

Teorie informace a kódování (KMI/TIK)

Bezpečnostní kódy



Lukáš Havrlant

Univerzita Palackého

13. listopadu 2012

Konzultace

- V pracovně 5.076.
- Každý čtvrtek 9.00–11.00.
- Emaily:
 - `lukas@havrlant.cz`
 - `lukas.havrlant@upol.cz`
- Web: `http://tux.inf.upol.cz/~havrlant/`

Doporučená literatura

- Adámek J.: Foundations of Coding. J. Wiley, 1991.
- Adámek J.: Kódování. SNTL, Praha, 1989. (lze stáhnout na <http://notorola.sh.cvut.cz/~bruxy/>)
- Ash R. B.: Information Theory. Dover, New York, 1990.

Úvod

- Při přenosu dat může dojít k chybě → jak odhalit chybu? (znáte: CRC, hash) Jak opravit chybu?
- Chyba může být dvojího typu:
 1. Záměna znaku: 101010 → 101000 (o tento typ se budeme zajímat)
 2. Vytvoření nového znaku: 101010 → 1010101. Pohlčení znaku: 101010 → 10110
- Základní princip: redundance informace.
- Čeština je hodně redundantní: „opakkvání“ → „opakování“, ale „sekers“ → „sekera“ nebo „sekery“?
- Všechna rodná čísla vytvořená od roku 1986 včetně jsou dělitelná jedenácti.

Příklad kódování objevující jednu chybu

- Kontrola sudé parity: kód doplní na konec binárního slova 0/1 tak, aby slovo mělo sudý počet jedniček.
- Vstupem je slovo $w \in \{0,1\}^+$, doplníme $c \in \{0,1\}$, vznikne slovo wc .
- Pokud $w = 001$, pak $c = 1$ a $wc = 0011$.
- Pokud $w = 101$, pak $c = 0$ a $wc = 1010$.
- Kód objevuje 1 chybu: pokud přijmeme 1011, nastala chyba.
- Kód nerozpozná dvě chyby! Vyšleme 1010, přijmeme 1111.

Příklad kódování opravující jednoduché chyby

- Opakující kód – symbol odešleme třikrát.
- $0 \longrightarrow 000$; $1 \longrightarrow 111$.
- Pokud přijmeme 010, poslali jsme původně 000 (nebo nastala více než jedna chyba).

Cíl

- Cíl: sestrojit kód, který objeví/opraví co nejvíce chyb při co nejnížší redundanci.
- Uvažujeme pouze blokové kódy.

Kódová a nekódová slova

- Zdrojová abeceda A , kódová abeceda B , délka kódového slova je n , pak pro kód C platí: $C \subseteq B^n$.
- $B^n = \{b_1 b_2 \dots b_n \mid b_i \in B \text{ pro } i = 1, 2 \dots n\}$
- C je množina kódových slov, $B^n \setminus C$ množina nekódových slov.
- Přijmeme-li nekódové slovo, tak během přenosu nastala chyba.
- Pokud přijmeme kódové slovo, tak během přenosu...?
- Kontrola sudé parity: pro $B = \{0, 1\}$, $n = 4$ máme
 - $B^4 = \{0000, 0001, 0010, 0011, 0100, 0101, \dots\}$; $|B^4| = 2^4 = 16$
 - $C = \{0000, 0011, 0101, 0110, \dots\}$; $|C| = 8$

Informační poměr kódu (rate)

- Máme slovo délky k . Přidáme symboly z kódové abecedy a získáme slovo o délce n .
- Původním k symbolům říkáme *informační symboly*.
- $n - k$ symbolům říkáme *kontrolní symboly*.
- Poměr $R(C) = \frac{k}{n}$ měří redundanci kódu.

Define: Necht' $C \subseteq B^n$ je blokový kód o délce n . Jestliže existuje bijektivní zobrazení $\varphi : B^k \rightarrow C$, řekneme, že kód C má k informačních symbolů a $n - k$ kontrolních. *Informační poměr kódu* $R(C)$ je

$$R(C) = \frac{k}{n}.$$

Příklady informačních poměrů

- **Opakovací kód**

- Kód C_n má na abecedě B kódová slova $\{a^n \mid a \in B\}$
- Pro $B = \{0, 1\}$ máme $C_3 = \{000, 111\}$, $C_5 = \{00000, 11111\}$
- $k = 1$; $R(C_n) = \frac{1}{n}$

- **Kontrola sudé parity**

- Pro kód délky n platí: $C_n = \{w \text{ má sudý počet } 1 \mid w \in \{0, 1\}^n\}$
- $C_3 = \{000, 011, 101, 110\}$
- Pro C_3 : $|C| = 4$, $k = 2$; $R(C_3) = \frac{2}{3}$
- Pro C_n : $|C| = 2^{n-1}$, $k = n - 1$; $R(C_n) = \frac{n-1}{n}$

Hammingova vzdálenost

Definice: Pro slova $u = u_1 \cdots u_n, v = v_1 \cdots v_n \in B^n$ definujeme zobrazení

$$\delta(u, v) = |\{i \mid 1 \leq i \leq n, u_i \neq v_i\}|.$$

Tj. $\delta(u, v)$ vrací počet pozic, na kterých se slova u a v liší. Zobrazení δ nazýváme **Hammingova vzdálenost**. δ je metrikou na B^n :

$$\delta(u, v) \geq 0 \quad (= 0 \text{ iff } u = v),$$

$$\delta(u, v) = \delta(v, u),$$

$$\delta(u, v) \leq \delta(u, w) + \delta(w, v).$$

$$\delta(011, 101) = 2, \quad \delta(011, 111) = 1, \quad \delta(0110, 1001) = 4, \\ \delta(0123, 0145) = 2, \quad \delta(01, 01) = 0$$

Důkaz trojúhelníkové nerovnosti

Chceme dokázat: $\delta(u, v) \leq \delta(u, w) + \delta(w, v)$. Označme:

$$U = \{i \mid 1 \leq i \leq n, u_i \neq w_i\},$$

$$V = \{i \mid 1 \leq i \leq n, w_i \neq v_i\}.$$

Pak $U \cup V$ je množina indexů, na kterých se u může lišit od v . Pro všechna $i \in \{1, 2, \dots, n\} \setminus (U \cup V)$ platí, že $u_i = w_i = v_i$. Dále

$$\delta(u, v) \leq |U \cup V|.$$

$\delta(u, v)$ může být menší, pokud $u_i \neq w_i \wedge w_i \neq v_i$, ale $u_i = v_i$. Pak $i \in U \cup V$, ale $i \notin \{j \mid 1 \leq j \leq n, u_j \neq v_j\}$. Dostáváme:

$$\delta(u, v) \leq |U \cup V| \leq |U| + |V| = \delta(u, w) + \delta(w, v). \quad \square$$

Minimální vzdálenost

Definice: Minimální vzdálenost kódu C , zapíšeme $d(C)$, je minimální vzdálenost dvou různých slov z C :

$$d(C) = \min\{\delta(u, v) \mid u, v \in C, u \neq v\}.$$

Příklady:

$$C_1 = \{0011, 1010, 1111\}, \quad d(C_1) = 2$$

$$C_2 = \{0011, 1010, 0101, 1111\}, \quad d(C_2) = 2$$

$$C_3 = \{000, 111, 100\}, \quad d(C_3) = 1$$

Definice: Necht' C je nějaký kód o délce n a $u \in C$. Pak $B_q(u) = \{v \in B^n \mid \delta(u, v) \leq q\}$ nazveme koule se středem v u a poloměrem q .

Důsledek: Kód C má minimální vzdálenost $d(C) \geq d$ právě tehdy, když pro všechna $u \in C$ platí

$$B_{d-1}(u) \cap C = \{u\}.$$

Objevování chyb

- t chyb nastane, pokud pošleme slovo u , přijmeme slovo v a $\delta(u, v) = t$. (Slovo u se od v liší v t symbolech.)

Definice: Kód C objevuje d chyb, pokud při odeslání kódového slova a vzniku $t \leq d$ chyb, přijmeme vždy nekódové slovo.

Věta: Kód C objevuje d chyb právě když $d(C) > d$.

Největší d takové, že C objevuje d chyb je $d(C) - 1$.

Příklady

- **Kontrola sudé parity:**

- $d(C) = 2$, takže kód objevuje 1 chybu.
- Proč $d(C) = 2$? Nechť $u, v \in B^{n-1}$, $u \neq v$. Slova zakódujeme, dostaneme ua, vb , kde $a, b \in B$.
 - Pokud $\delta(u, v) \geq 2$, pak jistě $\delta(ua, vb) \geq 2$.
 - Pokud $\delta(u, v) = 1$, pak právě jedno ze slov obsahuje lichý počet jedniček a tedy $a \neq b$ a $\delta(ua, vb) = 2$.

- **Opakovací kód:**

- $d(C_n) = n$, takže kód objevuje $n - 1$ chyb.

Oprava chyb

- Pokud přijmeme nekódové slovo v , *opravíme* ho na takové kódové slovo u , které je slovu v nejbližší.
- $\longrightarrow$ pro každé nekódové slovo musí existovat jediné kódové slovo, které je mu nejbližší.

Definice: Kód C opravuje d chyb pokud pro každé kódové slovo $u \in C$ a každé slovo $v \in B^n$ takové, že $\delta(u, v) \leq d$, platí, že slovo u má od slova v nejmenší vzdálenost ze všech slov z C .

Tedy pokud $u \in C$ a $\delta(u, v) \leq d$, pak $\delta(u, v) < \delta(w, v)$ pro všechna $w \in C, w \neq u$.

Oprava chyb podruhé

Věta: C opravuje d chyb právě tehdy, když $d(C) > 2d$.

- Největší d takové, že C opravuje d chyb je $d = \lfloor (d(C) - 1)/2 \rfloor$. Příklady:
- Kontrola sudé parity má $d(C) = 2$, takže opravuje 0 chyb.
- Opakovací kód C_3 opravuje $(3 - 1)/2 = 1$ chybu.

Důkaz

Věta: C opravuje d chyb právě tehdy, když $d(C) > 2d$.

„ $\Leftarrow$ “: Nechť $d(C) > 2d$. Dále nechť $u \in C, v \in B^n : \delta(u, v) \leq d$.

Sporem předpokládejme, že pro $w \in C, w \neq u$ platí $\delta(w, v) \leq \delta(u, v)$. Použijeme trojúhelníkovou nerovnost:

$$\delta(w, v) + \delta(u, v) \geq \delta(w, u)$$

Přitom ale $\delta(u, v) \leq d$ a zároveň $\delta(w, v) \leq d$, takže:

$$2d = d + d \geq \delta(w, v) + \delta(u, v) \geq \delta(w, u).$$

Dostáváme $2d \geq \delta(w, u)$, což je spor s předpokladem, protože $u, w \in C$.

Důkaz: pokračování

Věta: C opravuje d chyb právě tehdy, když $d(C) > 2d$.

„ $\Rightarrow$ “: Nechť C opravuje d chyb. Opět sporem předpokládejme, že $d(C) \leq 2d$ a že tak existují $u, v \in C$; $u \neq v$ taková, že $t = \delta(u, v) \leq 2d$. Rozdělme důkaz na dva případy.

Když je t sudé. Dále nechť w je slovo, které má vzdálenost $t/2$ od obou slov u, v . Slovo w sestojíme takto: nechť

$$I = \{i \mid u_i \neq v_i\} = \{i_1, i_2, \dots, i_t\},$$

takže $|I| = t$. Pak $w_i = v_i$ pro všechna $i \in \{i_1, i_2, \dots, i_{t/2}\}$ a $w_i = u_i$ pro všechna ostatní i . Potom $\delta(u, w) = t/2 = \delta(v, w)$, tedy $\delta(u, w) = \delta(v, w) \leq d$, což je spor s tím, že C opravuje d chyb.

Důkaz: finále

Když je t liché. Pak $t + 1$ je sudé, přitom $t + 1 \leq 2d$. Stejně jako v předchozím kroku sestavíme slovo w tak, že $w_i = v_i$ pro všechna $i \in \{i_1, i_2, \dots, i_{(t+1)/2}\}$ a $w_i = u_i$ pro všechna ostatní i . Tzn., že

$$\delta(u, w) = (t + 1)/2$$

$$\delta(v, w) = (t - 1)/2$$

To je ale průšvih, protože $\delta(u, w) > \delta(v, w)$, přitom $\delta(u, w) = (t + 1)/2 \leq d$. Jinými slovy: vyslali jsme slovo u , cestou se poškodilo $(t + 1)/2 \leq d$ symbolů a my přijali slovo w , které má ale blíže ke slovu v ! Kód tak nemůže opravovat d chyb. □

Příklady

- **Kontrola sudé parity:** $d(C) = 2$, kód neopravuje žádnou chybu.
- **Opakovací kód:** $d(C_n) = n$, kód opravuje $\lfloor (n - 1)/2 \rfloor$ chyb.

Dvoudimenzionální kontrola sudé parity

Informační symboly umístíme do matice $p - 1 \times q - 1$, ke které přidáme jeden sloupec a jeden řádek pro kontrolní součty, tj. aby celý řádek/sloupec obsahoval sudý počet jedniček. Poslední bit nastavíme tak, aby celé slovo mělo sudou paritu. Příklad pro $(p, q) = (4, 5)$:

0	1	1	0	0
0	0	1	0	1
1	1	1	0	1
1	0	1	0	0

Kód opravuje 1 chybu.

Kód dva-z-pěti

- Binární blokový kód s délkou $n = 5$.
- Kódová slova obsahují 2 jedničky. Celkem: $\binom{5}{2} = 10$ možností.
- 00011, 00101, 00110, 01001, 01010, ...
- Populární na reprezentování číslíc.
- Dekódování: Můžeme přidat váhu jednotlivým pozicím: 0, 1, 2, 3, 6. Potom můžeme kódové slovo 00101 dekodovat jako:
$$0 \cdot 0 + 0 \cdot 1 + 1 \cdot 2 + 0 \cdot 3 + 1 \cdot 6 = 8.$$
- Trojka je reprezentována dvěma slovy: 10010 a 01100. Druhá možnost se používá pro nulu.

Systematický kód

Definice: Blokový kód C o délce n nazýváme **systematický** kód, pokud pro nějaké $k < n$, pro všechna slova $u \in B^k$, existuje právě jedno slovo $v \in B^{n-k}$ takové, že $uv \in C$ je kódové slovo.

- Opakovací kód a kód sudé parity jsou systematické kódy.
- $R(C) = \frac{k}{n}$

Cvičení

- Jaká je $d(C)$ kódu dva-z-pěti? Jaký je informační poměr?
- Jaká je $d(C)$ dvoudimenzionální kontroly sudé parity? Jaký je informační poměr?
- Za použití dvoudimenzionálního $(4, 5)$ -kódu (4 řádky, 5 sloupců) jsme přijali slovo 01001 11101 11100 01100. Je to kódové slovo? Pokud ne, lze opravit?
- Trojúhelníková nerovnost říká, že $\delta(u, v) \leq \delta(u, w) + \delta(w, v)$. Existují navzájem různá slova $u, v, w \in B^n$ pro nějaké B a nějaké n tak, že $\delta(u, v) = \delta(u, w) + \delta(w, v)$?