TESTOVÁNÍ SLOŽENÉ HYPOTÉZY O ENTROPII

Jaroslava FEISTAUEROVÁ, ÚTIA ČSAV, Praha

Shannonova entropie hraje důležitou roli nejen v teorii přenosu, ale také v teorii rozhodování, klasifikaci, rozpoznávání atd.

V některých aplikacích je třeba rozhodnout, zda entropie leží či neleží v daném intervalu. Problémem testování hypotézy tohoto typu se zabývá tento příspěvek.

Nejprve zavedeme označení. Dvojicí (S, P) značíme diskrétní stochastický systém. $S = \{s_1, \dots, s_m\}$, $m > 1$, je stavový prostor a $P = (p_1, \dots, p_m)$ je pravděpodobnostní distribuce na S. Množinu všech možných pravděpodobnostních distribucí na S označíme $\mathbf{P}$. Shannonova entropie $H(P) = -\sum_{i=1}^{n} p_i \log p_i$, $P \in \mathbf{P}$, nabývá hodnot z intervalu $\langle 0, \log m)$.
Hypotéza o entropii musí mít tudíž tvar podmnožiny $\mathcal{H} \subset \langle 0, \log m)$. Tuto podmnožinu nazveme entropickou hypotézou.

Klasická statistická hypotéza o P má tvar podmnožiny $\mathbf{H} \subset \mathbf{P}$. Ke každé entropické hypotéze $\mathcal{H}$ existuje statistická hypotéza $\mathbf{H}$ ekvivalentní s $\mathcal{H}$ v následujícím smyslu: $P \in \mathbf{H} \iff H(P) \in \mathcal{H}$. Tato hypotéza je dána vztahem

$$\mathbf{H} = \{P \in \mathbf{P} : \ H(P) \in \mathcal{H}\}. \tag{1}$$

Existuje však statistická hypotéza $\mathbf{H}$, která není ekvivalentní s žádnou entropickou hypotézou.

Uvažujme empirické distribuce $P_n = (p_{n1}, \dots, p_{nm}) \in \mathbf{P}$ na S, $n = 1, 2, \dots$, kde

$$p_{ni} = \frac{\#\{1 \le k \le n : S_k = s_i\}}{n}, \quad i = 1, \dots, m$$

a $S_1, \dots, S_n$ je n nezávislých náhodných výběrů z (S, P). Dále uvažujme odpovídající entropie $H(P_n)$, neprázdný interval $\mathcal{H} = \langle h_1, h_2 \rangle \subset \langle 0, \log m)$ a odpovídající uzavřenou množinu $\mathbf{H} \subset \mathbf{P}$ definovanou vztahem (1). Jestliže $P \in \mathbf{H}$ a $H(P) \neq 0$, pak $H(P_n)$ má asymptoticky normální rozdělení se střední hodnotou $H(P) \in \langle h_1, h_2 \rangle$ a variancí $\sigma^2(P) \le s^2$, kde

$$s^2 = \sup_{P \in \mathbf{H}} \left[\sum_{i=1}^{m} p_i \log^2 p_i - H^2(P) \right]. \tag{2}$$

Veličina s^2 je kladná a platí

$$\sup_{P \in \mathbf{H}} \limsup_{n \to \infty} P^n \left(H(P_n) \notin \left\langle h_1 - \frac{sx}{\sqrt{n}}, h_2 + \frac{sx}{\sqrt{n}} \right\rangle \right) \le 2\Phi(-x), \quad x > 0,$$

kde $\varphi(x)$ je standardní normální distribuční funkce. Můžeme tedy formulovat následující tvrzení:

Věta 1: Jestliže s je definováno pomocí (2) pak test, který zamítá hypotézu $\mathcal{H}$, jestliže

$$H(P_n) \notin \left\langle h_1 - \frac{s\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)}{\sqrt{n}}, h_2 + \frac{s\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)}{\sqrt{n}} \right\rangle,$$

je asymptotický test na hladině α. ($\alpha \in (0, 1)$ a $\Phi^{-1}(\alpha)$ je kvantilová funkce standardního normálního rozdělení.)

Zbývá vyřešit problém stanovení hodnoty s. Postačí alespoň nalézt horní odhad $s_n \geq s$. Protože síla testu klesá s rostoucím $s_n - s$, snažíme se o nalezení s_n co možná nejmenšího. Toto s_n je dáno následujícím vztahem:

$$s_n^2 = \begin{cases} s^2(h_2) & \text{jestliže } h_2 \leq h_{(m)} \\ s^2(h_1) & \text{jestliže } h_1 \geq h_{(m)} \\ s^2(h_{(m)}) & \text{jindy,} \end{cases}$$

kde

$$s^2(h) = g(2^{-h}) - h^2 \quad \text{pro} \quad h \in (0, \log m),$$

$$g(x) = x \log^2 x + (1-x) \log^2 \frac{1-x}{m-1} \quad \text{pro} \quad x \in \left\langle \frac{1}{m}, 1 \right\rangle \quad \text{a}$$

$h_{(m)}$ je bod maxima funkce $s^2(h)$.

Funkce $s^2(h)$ není obtížné tabelovat pro různé hodnoty m. Jejich průběh je vidět na obr. 1. Některé hodnoty funkce $s^2(h)$ jsou uvedeny v tabulce 1.

Příklad 1.
Mějme 10-ti bitový stavový prostor (tj. $m = 1024$). Výběr o rozsahu $n = 10000$ nám poskytne empirickou distribuci P_n na $S = \{1, \ldots, 1024\}$ s entropií $H(P_n)$. Pro asymptotický test hypotézy $\mathcal{H} = (9, 10)$ na hladině významnosti $\alpha = 0.01$ je hodnota $s_n^2 = s^2(9) = 18.991$. Tedy hypotézu $\mathcal{H}$ zamítáme na hladině významnosti 0.01, jestliže

$$H(P_n) < 9 - \frac{\sqrt{18.991}\, \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)}{\sqrt{n}} = 8.8878.$$

Pokud $H(P_n) \geq 8.8878$, pak ještě nelze hypotézu $\mathcal{H}$ přijmout. Nicméně problém přijetí $\mathcal{H}$ na dané hladině můžeme vyřešit zamítnutím alternativy $\mathcal{K}$: $(0, 10) - \mathcal{H}$ na této hladině. Test pro tuto alternativu má tvar

$$H(P_n) > 9 + \frac{\sqrt{83.016}\, \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)}{\sqrt{n}} = 9.2347.$$

Jestliže je tedy tato nerovnost splněna, hypotézu $\mathcal{H}$ přijímáme. Pokud máme smůlu a hodnota $H(P_n)$ splňuje nerovnost

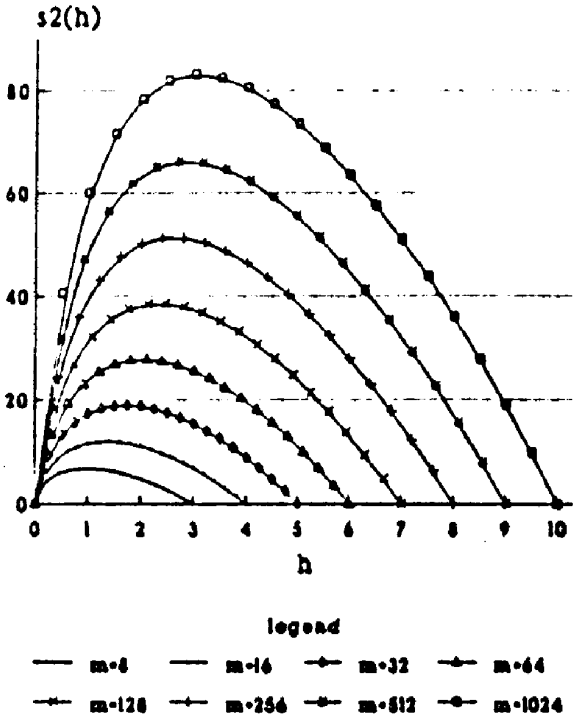
$$8.8878 \leq H(P_n) \leq 9.2347,$$

pak na dané úrovni 0.001 nelze rozhodnout.

Na závěr poznamenejme, že uvedený test byl v praxi použit při testování optimality kódové knihy pro kompresi řečového signálu. V této oblasti se ukazuje, že rozumným kritériem optimality je rovnoměrnost užití všech řečových segmentů z kódové knihy, tzn. je třeba testovat, zda entropie kódové knihy je blízká maximální hodnotě.

Tab. 1.

m	2	4	8	16	32	64	128	256	512	1024
$h_{(m)}$	0.203	0.560	0.972	1.360	1.714	2.034	2.326	2.592	2.836	3.062
$s^2(h_{(m)})$	1.121	3.237	6.750	11.894	18.825	27.650	38.443	51.252	66.105	83.016
$s^2(1)$	0.000	2.841	6.748	11.539	17.226	23.841	31.410	39.949	49.472	59.984
$s^2(\log m - 1)$	0.000	2.841	4.788	6.830	8.883	10.925	12.954	14.973	16.984	18.991



Obr. 1

Literatura

[1] J. Feistauerová, Statistická teorie kódových knih pro kompresi řeči, Kandid. disert. práce, ÚTIA ČSAV, 1989

[2] J. Feistauerová, I. Vajda, Testing System Entropy and Prediction Error Probability (zasláno k publikaci do IEEE Trans. Syst. Man, Cybern.), 1992 - zde uveden obsáhlý seznam literatury