac{H(A)}{\log_b m} \leq \bar{l}(C(A)) < \frac{H(A)}{\log_b m} + \frac{1}{k}$$

- abeceda A^k ale může mít velikost až n^k (n je velikost A)!

Kódování

Optimální kód

Příklad

$A = \{a_1, a_2, a_3\}$
$P(a_i) = 0.8, P(a_2) = 0.02, P(A_3) = 0.18$
$H(A) = -\sum_{i=1}^{3} P(a_i) \log_2 P(a_i) \doteq 0.816$ bitů/symbol
$C(A) = \{\langle a_1, \mathbf{0}\rangle, \langle a_2, \mathbf{II}\rangle, \langle a_3, \mathbf{I0}\rangle\}$
$\bar{l}(C(A)) = \sum_{i=1}^{3} P(a_i) l(a_i) = 1.2$ b/symbol
$\bar{l}(C(A)) - \frac{H(A)}{\log_2 2} \doteq 0.384$ b/symbol $\doteq 47\,\%$

Kódování

Optimální kód

Příklad

$A = \{a_1, a_2, a_3\}$
$P(a_i) = 0.8, P(a_2) = 0.02, P(A_3) = 0.18$
$H(A) = -\sum_{i=1}^{3} P(a_i) \log_2 P(a_i) \doteq 0.816$ bitů/symbol
$C(A) = \{\langle a_1, \mathbf{0}\rangle, \langle a_2, \mathbf{II}\rangle, \langle a_3, \mathbf{I0}\rangle\}$
$\bar{l}(C(A)) = \sum_{i=1}^{3} P(a_i) l(a_i) = 1.2$ b/symbol
$\bar{l}(C(A)) - \frac{H(A)}{\log_2 2} \doteq 0.384$ b/symbol $\doteq 47\,\%$

$A^2 = \{a_1a_1, a_1a_2, a_1a_3, a_2a_1, a_2a_2, a_2a_3, a_3a_1, a_3a_2, a_3a_3\}$
$P(a_1a_1) = 0.64, P(a_1a_2) = P(a_2a_1) = 0.016, P(a_1a_3) = P(a_3a_1) = 0.144, P(a_2a_2) = 0.0004, P(a_2a_3) = P(a_3a_2) = 0.0036, P(a_3a_3) = 0.0324$
$C(A^2) = \{\langle a_1a_1, \mathbf{0}\rangle, \langle a_1a_2, \mathbf{I0I0I}\rangle, \langle a_1a_3, \mathbf{II}\rangle, \langle a_2a_1, \mathbf{I0I000}\rangle, \langle a_2a_2, \mathbf{I0I00I0I}\rangle, \langle a_2a_3, \mathbf{I0I00II}\rangle, \langle a_3a_1, \mathbf{I00}\rangle, \langle a_3a_2, \mathbf{I0I00I00}\rangle, \langle a_3a_3, \mathbf{I0II}\rangle\}$
$\bar{l}(C(A)) = \frac{\bar{l}(C(A^2))}{2} = \frac{\sum_{i=1,j=1}^{3,3} P(a_ia_j) l(a_ia_j)}{2} \doteq \frac{1.723}{2} \doteq 0.862$ b/symbol
$\bar{l}(C(A)) - \frac{H(A)}{\log_2 2} \doteq 0.046$ b/symbol $\doteq 5.6\,\%$

Kódování

Optimální kód

Věta

Pro optimální prefixový kód ze zdrojové abecedy A do kódové abecedy B platí

1. *Symboly z A s větší pravděpodobností výskytu mají kratší kódová slova.*
2. $m' \in \{2, 3, \ldots, m\}, m' \equiv n \pmod{(m-1)}$ *symbolů z A s nejmenší pravděpodobností výskytu, kde $n \geq 2$ je velikost A a $m \geq 2$ je velikost B, má stejně (maximálně) dlouhá kódová slova a ta se liší pouze v jednom symbolu.* $\square$

Proč m' a ne m? Odpověď u Huffmanova kódování (viz dále).

Základní techniky a kódování čísel

Run-length encoding (RLE)

= kódování posloupností stejných zdrojových symbolů (runs) kódy příznaku kódování opakování, délky posloupnosti a jednoho symbolu místo samotných symbolů

- podle délky kódů příznaku a délky až pro posloupnosti delší než určitý počet k symbolů, např. 3
- kód příznaku může být zaměnitelný s kódem symbolu $\rightarrow$ kódování s kódem délky zmenšené o k za kódy určitého počtu k symbolů
- aplikace: text, obraz (BMP)

Diferenční kódování

= kódování (malého) rozdílu symbolu/čísla od předchozího (nebo predikce z několika předchozích) kódy příznaku kódování rozdílu a rozdílu místo samotného symbolu/čísla, s výjimkou prvního

Run-length encoding (RLE)

Input: číslo k
$r \leftarrow 0$;
while načti ze vstupu symbol a **do**
    if $r = 0$ **then**
        $x \leftarrow a$;
        $r \leftarrow 1$;
    else
        if $a = x$ **then**
            $r \leftarrow r + 1$;
        else
            if $r \leq k$ **then**
                zapiš na výstup r kódů symbolu x;
            else
                zapiš na výstup kódy příznaku, čísla r a symbolu x / k kódů symbolu x a kód čísla $r - k$;
            $x \leftarrow a$;
            $r \leftarrow 1$;

PRIKLAD: $k = 3$, vstup $bbbbaaaarrrbbaaaaara$, kod priznaku x, kod opakovani cislo, kod symbolu symbol, obe varianty

Move-to-front (MTF) kódování

- kódování často se opakujících symbolů malými čísly (speciálně posloupností stejných symbolů posloupností čísel 0)
- lokálně adaptivní = adaptace podle lokálních četností výskytu symbolů

Uses: zdrojová abeceda $A = \{a_1, \ldots, a_n\}$, (volitelně) pravděpodobnosti $\{p_1, \ldots, p_n\}$ výskytu a_i

(volitelně) setřiď a_i a p_i tak, že $p_i \geq p_j$ pro $i < j$;

while načti ze vstupu symbol $a \in A$ **do**

  zapiš na výstup číslo $i - 1$, kde $a_i = a$;

  if $i > 1$ **then**

    $x \leftarrow a_i$;

    $a_j \leftarrow a_{j-1}$ pro $j = 2, 3, \ldots, i$;

    $a_1 \leftarrow x$;

PRIKLAD: $A = \{a, b, r\}$, $p(a) = \frac{10}{20}, p(b) = \frac{6}{20}, p(r) = \frac{4}{20}$, vstup $bbbbaaaarrrbbaaaaara$, se setrizenim i bez

Kódování čísel

- přirozených čísel – celá lze bijektivně zobrazit na přirozená (např. $-2i$ pro $i < 0$ a $2i + 1$ pro $i \geq 0$)
- předpoklad nižší pravděpodobnosti výskytu u větších čísel
- binární kódy s proměnnou délkou (variable-length codes, fixed-to-variable codes) $=$ proměnná délka kódových slov pro zdrojová slova pevné délky (symboly nebo jejich k-tice), nízká průměrná délka kódu vs. náročnější manipulace s kódem (v porovnání s blokovým kódem, s využitím bufferu)

Unární kód

- kódování přirozených čísel

$=$ pro $i \geq 0$: zřetězení i $\mathbf{I}$ a $\mathbf{0}$ (nebo opačně), např. $\mathbf{IIIII0}$ pro 5

- prefixový, délka $i + 1$, optimální při $P(i) = \frac{1}{2^i}$

Další

- start-step-stop (obecné unární) kódy, start/stop kód, Levensteinův kód, Stoutovy kódy, Yamamotovy kódy, taboo kódy, Goldbachovy kódy, aditivní kódy aj.

Eliasovy kódy

- kódování přirozených čísel, P. Elias
- **Alpha** $= \alpha(i)$ pro $i \geq 0$: unární kód i, s $\mathbf{I}$ na konci
- **Beta** $= \beta(i)$ pro $i \geq 0$: reprezentace i ve dvojkové soustavě ($=$ binární reprezentace) – neprefixový
- další – pro $i \geq 1$: zřetězení kódu $l(\beta(i)) = \lfloor \log_2 i \rfloor + 1$ a $\beta(i)$, prefixové
- pro každé $i \geq 1$: $i = 2^{l(\beta(i))-1} + k$, $0 \leq k < 2^{l(\beta(i))-1}$

Gamma

= $\gamma(i)$ pro $i \geq 1$: zřetězení $l(\beta(i)) - 1$ $\mathbf{0}$ a $\beta(i)$ nebo $\alpha(l(\beta(i)) - 1)$ a $\beta(k)$, např. $\mathbf{00I0I}$ pro 5

- délka $2\lfloor \log_2 i \rfloor + 1$, optimální při $P(i) = \frac{1}{2i^2}$

Delta

= $\delta(i)$ pro $i \geq 1$: zřetězení $l(\beta(l(\beta(i)))) - 1$ $\mathbf{0}$, $\beta(l(\beta(i)))$ a $\beta(i)$ bez první $\mathbf{I}$ nebo $\gamma(l(\beta(i)))$ a $\beta(k)$, např. $\mathbf{0II0I}$ pro 5

- délka $2\lfloor \log_2 \log_2 2i \rfloor + \lfloor \log_2 i \rfloor + 1$, optimální při $P(i) = \frac{1}{2i(\log_2 2i)^2}$

Eliasovy kódy

Omega (rekurzivní)

= $\omega(i)$ pro $i = 1$: $\mathbf{0}$, a pro $i \geq 2$: zřetězení odzadu $\mathbf{0}$ a počínaje $k := i$ pokud $k \geq 2$ rekurzivně $\beta(k)$, $k := l(\beta(k)) - 1$, např. $\mathbf{I0I0I0}$ pro 5

- dekódování: $i := 1$ a opakovaně jestliže je další bit $\mathbf{I}$ tak s dalšími i bity tvoří kód $\beta(i)$
- délka $\sum_{j=1}^{k}(\lfloor \log_2 i \rfloor^j + 1) + 1$, $\lfloor \log_2 i \rfloor^k = 1$

Fibonacciho kódy

- kódování přirozených čísel, L. Pisano (Fibonacci)
- Fibonacciho reprezentace $a_1 a_2 \ldots$ přirozeného čísla $i \geq 1$: $i = \sum_{j=1} a_j F_j$, $a_j \in \{0, 1\}$, F_j j-té Fibonacciho číslo ($F_1 = 1, F_2 = 2, F_j = F_{j-1} + F_{j-2}$) – neobsahuje sousední 1
- = pro $i \geq 1$: zřetězení Fibonacciho reprezentace i (jako bitů) a $\mathbf{I}$, např. $\mathbf{000II}$ pro 5 – končí $\mathbf{II}$
- délka $\leq \lfloor \log_\phi \sqrt{5} n \rfloor + 1$, $\phi = \frac{1}{2}(1 + \sqrt{5}) \approx 1.618$ tzv. zlatý řez
- prefixové, robustnější než jiné kódy čísel
- další (zobecněné) založené na k-krokových (zobecněných) Fibonacciho číslech

Golombovy kódy

- kódování přirozených čísel, S. W. Golomb
- parametr přirozené číslo $j > 0$

$=$ pro $i \geq 0$ zřetězení dvou kódů:

1. unární kód $q = \lfloor \frac{i}{j} \rfloor$ ($=$ celé části $\frac{i}{j}$)
2. $\lfloor \log_2 j \rfloor$-bitová binární reprezentace $r = i - qj$ ($=$ zbytek po celočíselném dělení $\frac{i}{j}$) pro $r = 0, 1, \ldots, 2^{\lceil \log_2 j \rceil} - j - 1$ a $\lceil \log_2 j \rceil$-bitová binární reprezentace $r + 2^{\lceil \log_2 j \rceil} - j$ pro $r = 2^{\lceil \log_2 j \rceil} - j, \ldots, j - 1$

Příklad

$j = 5$

$\lfloor \log_2 5 \rfloor = 2$-bitové binární reprezentace $r = 0, 1, 2$ a $\lceil \log_2 5 \rceil = 3$-bitové binární reprezentace $r + 3$ pro $r = 3, 4$

$0 \mapsto \mathbf{000}$, $1 \mapsto \mathbf{00I}$, $2 \mapsto \mathbf{0I0}$, $3 \mapsto \mathbf{0II0}$, $4 \mapsto \mathbf{0III}$, $5 \mapsto \mathbf{I000}$, $6 \mapsto \mathbf{I00I}$, $\ldots$

Golombovy kódy

- délka pro malé j z malé rychle narůstá, pro velké j z delší narůstá pomalu
- prefixové, pro $j = \lceil -\frac{1}{\log_2 p} \rceil$ (přesněji $j = \lceil -\frac{\log_2(1+p)}{\log_2 p} \rceil$) optimální při $P(i) = p^{i-1}(1-p)$ – geometrické rozdělení pravděpodobnosti, např. posloupnost (run z RLE) $i-1$ výskytů symbolu s vysokou pravděpodobností výskytu p ukončená jedením výskytem jiného symbolu s nízkou pravděpodobností $1-p$ (např. prohra a výhra) $\rightarrow$ (adaptivní) Golomb RLE
- použití např. v bezeztrátové kompresi obrazu (JPEG-LS)

Riceovy kódy

- $\sim$ Golomb-Riceovy kódy, R. F. Rice (Rice machine)
- $=$ Golombovy kódy pro $j = 2^k$ pro nějaké (celé nezáporné) k
- jednodušší kódování (a dekódování) pro $i \geq 0$: zřetězení unárního kódu pro zbývajících $q = \lfloor \frac{i}{j} \rfloor$ bitů a k nejméně významných bitů binární reprezentace i
- délka $\lfloor \frac{i}{j} \rfloor + k + 1$, optimální při $P(i) = \frac{1}{2^{\frac{i}{j}+k+1}}$
- použití např. v bezeztrátové kompresi audia (MPEG-4 ALS, FLAC)