KATEDRA
INFORMATIKY
UNIVERZITA PALACKÉHO V OLOMOUCI

KMI/KOM - Komprese dat

Poznámky z výuky *(2025 – 2026)*
Verze z *01. 04. 2026*

Vojtěch Netrh
vojtanetrh@gmail.com

Obsah

1 Režijní informace

Vyučující: Jan Outrata

Otázky jsou dány tím co probereme v přednáškách, asi to co má na stránkách rozvržené do 7 kapitol.

1.1 Týdenní rozpis

1. týden – 1.-16. slajd

1.2 Státnicové okruhy

1. Základní pojmy, taxonomie metod, míry komprese, typy modelů dat, pravděpodobnostní a Markovův model dat.
2. Run-length encoding a Move-to-front kódování.
3. Kódování čísel: unární kód, Eliasovy, Fibonacciho a Golombovy kódy.
4. Tunstallův kód a Shannon-Fanovo kódování.
5. Huffmanovo kódování se semi-adaptivním modelem.
6. Huffmanovo kódování s adaptivním modelem.
7. Aritmetické kódování.
8. Kontextové kódování (PPM).
9. Blokové třídění.
10. Třída slovníkových metod LZ77.
11. Třída slovníkových metod LZ78, reprezentace slovníku.
12. Slovníková metoda LZW.

2 Obecně o kompresi dat

Jedná se o proces, při kterém se zmenší velikost informačního obsahu, ale zachová se daná míra obsažené informace. Jedná se dost o experimentální obor. Šetří místo (vhodné pro ukládání i přenos). Dnes je samozřejmostí úplně všude.

Příklady z minulosti: Brailovo písmo, morseovka.

Proces se dá rozdělit na 2 fáze:

1. Identifikování a modelování struktury dat (s vynecháním redundancí)
2. Kódování dat podle modelu

Strukturou se bere opakování vzorů (takže statistická = frekvence atd.), korelace mezi vzory. Pro různé části dat můžeme mít různé modely (nemusíme to dělat celé najednou).

> **Příklad 1**
>
> $x_1, x_2, ..., x_{12} = 2, 2, 4, 6, 7, 7, 7, 10, 10, 11, 11, 14$
>
> 1. čísla od 12 do 14 binárně $\Rightarrow$ 48b
> 2. 7 různých čísel binárně $\Rightarrow$ 36b
> 3. častější číslo = kratší kód $\Rightarrow$ **0I** pro 7, **III** pro 11, **II0** pro 10, … $\Rightarrow$ 33b
> 4. kódování opakování čísla 32b
> 5. malé rozdíly mezi sousedními čísly $\rightsquigarrow$ predikce $\Rightarrow$ 26b
> 6. vztah mezi čísly $\rightsquigarrow$ predikce $\Rightarrow$ 19b

Zachováním veškeré zachované informace se myslí: odstranění redundance (obsah co nemá žádnou „hodnotu“).

3 Kompresní techniky a metody

Používáme 2 algoritmy – kompresní a dekompresní (rekonstrukční).

3.1 Bezztrátová komprese

Dekomprimovaná data jsou stejná jako originální (nic se neztratí). Vhodné pro text, programové soubory, citlivé záznamy. Např. Huffmanovo kódování, aritmetické kódování, PPM, BWT, …

3.2 Ztrátová komprese

Při kompresi se nějaká data vynechají (dekomprimovaná data nejsou totožná s originálními). Díky této ztrátě je poskytnuta vyšší komprese. Hodí se pro obraz, video, zvuk, … Vzorkování a kvantizace, diferenční kódování, transformační a podpásmové kódování. Používá se v řadě klasických typech souborů (`jpeg`).

Míry kompresních algoritmů **Poznámka 2**

Kompresní algoritmy jsou lineární a konstantní, případně logaritmické.

Zkoumá se: **compression ratio** (poměr velikosti originálních a komprimovaných dat), **compression rate** (průměrná velikost komprimovaných dat na vzorek originálních) a **distortion** (míra zkreslení dat).

3.3 Fyzický model dat

Popis zdroje nebo procesu generování dat z hlediska fyzického fungování. Obecně je příliš složitý až nemožný.

3.4 Pravděpodobnostní model dat

Pravděpodobnost výskytu symbolů dat je nezávislá na výskytu ostatních symbolů (jedná se o nejjednodušší předpoklad). Statistický popis se zjistí empiricky (pozorováním). Používá se pro statistické bezeztrátové kompresní metody. Používá se tzv. *klasická pravděpodobnost*.

3.5 Markovův model dat

Výskyt symbolu x_j závisí na výskytu předchozích symbolů x_i (kde $i < j$). Vychází z pravděpodobnostního modelu. Tohoto modelu mohou být různé řády (od nultého). Vychází z *Markovova řetězce*, který pro k-tý Markův model je:

$$P(x_n \mid x_{n-1}, ..., x_{n-k}) = P(x_n \mid x_{n-1}, ..., x_{n-k}, ...)$$

Pokud má abeceda počet symbolů l, pak l^k je počet stavů. Pro *první Markovův model* platí

$$P(x_n \mid x_{n-1}) = P(x_n \mid x_{n-1}, x_{n-2}, x_{n-3}, ...)$$

Obecně modely vyšších k řádů mají vyšší míru komprese než nezávislé výskyty symbolů.

Markovův model **Příklad 3**

$x_1x_2\dots x_{10} = aababbabaa$

stavy modelu 1. řádu = posloupnosti (bezprostředně) předchozích symbolů délky 1 pro všechny symboly: a, b
$P(a) = \frac{5}{9}$, $P(b) = \frac{4}{9}$,
$P(a|a) = \frac{2}{5}$, $P(b|a) = \frac{3}{5}$, $P(a|b) = \frac{3}{4}$, $P(b|b) = \frac{1}{4}$

stavy modelu 2. řádu = posloupnosti (bezprostředně) předchozích symbolů délky 2 pro všechny symboly: aa, ab, ba, bb
$P(aa) = \frac{1}{8}$, $P(ab) = \frac{3}{8}$, $P(ba) = \frac{3}{8}$, $P(bb) = \frac{1}{8}$,
$P(a|aa) \to 0$, $P(b|aa) \to 1$, $P(a|ab) = \frac{2}{3}$, $P(b|ab) = \frac{1}{3}$, $P(a|ba) = \frac{1}{3}$, $P(b|ba) = \frac{2}{3}$,
$P(a|bb) \to 1$, $P(b|bb) \to 0$

3.6 Další modely dat

Slovníkový – odkazy do paměti předchozích symbolů v datech

Prediktivní – predikce symbolů na základě okolních symbolů v datech

4 Teorie informace

Jedná se o vědu, jak měřit, ukládat a přenášet data. Řeší problém náhodnosti a nejistoty pokud se bavíme o tom „jak moc informací“ je uloženo v datech. Informace totiž nemůže být měřena pouze její velikostí (třeba kolik místa zabírá na disku). Její zakladatelem je Claude Shannon. Používá se jako rámec pro bezeztrátové kompresní metody. Úzce souvisí s pravděpodobnostním modelem dat.

4.1 Klasická (Shannonova) teorie

Míra průměrná informace experimentu = výsledky experimentu (data nebo jevy). Informaci jevu A značíme jako $i(A)$, její pravděpodobnost je $P(A)$. Potom definujeme míru její informace jako

$$i(A) = -\log_b P(A)$$

Jednotka i je pro: 2 – bit, e – nat, 10 – hartley.

4.1.1 Entropie

Jde o průměrnou hodnotu vlastní informace pro všechny možné výstupy náhodného zdroje dat. Značí se H.

$$H(A_i) = \sum_i P(A_i)i(A_i) = -\sum_i P(A_i)\log P(A_i)$$

Shannon dokázal že pouze toto splňuje 3 požadavky:

1. spojitá funkce
2. pro stejně pravděpodobné jevy musí monotónně růst s počtem možností
3. musí být stejná při rozdělení na k podexperimentů

Shannov objevil tzv. *Noiseless Source Coding Theorem*, který říká, že entropie představuje (teoretickou) spodní hranici pro bezztrátovou kompresi. Čili pokud si entropii spočítáme, víme že se nikdy nemůžeme dostat pro výsledný počet bitů na číslo. Dá se aplikovat i na přirozený jazyk, kde to jaké písmeno bude následující je silně dáno předchozím znakem (např. často po q následuje u).

[Příklady v prezentacích]

4.2 Algoritmická (Kolmogorova) teorie

Základem jak plyne z názvu je *Kolmogorova složitost* $K(x)$. Jde o velikost nejkratšího algoritmu (programu) potřebného k vygenerování posloupnosti x. Velikost tohoto programu by teoreticky měla odpovídat nejvyšší míře komprese, které jde dosáhnout. Avšak neexistuje žádný systematický způsob jak tento algoritmus sestrojit nebo se k němu alespoň přiblížit.

5 Kódování

Základem kódování je kód, takže musíme znát abecedu a její symboly. Známe např. blokový kód, kde všechna slova mají stejnou délku (v praxi ASCII).

Jednoznačně dekódovatelný kód **Definice 4**

Každá neprázdná posloupnost symbolů z kódované abecedy je zřetězením nejvýše jedné posloupnosti kódových slov.

Prostě že daný kód můžeme namapovat jen na jediný řetězec.

Každý blokový je jednoznačný. Konkrétní příklady: $\{0, 01, 011, 111\}$ **ano**, $\{0, 01, 10, 11\}$ **ne**.

Prefixový kód **Definice 5**

Žádné kódové slovo není prefixem jiného kódového slova.

Např. $\{0, 10, 110, 111\}$

Kraftova věta **Theorem 6**

Prefixový kód s k kódovými slovy délek $l_1, l_2, ..., l_k$ nad kódovanou abecedou velikosti m existuje právě když

$$\sum_{i=1}^{k} m^{-l_i} \leq 1$$

Vzorec se nazývá *Kraftovou nerovností*.

5.1 Optimální kód

McMillanova věta **Theorem 7**

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

$$\sum_{i=1}^{k} m^{-l_i} \leq 1$$

Shannon noisless coding theorem **Theorem 8**

6 Základní techniky a kódování čísel

6.1 Run-length kódování

6.2 Move-to-front kodování

6.3 Kódování čísel

6.3.1 Unární kódování

6.3.2 Eliasovy kódy

6.3.3 Fibonacciho kódy

6.3.4 Golombovy kódy

6.3.5 Riceovy kódy

7 Statistické metody

7.1 Hufmanovo kódování

Sestavuje se strom podle toho, které znaky se kolikrát vyskytují. Strom obsahuje prázdný znak, jednotlivé znaky i různé kombinace. Uzel je tím blíže ke kořenu čím více má výskytů. Provádí se různé operace, které „upravují“ strom.