Úvod do teorie informace

Ing. Jan Přichystal, Ph.D.

PEF MZLU v Brně

24. září 2007

Úvod

Výměna informací s okolím nám umožňuje udržovat vlastní existenci.

Proces zpracování informací je trvalý, nepřetržitý, ale ovlivnitelný.

Zabezpečení informací je spojeno s lidským jednáním a je údělem celé společnosti, bez ohledu na vývojový stupeň materiálních podmínek.

Problémy se zpracováním informací se prohloubily během 20. století a neexistují lidé, kteří by je všechny byli schopni zpracovat nebo evidovat.

Úvod

Využívání a efektivní práce s informacemi vyžaduje o nich něco vědět.

- Co jsou to informace?
- Co jsou relevantní informace
- Jak je získáme a jak zhodnotíme jejich využitelnost?
- Jak se přenášejí a jak jsou uloženy?

Základní podmínkou úspěšnosti jednotlivců je permanentní osvojování nových znalostí vytvořených jinými a tvorba znalostí vlastních. Abychom byli připraveni, musíme mít dostatek informací o informacích a možnostech manipulace s nimi.

Pojem informace

Informace je obsah jakéhokoli oznámení, údaje o čemkoli, s určením pro přenos v prostoru a čase. V nejširším slova smyslu je to obsah vztahů mezi materiálními objekty, projevující se změnami těchto objektů. (Terminologický slovník informatiky)

Informace je obsah zprávy, sdělení, objasnění, vysvětlení, poučení. (Slovník cizích slov)

Informace jsou údaje, čísla, znaky, povely, instrukce, příkazy, zprávy apod. Za informace považujeme také podněty a vjemy přijímané a vysílané živými organismy. (Oborová encyklopedie VT)

Jak informace chápat?

Informace – z hlediska kvalitativního (obsah sdělení, význam zprávy) tím se zabývá INFORMATIKA

Informace – z hlediska kvantitativního (množství a jeho měření) tím se zabývá TEORIE INFORMACE

Teorie informace

Claude Shannon – základy teorie informace, stanovil možnosti měření informačního množství

Shannonova definice informace:

Informace je míra množství neurčitosti nebo nejistoty o nějakém náhodném ději odstraněná realizací tohoto děje.

Informace rozšiřuje okruh znalostí příjemce.

Měření informačního množství

Entropie – název vypůjčený z fyziky, použitý pro měření informačního množství.

Jak kvantifikovat rozšíření okruhu znalostí příjemce?

Pravděpodobnost jevu – spojeno s individuálními vlastnostmi příjemce (Shannon)

Jevy a jejich realizace

Jev – náhodný proces s n možnými realizacemi (tah sportky, účast na přednášce, semafor na křižovatce apod.)

Realizace jevu – jeden projev, získání výsledku (vytažení 6 čísel, konkrétní počet osob na přednášce, svítící zelená na křižovatce apod.)

Požadované vlastnosti funkce pro výpočet množství informace

- Jev X má n realizací, množství informace je funkcí n.
- Je-li $n = 1$, jedná se o jev jistý, množství informace je rovno nule.
- Jevy X a Y probíhající současně a nezávisle, $p(x, y) = p(x) \cdot p(y)$: množství informace je dáno součtem množství jednotlivých jevů: $f(x, y) = f(x) + f(y)$
- Jev X má n realizací, jev Y má m realizací. Je-li $m > n$, pak chceme i $f(m) > f(n)$

Výpočet vlastní informace

Funkce, která vyhovuje uvedeným podmínkám, je logaritmus.

$$I(x) = \log n$$

Zde předpokládáme, že pravděpodobnost každé realizace je stejná. Má-li jev n realizací, pak můžeme psát

$p(x) = 1/n$, odsud pak $n = 1/p(x)$

Výpočet vlastní informace

Buď X množina výsledků náhodného děje, x výsledek realizace a $p(x)$ pravděpodobnost tohoto výsledku. Každému x z X pak lze přiřadit reálné číslo $I(x)$ nazývané vlastní informace o výsledku x, pro než platí:

$$I(x) = -\log p(x), \qquad (0 \geq p(x) \geq 1)$$

Číslo $I(x)$ představuje množství informace obsažené ve výsledku x.

Základ logaritmu – principiálně není podstatný. Ale používají se logaritmy o základu 2. Pak dostáváme výsledek v bitech.

Entropie

Jak spočítat informační množství celého jevu? Pomůžeme si shrnutím všech vlastních informací jednotlivých realizací.

Předpokládejme, že jev X má n realizací

$$X = x_1, x_2, \ldots, x_n$$

s pravděpodobnostmi

$$p(x_1), p(x_2), \ldots, p(x_n)$$

Výpočet entropie jevu

Entropie $H(X)$ je dána určitou střední hodnotou vlastních informací všech realizací jevů:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \cdot \log p(x_i)$$

Entropie zahrnující informační množství celého jevu se nazývá též úplná informace.

Kódování informace

Základní podmínkou komunikace je vytvoření signálního komunikačního kanálu.

Informaci je pro tento účel nutné transformovat, tj. vyjádřit v jiném jazyce s jinou abecedou.

Přiřazení znaků jedné abecedy znakům jiné abecedy se nazývá kódování, inverzní postup pak dekódování.

Předpis, který toto přiřazování definuje, se nazývá kód.

Kvalita kódování, redundance

Z hlediska optimálního přenosu je efektivní kód, který obsahuje minimální počet informačních prvků, každý znak kódu tedy má maximální entropii.

Kvantitativně je hospodárnost kódu vyčíslitelná redundancí (nadbytečností), podle vztahu:

$$R = 1 - H/H_{max}$$

Způsoby kódování

Rovnoměrné kódování – každému znaku je přiřazen stejně dlouhý kód. Obvykle je jednodušší, rychlejší na zpracování, ale méně hospodárné. (Baudot)

Nerovnoměrné kódování – každému znaku je přiřazen jinak dlouhý kód. Pro konstrukci a zpracování je obtížnější, může však být maximálně hospodárné. (Shannon-Fano, Huffman)

Příklady kódů

Zdroj produkuje 4 znaky A, B, C, D.

Předpokládáme pravděpodobnosti znaků:

znak	$p_1(x)$	kód 1	kód 2
A	0,25	00	0
B	0,25	01	10
C	0,25	10	110
D	0.25	11	111

znak	$p_2(x)$	kód 1	kód 2
A	0,5	00	0
B	0,25	01	10
C	0,125	10	110
D	0.125	11	111

Shannon-Fanův algoritmus

1. Znaky uspořádáme sestupně podle pravděpodobnosti jejich výskytu.
2. Vypočteme kumulativní pravděpodobnosti.
3. Rozdělíme znaky do dvou skupin tak, aby jejich součtové pravděpodobnosti byly blízké 0,5.
4. Krok 3 opakujeme tak dlouho, dokud existují vícečlenné skupiny znaků.

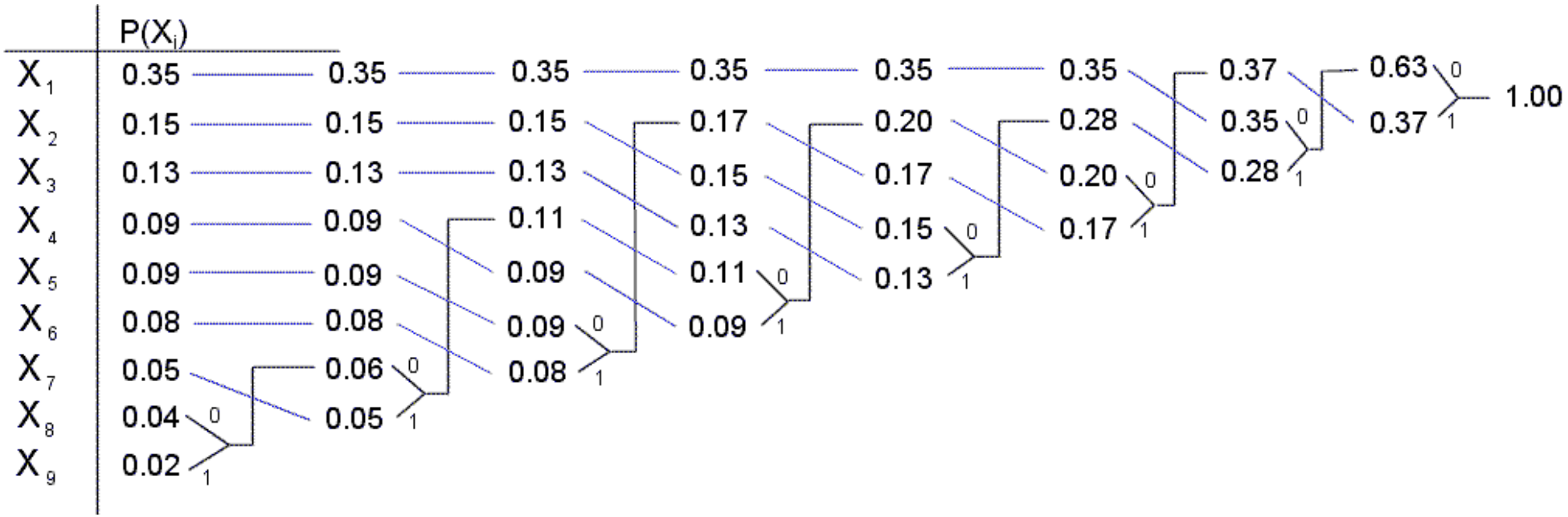
Shannon-Fanův algoritmus

znak	p(x)	s	skupiny			výsledek
x_1	0,30	1,00		0		00
x_2	0,24	0,70	0	1		01
x_3	0,20	0,46		0		10
x_4	0,15	0,26	1		0	110
x_5	0,11	0,11		1	1	111

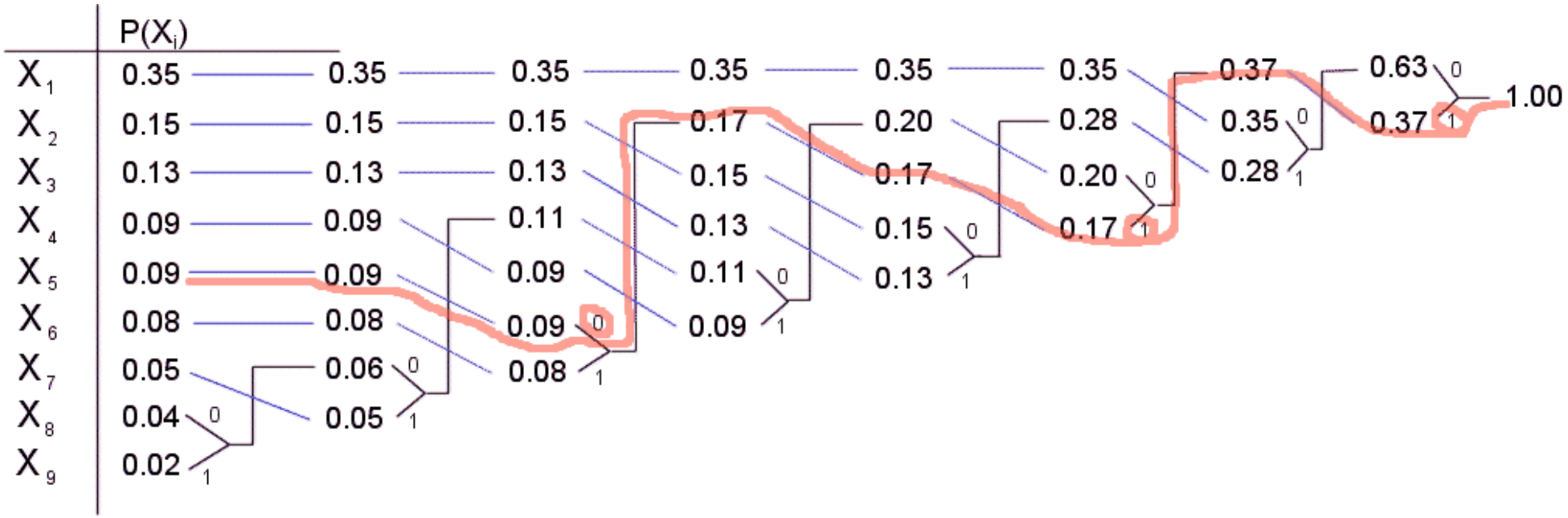
Huffmanovo kódování

1. Sečteme poslední dvě pravděpodobnosti a vytvoříme nový sloupec pravděpodobností, kde ty dvě, které jsme sčítali nahradí jejich součet.
2. Všechny pravděpodobnosti v novém sloupci seřadíme sestupně podle velikosti a propojí se spojnicemi s hodnotami v původním sloupci.
3. Spojnice pravděpodobností $p(x_{n-1})$ a $p(x_n)$ se sjednotí, ale předtím přiřadíme $p(x_n)$ bit kódového slova s hodnotou 1 a $p(x_{n-1})$ bit s hodnotou 0.
4. Takto postupujeme, dokud se součet posledních dvou čísel nerovná 1.
5. Závěrečné kódování každého slova pak probíhá po spojnicích jako „sbírání“ zapsaných bitů kódového slova tak, že jdeme po spojnicích a zapisujeme všechny bity, které po cestě potkáme.
6. Nakonec se celý zápis obrátí odzadu dopředu a výsledkem je kódové slovo pro danou událost.

Huffmanovo kódování 1.



Huffmanovo kódování 2.



Výsledný kód

událost	$p(x_i)$	kód
x_1	0.35	00
x_2	0.15	010
x_3	0.13	011
x_4	0.09	101
x_5	0.09	110
x_6	0.08	111
x_7	0.05	1001
x_8	0.04	10000
x_9	0.02	10001

Zabezpečení informace při přenosu

Detekce chyby

- zabezpečení paritou
- kontrolní součet (CRC)
- Hammingův kód

Zabezpečení proti neoprávněnému čtení

- šifrování
- podepisování

Zabezpečení paritou

Ke každému úseku dat je připojen další bit, který svou hodnotou doplňuje počet binárních jedniček na počet lichý nebo sudý (sudá/lichá parita)

$$10010011\,10110101 \longrightarrow 10010011\mathbf{0}\,10110101\mathbf{1}$$

Kontrolní součet

Data se rozdělí na úseky požadované délky (8, 16, 32 bitů) a tyto úseky se sečtou po bitech bez přenosu. Vzniklý úsek dat se připojí k datům přenášeným.

10100010	Data
11010111	
11010101	
01100110	
11000110	CRC

Závěr

Děkuji za pozornost

Dotazy?