

INFORMACE A DEZINFORMACE - STATISTICKÝ POHLED

M. Kotlíková, H. Mašková, A. Netrvalová, P. Nový, D. Spíralová, F. Vávra, D. Zmrhal

Klíčová slova: Informace, entropie, divergence, dezinformace.

Abstrakt: Při práci se statistickými odhady sdílené informace

$$I(X:T)=\sum_{t\in T}\sum_{x\in X}p(x,t)\lg\frac{p(x,t)}{p(x)p(t)}$$

nastávají některé zdánlivě neočekávatelné situace. Protože pravděpodobnosti (hustoty) $p(x,t),p(x),p(t)$ pro nás nejsou dostupné, pracujeme s jejich odhady $e(x,t)\approx p(x,t),e(x)\approx p(x),e(t)\approx p(t)$ nebo s jejich některou parametrickou reprezentací (ale ta je také odhadem). Potom odhad $\hat{I}_n(X:T)=\frac{1}{n}\sum_{i=1}^{n}\lg\frac{e(x_i,t_i)}{e(x_i)e(t_i)}$ „konverguje k" $\sum_{t\in T}\sum_{x\in X}p(x,t)\lg\frac{e(x,t)}{e(x)e(t)}$. Tento výraz budeme značit $I(X:T;e)$ a bude dále nazýván sdílenou informací mezi náhodnými proměnnými X a T při využití modelů $e(x,t),e(x),e(t)$. Obdobná situace nastává i pro další pojmy z teorie informace jako je entropie a divergence. V tomto sdělení jsou prezentovány některé vybrané možnosti využití uvedeného jevu a vlastnosti předložených „výběrových měr" (např. $I(X:T;e)$ může být i záporná). Rozdíl $Di(X:T;e)=I(X:T;e)-I(X:T)$ pak může být užitečnou charakteristikou „přijatelnosti" (i např. vychýlenosti) zvolených modelů $e(x,t),e(x),e(t)$. Pro tento rozdíl jsme zvolili pracovní název „dezinformace". Název je motivován tím, že „vhodnou" volbou modelů lze dosáhnout poměrně širokého spektra hodnot „výběrové" informace $I(X:T;e)$ a tím i případně možných rozhodnutí (manipulace s daty a modelem).

1 Úvod

Výše uvedený koncept výběrové informace je využitelný jak pro spojitá, tak i diskrétní rozdělení pravděpodobnosti. Autoři se domnívají, že otevírá prostor zkoumání a popisování jevů v úlohách klasifikace, rozhodování a i testování hypotéz. V tomto příspěvku budou prezentovány některé výsledky pro diskrétní rozdělení, zaměřené na využití při odhadování. Prezentace je zaměřena do dvou oblastí. První oblastí budou náznaky toho, jak lze získat klasické formule odhadů s využitím navrhovaného aparátu. Druhou oblastí bude náznak některých neotřelejších postupů. Příspěvek však bude z důvodů přehlednosti členěn podle klasických prostředků teorie informace, obě oblasti se budou prolínat.

2 Výběrová entropie

Mějme abecedu $X = \{x_a, x_b, ..., x_m\}$ - soustavu elementárních jevů a n nezávislých pozorování $x_1, x_2, ..., x_n$ - "produkce" zdroje s touto abecedou. Nechť je takový zdroj stacionární s pravděpodobnostmi výskytu jednotlivých písmen $p(x)$. Toto pravděpodobnostní rozdělení je nám nedostupné a je modelováno námi předpokládaným rozdělením $e(x)$. Potom budeme za výběrovou entropii $H_n(X;e)$ považovat výraz:

$$H_n(X;e) = -\frac{1}{n}\sum_{i=1}^{n} \lg\Big(e(x_i)\Big).$$

Ten lze jinak psát:

$$H_n(X;e) = -\sum_{x\in X}\frac{n(x)}{n}\lg\Big(e(x)\Big),$$

kde $n(x)$ je počet pozorování písmene (jevu) x. S rostoucím n výběrová entropie $H_n(X;e)$ konverguje k výrazu $-\sum_{x\in X} p(x)\lg(e(x))$. Konvergence je ve smyslu pravděpodobnosti [4] (zákon velkých čísel). „Limitní výběrovou entropií“ budeme pak rozumět:

$$H(X;e) = -\sum_{x\in X} p(x)\lg(e(x)).$$

Tento pojem není originální. Je znám pod názvem „Cross Entropy“, např. [7]. Nejprve vlastnosti:
1. $H(X;e)$ je konvexním funkcionálem $e()$.
2. $H(X;e)$ je spojitým funkcionálem $e(), \forall e()$, pro které je $e(x) > 0, \forall x \in X$.
3. $H(X;e)$ nabývá svého minima pro $e() = p()$, pokud $p(x) > 0, \forall x \in X$.

Vlastnost 1. je zřejmá, je důsledkem konvexity funkce $-\lg()$. Vlastnost 2. dokážeme pomocí konceptu sub-derivace ve směru.

Rozdělení $e_{\lambda,g}(x) = (1-\lambda)e(x) + \lambda g(x)$ budeme pro $0 \le \lambda \le 1$ nazývat (λ) variací rozdělení $e()$ ve „směru“ $g()$.

Sub-derivací $H(X;e)$ ve směru g budeme rozumět výraz [2,3]:

$$H'_g(X;e) = \lim_{\lambda\to 0_+}\frac{H(X;e_{\lambda,g}) - H(X;e)}{\lambda}.$$

Potom:

$$\begin{aligned} H'_g(X;e) &= \lim_{\lambda\to 0_+}\frac{d}{d\lambda}H(X;e_\lambda) = \lim_{\lambda\to 0_+}\left\{-\sum_{x\in X}p(x)\frac{g(x)-e(x)}{e_{\lambda,g}(x)}\right\} = \\ &= -\sum_{x\in X}p(x)\frac{g(x)-e(x)}{e(x)}. \end{aligned}$$

Derivace je definována pro všechny modely $e()$, pro které je $e(x) > 0, \forall x \in X$. Tedy je i funkcionál $H(X;e)$ pro takové modely spojitý v $e()$ [2,3]. Vlastnost 3. je důsledkem obou předchozích a faktu, že: $H'_g(X;p) = 0$ pro taková $p()$, pro které je $p(x) > 0, \forall x \in X$. Požadavek nenulovosti je, u diskrétní verze, přirozený. Mohou s ním však být problémy u odhadů při nulovém počtu pozorování jevu (písmene), u kterého je předpoklad nenulové pravděpodobnosti (zero frequency problem). Minimalita $H(X;e)$ pro $e() = p()$ vede na následující heuristiku při odhadování:

$$\text{Hledáme takové } \mathrm{e}^*(), \text{ pro které je } \mathrm{H_n(X;e^*)} = -\frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{lg}\Big(\mathrm{e}^*(\mathrm{x_i})\Big) \text{ minimální.}$$

To však není nic jiného než metoda maximální věrohodnosti, protože, pokud takové $e^*()$ existuje, pak:

$$-\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \lg\Big(e^*(x_i)\Big) \le -\tfrac{1}{n}\sum_{i=1}^{n} \lg\Big(e(x_i)\Big) \Leftrightarrow$$
$$\Leftrightarrow \textstyle\sum_{i=1}^{n} \lg\Big(e^*(x_i)\Big) \ge \sum_{i=1}^{n} \lg\Big(e(x_i)\Big) \Leftrightarrow \prod_{i=1}^{n} e^*(x_i) \ge \prod_{i=1}^{n} e(x_i).$$

A to je velice užitečný prostředek, zvláště pro parametrické odhadování. Dokonce lze i pomocí klasického diferenciálního počtu nalézt (v případě konečného souboru elementárních jevů) odhad $\hat{e}(x)$ minimalizující $H_n(X;e)$. Tedy hledáme čísla $\hat{e}(x_a), \hat{e}(x_b), ..., \hat{e}(x_m)$, pro která platí: $\sum_{x\in X} \hat{e}(x) = 1$ a $H_n(X;\hat{e}) \to \min$. Potom s použitím metody Lagrangeových multiplikátorů se jedná o nalezení minima:

$$Q\{\hat{e}(x_a), \hat{e}(x_b), ..., \hat{e}(x_m), \psi\} = -\sum_{x\in X} \frac{n(x)}{n} \lg\Big(\hat{e}(x)\Big) + \psi\Big(1 - \sum_{x\in X} \hat{e}(x)\Big),$$

pak

$$\frac{\partial Q}{\partial \hat{e}(x)} = -\frac{n(x)}{n}\frac{1}{\hat{e}(x)} - \psi \quad \text{a} \quad \frac{\partial^2 Q}{\partial \hat{e}^2(x)} = \frac{n(x)}{n}\frac{1}{\hat{e}^2(x)} \ge 0, \quad \text{pokud} \quad \hat{e}(x) > 0.$$

Řešením rovnic:

$$\frac{\partial Q}{\partial \hat{e}(x)} = 0 \;\text{ a }\; \sum_{x\in X} \hat{e}(x) = 1 \;\text{ dostaneme }\; \hat{e}(x) = \frac{n(x)}{n}, \quad \text{pokud} \quad n(x) > 0.$$

A to není nic jiného než klasický frekvenční odhad pravděpodobnosti za předpokladu, že pro všechny možné jevy (písmena) máme alespoň jedno pozorování.

3 Výběrová divergence a informace ve speciálních případech

Divergence pravděpodobnostních modelů je definována [1,2]:

$$D(p\|s) = \sum_{x\in X} p(x) \lg \frac{p(x)}{s(x)}.$$

Z obdobných důvodů jako výše, můžeme zavést:

$$D_n(p\|s;e) = \sum_{x\in X} \frac{n(x)}{n} \lg \frac{e(x)}{s(x)}$$

výběrovou divergenci modelu $e()$ proti modelu $s()$ a její limitní formu:

$$D(p\|s;e) = \sum_{x\in X} p(x) \lg \frac{e(x)}{s(x)}.$$

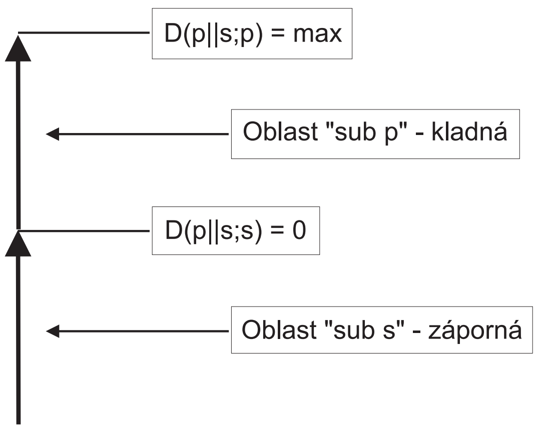
Za použití shodných prostředků jako u entropie lze dokázat obdobné vlastnosti:
1. $D(p\|s;e)$ je konkávním funkcionálem $e()$.
2. $D(p\|s;e)$ je spojitým funkcionálem $e(), \forall e()$, pro které je $e(x) > 0, \forall x \in X$.
3. $D(p\|s;e)$ nabývá svého maxima pro $e() = p()$, pokud $p(x) > 0, \forall x \in X$.
4. $D(p\|s;s) = 0$.
Vlastnosti 1. a 4. jsou triviální. Vlastnosti 2. a 3. jsou dokazatelné opět za použití sub-derivace ve směru:

$$D'_g(p\|s;e) = \lim_{\lambda\to 0_+} \frac{D(p\|s;e_{\lambda,g}) - D(p\|s;e)}{\lambda} = \sum_{x\in X} p(x) \frac{g(x) - e(x)}{e(x)}.$$

Ta je až na znaménko shodná se sub-derivací u limitní výběrové entropie. To je přirozené, neboť:

$$\begin{aligned} D(p\|s;e) &= \sum_{x\in X} p(x) \lg \frac{e(x)}{s(x)} = \sum_{x\in X} p(x) \lg \Big(e(x)\Big) - \sum_{x\in X} p(x) \lg \Big(s(x)\Big) = \\ &= -H(X;e) - \sum_{x\in X} p(x) \lg \Big(s(x)\Big). \end{aligned}$$

Jak bylo uvedeno v práci [8], uvedené vlastnosti umožňují klasifikovat možné modely $e()$ ve vztahu ke srovnávacímu $s()$ podle následující stupnice:



Samozřejmě se též nabízí problém nalezení „nejvíc odlišného“ modelu od $s()$. Je však namístě poznamenat, že divergence není metrikou.

To můžeme naznačit následující úlohou: Hledáme sdružené rozdělení pravděpodobnosti $p(x,y)$ za předpokladu, že jsou dány a známy jeho marginály $p(x)$ a $p(y)$, a to takové, že „maximalizuje“ divergenci „od“ sdruženého rozdělení za předpokladu nezávislosti $p(x)p(y)$. Tedy:

$$\sum_{x\in X}\sum_{y\in Y}\frac{n(x,y)}{n}\lg\frac{p(x,y)}{p(x)p(y)} \xrightarrow{p(x,y)} \max$$

při vazbách:

$$p(x)=\sum_{y\in Y}p(x,y);\forall x\in X \qquad \text{a} \qquad \mathrm{p(y)}=\sum_{\mathrm{x\in X}}\mathrm{p(x,y);\forall y\in Y}.$$

Řešení tohoto problému klasickou cestou vede opět na využití Lagrangeových multiplikátorů:

$$Q(p(,),\alpha,\psi)=\sum_{x\in X}\sum_{y\in Y}\frac{n(x,y)}{n}\lg\frac{p(x,y)}{p(x)p(y)}+$$

$$+\sum_{x\in X}\alpha_x\Big(p(x)-\sum_{y\in Y}p(x,y)\Big)+\sum_{y\in Y}\psi_y\Big(p(y)-\sum_{x\in X}p(x,y)\Big).$$

Z toho dostaneme rovnici „sedlových bodů“:

$$0=\frac{\partial Q}{\partial p(x,y)}=\frac{n(x,y)}{np(x,y)}-\alpha_x-\psi_y$$

a pro matici druhých parciálních derivací (přesněji její diagonálu):

$$\frac{\partial^2 Q}{\partial p^2(x,y)}=-\frac{n(x,y)}{np^2(x,y)}\,,$$

tedy v případných sedlových bodech bude maximum. Řešením rovnice sedlových bodů dostaneme:

$$p(x,y)=\frac{n(x,y)}{n(\alpha_x+\psi_y)}\,.$$

Hodnoty α_x,ψ_y zjistíme numerickým řešením vazebních rovnic:

$$p(x)=\frac{1}{n}\sum_{y\in Y}\frac{n(x,y)}{(\alpha_x+\psi_y)},\forall x\in X;\quad p(y)=\frac{1}{n}\sum_{x\in X}\frac{n(x,y)}{(\alpha_x+\psi_y)},\forall y\in Y,$$

pokud takové řešení existuje.

Uvedenou úlohu lze modifikovat, např.:

$$\sum_{x\in X}\sum_{y\in Y}\frac{n(x,y)}{n}\lg\frac{p(x,y)}{p(x)p(y)}\xrightarrow{p(x,y)}\max,$$

$$\text{při vazbě}: \ \mathrm{p(x)} = \sum_{\mathrm{y}\in \mathrm{Y}} \mathrm{p(x,y)}, \forall \mathrm{x} \in \mathrm{X}.$$

Tj. příliš nám nezáleží na splnění podmínek:

$$p(y) = \sum_{x\in X} p(x,y), \forall y \in Y.$$

$$\text{Pak dostaneme řešení pro odhad } \mathrm{p(x,y)} \text{ ve tvaru}: \ \mathrm{p(x,y)} = \frac{\mathrm{n(x,y)}}{\mathrm{n(x)}}\mathrm{p(x)}$$

$$\text{a odpovídající marginální pravděpodobnost } \bar{\mathrm{p}}(\mathrm{y}) = \sum_{\mathrm{x}\in\mathrm{X}}\frac{\mathrm{n(x,y)}}{\mathrm{n(x)}}\mathrm{p(x)}.$$

Tato řešení jsou vlastně klasickým (frekvenčním) odhadem podmíněné pravděpodobnosti $p(y/x)$.

Ještě volnější modifikace je:

$$\sum_{x\in X}\sum_{y\in Y}\frac{n(x,y)}{n}\lg\frac{p(x,y)}{p(x)p(y)}\xrightarrow{p(x,y)}\max.$$

Tj. vůbec nám nezáleží na dodržení marginálů. Pak dostaneme řešení pro odhad $p(x,y)$ ve tvaru:

$$p(x,y) = \frac{n(x,y)}{n}$$

a tomu odpovídající marginální pravděpodobnosti:

$$\bar{p}(x) = \frac{n(x)}{n}; \qquad \bar{p}(y) = \frac{n(y)}{n}.$$

A to je vlastně „nejklasičtější“ frekvenční tvar odhadu.

4 Závěr

Předložená práce má ambici prokázat, že nastíněné pojetí „výběrových“ entropií, divergencí a informací spolu s jejich limitními tvary je konzistentní s některými klasickými technikami odhadování. Nejen to, ukazuje se, že je v něm možná skryt další, vlastní, potenciál pro rozvoj. Nakolik je tento dojem přesvědčivý i mimo skupinu autorů, ponecháváme na čtenáři.

Reference

[1] Vajda, I.: *Teória informácie a štatistického rozohodovania.* Alfa, Bratislava, 1982.

[2] Cover, T.M., Thomas, J.A.: *Elements of Information Theory.* Wiley, 1991.

[3] Hiriart-Urruty, J.B., Lemaréchal, C.: *Fundamentals of Convex Analysis.* Springer, 2001.

[4] Rényi, A.: *Teorie pravděpodobnosti.* Academia, Praha, 1972.

[5] Cziszár, I.: *Information Theoretic Methods in Probability and Statistics.* Conference Fifty Years of Shannon Theory, 1998 and Shannon Lecture ISIT'97.

[6] Cziszár, I., Körner, J.: *Information Theory, Coding Theorems for Discrete Memoryless Systems.* KIADÓ, BUDAPEST, 1986.

[7] Zhai, C.X.: *Essential Probability and Statistics.* Department of Computer Science, University of Illinois, Urbana-Champaign, Lecture for CS397-CXZ Algorithms in Bioinformatics, 2004.

[8] Vávra, F., Nový, P.: *Informace a dezinformace. Seminář z aplikované matematiky.* Katedra aplikované matematiky, Přírodovědecká fakulta Masarykovy Univerzity, Brno, 2004.

Poděkování: ??????????????????????

Adresa: Západočeská univerzita v Plzni, Fakulta aplikovaných věd, Katedra informatiky a výpočetní techniky, Univerzitní 8, 306 14 Plzeň.

E-mail: `vavra@kiv.zcu.cz`