ZÁKLADY GNOSTICKÉ TEORIE DAT

P. Kovanic

SOUHRN

Gnostická teorie dat je pokusem o matematický model poznávání kvantitativních vlastností světa za působení neurčitosti, o které není známo, zda vyhovuje statistickým předpokladům. Z jediného axiomu se dokazují významné zákonitosti jednotlivých dat, z nichž lze za použití druhého axiomu odvodit vlastnosti datových souborů a jejich charakteristiky použitelné k robustnímu odhadování polohy, měřítka i distribučních funkcí datových souborů, k testování homogenity souborů apod.

MATEMATICKÁ DATA VERSUS REÁLNÁ DATA

Statistická teorie je vybudována pro data generovaná myšlenými matematickostatistickými modely, nebo pro data podle těchto modelů vypočtená, či fysikálně sestrojená tak, aby vlastnosti dat dobře aproximovaly vlastnosti matematicky definovaných dat. Praktická použitelnost statistických metod je pak podmíněna shodou vlastností skutečných dat s teoretickými. Takovou shodu často nelze zaručit, dokonce někdy ani rozpoznat pro nedostatečné množství dat, proměnnost situací, nestacionárnost atp. Nemůžeme-li ale přisoudit datům statistické vlastnosti a přesto potřebujeme vybudovat teorii vedoucí k metodám použitelným pro zpracování dat, musíme najít jejich jinou důležitou vlastnost, z níž lze teorii odvodit: Reálná data přijímaná za základ gnostické teorie jsou výsledky buď měření nebo čítání reálných kvantit. Z teorie měření vyplývá, že tato data jsou striktně kladná a konečná. V praxi se setkáváme i s daty, která byla získána dodatečnou transformací výsledků měření nebo čítání, mají např. posunutou nulu ap. Ta musí být před dosazením do gnostických vzorců transformována do požadované primární formy.

GNOSTICKÁ TEORIE JEDNOTLIVÝCH DAT

První axiom a jeho důsledky pro teorii kvantifikace

Shrňme stručně gnostickou teorii jednotlivých dat dle /1/, kde jsou uvedeny přesnější formulace tvrzení i s důkazy. Nechť "z" je skutečný výsledek kvantifikace, tj. měření nebo čítání, vyjádřený číslem a nechť "z_o" je (ideální) hodnota, kterou bychom dostali jako výsledek kvantifikace, kdyby byl proces kvantifikace zbaven jakékoliv neurčitosti. První axiom gnostické teorie pak charakterisuje model reálných dat takto:

$$z = z_o \xi \qquad (z, z_o, \xi \in R_+) \qquad /1/$$

Snadno zjistíme, že množina transformací zobrazujících jakoukoliv datovou položku tvaru /1/ na sebe či na jinou položku dat, lišící se pouze parametrem ξ charakterisujícím vliv neurčitosti tvoří komutativní grupu, jejíž strukturní operací je násobení operátorů . Veličina "z" dle /1/ je však bilineární funkcí proměnných z_o a ξ. Záměna proměnných

$$x = z_o \mathrm{ch}\Omega \qquad y = z_o \mathrm{sh}\Omega \qquad /\text{kde } \Omega = \ln \xi / \qquad /2/$$

pak dává možnost přiřadit každé položce dat (z) typu /1/ dvourozměrový vektor $\underline{u}^T := (x,y)^T$ a také ${}_o\underline{u}^T := (y,x)^T$, přičemž $z = x + y$. Pro neznalost z_o ani ξ nedovedeme takový rozklad dat uskutečnit, ale jeho teoretický tvar nám dovoluje odvozovat důležité závěry. Především zjišťujeme, že množina matic $\underline{K}$ zobrazujících vektory typu $\underline{u}$ na sebe nebo na jiné /$\underline{u}'$/, tvoří komutativní grupu, jejíž strukturní operací je běžné maticové násobení. Tato grupa je isomorfní jednak s grupou operátorů ξ na R_+ a jednak s grupou Lorentzových transformací na dvourozměrné varietě vektorů. Lze dokázat, že metriku na této varietě nemůžeme libovolně zadat, ta je již určena, neboť jediná metrika invariantní k Lorentzovým transformacím je pseudoeuklidovská metrika. Jako důsledek prvního axiomu tedy dostáváme, že vliv kvantifikační neurčitosti na reálná data lze matematicky modelovat jako ortogonální rotaci vektorů v Minkowkého rovině, při níž se body representující reálná data pohybují po pseudoeuklidovské kružnici, tj. po euklidovské hyperbole. Relativní délka dráhy tohoto pohybu může tedy být snadno vyčíslena, je rovna veličině $|\Omega|$ o níž se snadno přesvědčíme, že vyhovuje jednoduchému variačnímu principu: je to maximum z možných relativních délek drah pohybů mezi zadanými koncovými body po různých křivkách. V tomto smyslu tedy lze hovořit o tom, že vliv kvantifikační neurčitosti maximalizuje znehodnocení reálných dat.

Teorie estimačních transformací a ideální gnostický cyklus

Nás ale zajímá i další transformace dat, estimační. Ta má výsledku kvantifikace (z) přiřadit odhad $\tilde{z}_o$. V dvourozměrné representaci je stanovení estimační transformace ekvivalentní nalezení dráhy po níž se má pohybovat konec vektoru $\underline{u}$ nebo $_o\underline{u}$, aby se nakonec dostal co nejblíže k bodu $(z_o,0)^T$ nebo $(0,z_o)^T$. Abychom stanovili nejlepší dráhu estimační transformace, studujeme dále kvantifikační transformace Lorentzova typu, o nichž si dokážeme, že splňují podmínky homogenity, dvojité symetrie a rovnoměrné regularity. Matematický zápis těchto vlastností je však hyperbolickou verzí známých podmínek analytičnosti funkce komplexní proměnné. Estimační transformace plně duální ke kvantifikační je proto daná euklidovsky ortogonálním otáčením vektorů v Gaussově rovině. Vyhovuje podmínkám analytičnosti Cauchy-Riemanna a příslušná dráha odpovídá opačnému variačnímu principu: její relativní délka je minimální, duálně k (2) platí pro estimační dráhu

$$x = r\cos\omega \qquad y = -r\sin\omega \qquad /3/$$

kde

$$r = \sqrt{x^2+y^2} \qquad /4/$$

$$\operatorname{tg}\omega = -\operatorname{th}\Omega \qquad /5/$$

Z obou drah, kvantifikační a estimační, lze sestrojit uzavřený cyklus transformací, který nazveme ideálním gnostickým cyklem. Tento cyklus hraje významnou úlohu jako vzor pro postup skutečného procesu odhadování, lze dokázat jeho důležité extremální vlastnosti.

Charakteristiky nepodobnosti gnostických událostí

Gnostickou událostí je skutečnost, že byl získán výsledek kvantifikace tvaru /1/. Za gnostické události můžeme proto přijmout i vektor $\underline{u}$ a $_o\underline{u}$ mající složky x,y tvaru /2/ nebo /3/, jejichž součet dá datovou položku tvaru /1/. V předchozím odstavci jsme se přesvědčili o plodnosti geometrického přístupu, který dovolil odvodit důležité vlastnosti matematického modelu neurčitosti z úvah o metrice. Zůstaňme geometrickému přístupu věrni při posuzování vztahů mezi dvěma událostmi s využitím pojmu podobnosti, který však oproti běžnému zobecníme tak, abychom vyčerpali všechny aspekty podobnosti. Dvě gnostické události $\underline{u}$ a $\underline{u}'$ mající složky x,y a x',y' budeme považovat za σ,ε-podobné, jestliže platí

$$x/y = \sigma\cdot(x'/y')^{\varepsilon} \qquad /\text{kde } \sigma = \pm 1 \text{ a } \varepsilon = \pm 1/ \qquad /6/$$

Za číselnou charakteristiku σ,ε-nepodobnosti pak přijmeme rozdíl obou těchto poměrů. Dosazením z /2/ se pak přesvědčíme, že charakteristikami nepodobnosti v případě kvantifikace jsou hyperbolické siny a cosiny součtů a rozdílů parametrů neurčitosti Ω obou událostí. S použitím /3/ dostaneme pro estimaci další čtyři charakteristiky nepodobnosti, obyčejné siny a cosiny součtů a rozdílů parametrů ω spjatých s parametry Ω vztahem /5/. Obvyklá "trojúhelníková" podobnost odpovídající hodnotám $\sigma=1$ a $\varepsilon=1$ nás na tomto místě zajímá nejméně, protože pouze ostatní tři kombinace hodnot σ a ε umožňují kvantitativně charakterisovat kvalitu gnostické události novým způsobem, charakteristikami "nepodobnosti sobě". Pro $\underline{u}'=\underline{u}$ jsou totiž numerickými charakteristikami nepodobnosti mezi $\underline{u}$ a $\underline{u}'$ (tedy nepodobnosti mezi $\underline{u}$ a $\underline{u}$) hyperbolické siny a cosiny dvojnásobného argumentu Ω a obyčejné siny a cosiny dvojnásobného argumentu ω. Zapišme explicitní výrazy pro tyto charakteristiky i snadno odvoditelné vztahy mezi nimi vyplývající z /5/:

$$f := \cos 2\omega = 1/\operatorname{ch} 2\Omega = (x^2-y^2)/(x^2+y^2) = 2/(\xi^2+\xi^{-2}) \qquad /7/$$

$$h_q := \operatorname{sh} 2\Omega = -\operatorname{tg} 2\omega = 2xy/(x^2-y^2) = (\xi^2-\xi^{-2})/2 = -h_e/f \qquad /8/$$

$$h_e := \sin 2\omega = -\operatorname{th} 2\Omega = -2xy/(x^2+y^2) = -(\xi^2-\xi^{-2})/(\xi^2+\xi^{-2}) = -h_q f \qquad /9/$$

kde

$$\xi := z/z_e = (x+y)/z_e \qquad /10/$$

V předcházejícím odstavci jsme se již setkali s maticemi $\underline{K}$ typu

$$\underline{K}_q(\Omega) = \begin{pmatrix} ch\Omega & sh\Omega \\ sh\Omega & ch\Omega \end{pmatrix} \qquad /11/$$

zobrazujícími jednu gnostickou událost tvaru /2/ na jinou. Taková matice je prvkem komutativní grupy isomorfní s Lorentzovou grupou. Duální k /11/ je estimační operátor

$$\underline{K}_e(\omega) = \begin{pmatrix} \cos\omega & -\sin\omega \\ \sin\omega & \cos\omega \end{pmatrix} \qquad /12/$$

který je také prvkem komutativní grupy. Ta však již isomorfní s Lorentzovou grupou není: Součiny matic z obou grup již jednoznačně spjaty nejsou i když prvky obou grup jsou vzájemně jednoznačně přiřazeny vztahem /5/. Uvedené veličiny a vztahy nám postačují k vyslovení (a podle /1/ i k důkazům) těchto důležitých tvrzení:

1) Vliv neurčitosti na data lze modelovat jako ortogonální otáčení vektoru v rovině (operátorem $\underline{K}_q(\Omega)$ nad Minkowského rovinou nebo operátorem $\underline{K}_e(\omega)$ nad Euklidovou rovinou).
2) Veličiny $\underline{K}_q^2(\Omega) = \underline{K}_q(2\Omega)$ a $\underline{K}_e^2(\omega) = \underline{K}_e(2\omega)$ se při takových transformacích chovají jako <u>tensory</u>.
3) Veličiny f, f^{-1}, h_e a h_q /7/-/9/, číselné charakteristiky nepodobnosti gnostických událostí sobě se shodují se složkami těchto tenzorů.

Platí tedy identity

$$\underline{K}_q^2(\Omega) = \begin{pmatrix} f^{-1} & h_q \\ h_q & f^{-1} \end{pmatrix} \qquad \underline{K}_e^2(\omega) = \begin{pmatrix} f & -h_e \\ h_e & f \end{pmatrix} \qquad /13/$$

a my máme důvod nazvat tyto veličiny <u>tensory nepodobnosti</u>. Vidíme, že charakterizují nekvalitu dat způsobenou neurčitostí.

<u>Změny informace jednotlivé datové položky způsobené neurčitostí</u>

Složky tensorů nepodobnosti dále hrají významnou úlohu, stojí proto za to je pojmenovat:

f^{-1} ..... <u>nevěrnost</u>
f ..... <u>věrnost</u>
h_q ..... <u>kvantifikační irrelevance</u>
h_e ..... <u>estimační irrelevance</u>

Při kvantifikaci se kvantifikační irrelevance mění od 0 do hodnoty h_q. Vypočtěme střední hodnotu úhlu 2ω na tomto intervalu, veličinu

$$I_q = \int_0^{h_q} 2\omega \, dh_q \qquad /14/$$

Hodnotě h_q dosažené na konci kvantifikace odpovídá hodnota estimační irrelevance h_e počátku estimace. Při estimaci se pak tato irrelevance bude měnit od h_e do nuly. Vypočtěme střední hodnotu úhlu 2Ω na tomto intervalu:

$$I_e = \int_{h_e}^{0} 2\Omega \, dh_e \qquad /15/$$

Použijeme-li pro nějaké komplexní p $(p\neq 0,\ p\neq 1)$ funkci

$$H\,p = -p \ln p - (1-p)\ln(1-p) \qquad /16/$$

kde pod $\ln(.)$ rozumíme hlavní hodnotu a zavedeme-li označení

$$p_q = (1+\sqrt{-1}\,h_q)/2 \qquad p_e = (1+h_e)/2 \qquad /17/$$

dostaneme integrály /14/ a /15/ ve tvaru

$$I_q = H\,1/2 - H\,p_q \qquad /18/$$
$$I_e = H\,1/2 - H\,p_e \qquad /19/$$

Veličiny I_q a I_e interpretujeme jako <u>změny informace</u> v kvantifikační a estimační části ideálního gnostického cyklu. K takové interpretaci je několik závažných důvodů přesto, že jsme tyto veličiny odvodili pro jedinou položku dat a není dán žádný obvyklý pravděpodobnostní model:

1) Obě veličiny I_q i I_e jsme odvodili jako rozdíly funkce H mající formu shodnou se statistickou entropií termodynamického systému a tudíž i s Shannonovou informací binárního pravděpodobnostního systému.
2) Při vzdalování veličiny dat z od ideální hodnoty I_q klesá od nuly monotonně a při návratu veličiny z od zašené (pozorované) hodnoty k ideální (tento návrat se snažíme zajistit estimací) veličina I roste monotonně od nuly. Pro každou pozoro-

vanou hodnotu $z_i \neq z_o$ však platí

$$I_q(z_i/z_o)+I_e(z_i/z_o)<0 \qquad (0<z_i/z_o<\infty) \qquad /20/$$

což interpretujeme jako nemožnost "napravit" estimací zcela ztrátu informace způsobenou vlivem neurčitosti, a to ani v případě ideálního gnostického cyklu.

3) Parametry funkcí H podobně jako pravděpodobnost v případě Shannonovy informace jsou veličiny p_q a p_e použitelné jako číselné charakteristiky očekávání, že neznámá veličina z_o má hodnotu převyšující posorovanou hodnotu.

4) Zřídla polí veličin I_q a I_e jsou úměrná zřídlům polí změn termodynamické entropie. V gnostickém cyklu tedy dochází ke změnám informace na úkor změn entropie a naopak. Hypotésa o existenci konverse entropie na informaci a naopak byla vyslovena velice dávno, zde máme matematický model této konverse pro obecný případ neurčitosti každé jednotlivé datové položky.

Po přijetí informační interpretace funkcí I_q a I_e lze přistoupit k formulaci významného tvrzení o optimalitě ideálního gnostického cyklu.

Ideální gnostický cyklus a jeho informační optimalita

Ideální hodnotě z_o můžeme přiřadit dva vzájemně duální vektory $\underline{u}_o^T = (z_o,0)$ a ${}_c\underline{u}_o^T = (0,z_o)$. Byl-li výsledek skutečné kvantifikace z_s, jsou příslušné gnostické události v dvourozměrovém tvaru $\underline{u}_s^T = (x_s,y_s)$ a ${}_c\underline{u}_s^T = (y_x,x_s)$, přičemž složky $x_s y_s$ mají tvar /2/ při kvantifikaci s úhlem Ω_s a při estimaci tvar /3/ s úhlem ω_s a amplitudou $r_s = \sqrt{x_s^2+y_s^2}$. Kvantifikační část dráhy ideálního gnostického cyklu (IGC) představuje oblouk Minkowského kružnice vycházející z bodu $\underline{u}_o$ a končící v bodě $\underline{u}_s$ (z hlediska euklidovského zobrazení je to oblouk rovnoosé hyperboly). Z bodu $\underline{u}_s$ pak IGC pokračuje podél oblouku obyčejné kružnice o poloměru r_s až do bodu $\underline{u}_s'^T = (r_s,0)$ a do původního výchozího bodu $\underline{u}_o$ se cyklus uzavírá pohybem představujícího bodu podél úsečky $\underline{u}_s'\underline{u}_o$. Ke každému takovému IGC graficky znázorněnému jako

$$\mathrm{IGC} := \widehat{\underline{u}_o\underline{u}_s} \,\|\, \widehat{\underline{u}_s\underline{u}_s'} \,\|\, \overline{\underline{u}_s'\underline{u}_o} \qquad /21/$$

existuje ovšem duální cyklus

$${}_c\mathrm{IGC} := \widehat{{}_c\underline{u}_o\,{}_c\underline{u}_s} \,\|\, \widehat{{}_c\underline{u}_s\,{}_c\underline{u}_s'} \,\|\, \overline{{}_c\underline{u}_s'\,{}_c\underline{u}_o} \qquad /22/$$

který při zobrazení v rovině (x,y) je s IGC symetrický vzhledem k ose kvadrantu. Třetí, uzavírající přímkovou část cyklů nazýváme atenuací, protože pohyb podél přímky procházející počátkem souřadnic lze představit jako důsledek změny měřítka či "zeslabení" veličin nějakým atenuátorem. Cyklus je tedy definován jako posloupnost tří na sebe navazujících myšlených pohybů v rovině:

IGC:=Kvantifikace‖Estimace‖Atenuace

probíhajících po již stanovených drahách. Představme si však jiný uzavřený cyklus VGC vzniklý z IGC malou variací při zafixovaných koncových bodech z_o a z_s. Pro každý takový cyklus je v /1/ dokázáno, že platí

$$\oint_{\mathrm{VGC}} dI \leq \oint_{\mathrm{IGC}} dI < 0 \qquad /23/$$

kde I značí postupně při vyčíslení integrálu I_q, I_e a I_a, přičemž I_a je změna informace při atenuaci (ta je nulová, protože změna měřítka nemění poměr y(x)). Analogická nerovnost platí i pro duální cykly ${}_c\mathrm{IGC}$ a ${}_c\mathrm{VGC}$. Odtud plyne, že IGC je optimální v tom smyslu, že minimalizuje ztráty informace způsobené neurčitostí. V tomto smyslu je IGC vzorem pro praktické odhadování, k němuž se budeme snažit co nejvíce přiblížit.

GNOSTICKÁ TEORIE DATOVÉHO SOUBORU

Charakteristiky datového souboru by měly representovat společné vlastnosti jednotlivých dat. Mají být určeny tak, aby jejich neurčitost byla menší než "průměrná" neurčitost jednotlivých dat. K tomu je třeba data složit podle nějakého kompozičního zákona. Prosté sečítání dat běžné při vytváření průměrů se neosvědčilo v případech, kdy se požaduje robustnost odhadu vůči odlehlým pozorováním. V robustní statistice dostává každá položka dat při vytváření odhadu parametru polohy či měřítka svou individuální váhu závislou na vzdálenosti této položky od hodnoty parametru polohy. V gnostické teorii /2/ je taková individuální váha určena druhým axiomem této teorie, který stanovuje kompoziční zákon pro skládání dat.

Druhý axiom gnostické teorie

Nechť $Z(z_o,n)$ je datový soubor sestavený z dat $z_1,\dots,z_n$ majících tvar /1/. Charakteristikou souboru Z nazveme funkci všech těchto dat a z_o. Za charakteristickou událost souboru Z přijmeme událost $\underline{u}_c^T = (z_o \operatorname{ch}\Omega_c,\ z_o \operatorname{sh}\Omega_c)$ v případě kvantifikace a $\underline{u}_c^T = (r_c\cos\omega_c, -r_c\sin\omega_c)$ v případě estimace, přičemž veličiny Ω_c, ω_c jsou charakteristiky souboru Z.

Druhý axiom gnostické teorie říká, jak určit charakteristiky Ω_c a ω_c:

$$\underline{K}_q^2(\Omega_c) = \frac{1}{w_q^2}\sum_i^n \underline{K}_q^2(\Omega_i) \qquad \underline{K}_e^2(\omega_c) = \frac{1}{w_e^2}\sum_i^n \underline{K}_e^2(\omega_i) \qquad /24/$$

kde

$$w_q^2 = \operatorname{Det}\Big(\sum_i^n \underline{K}_q^2(\Omega_i)\Big) \qquad w_e^2 = \operatorname{Det}\Big(\sum_i^n \underline{K}_e^2(\omega_i)\Big) \qquad /25/$$

jsou normalizační váhy. Gnostický kompoziční zákon tedy skládá vlivy jednotlivých dat souboru silně nelineárně tak, aby tenzory nepodobnosti charakteristické události souboru byly normovanými součty tenzorů nepodobnosti jednotlivých dat. Tenzorový charakter sčítaných veličin zajišťuje formální správnost takového skládání, dále však uvidíme, že pro právě tento způsob skládání je i závažný přímý důvod - zajištění konsistence gnostické teorie s relativistickou fyzikou. Takový výběr axiomu skládání dat se ovšem podstatně liší od často heuristických a formálních důvodů vedoucích k pravidlům pro vážení dat přijímaným v mnoha desítkách metod robustní statistiky.

Důsledky druhého axiomu

Ze vztahů /17/ snadno získáme charakteristiky Ω_c a ω_c:

$$2\Omega_c = \operatorname{arcth}(\bar{h}_q/\overline{f^{-1}}) \qquad 2\omega_c = \operatorname{arctg}(\bar{h}_e/\bar{f}) \qquad /26/$$

přičemž symbol $\bar{q}$ značí zde i dále aritmetický průměr hodnot $q_1,\dots,q_n$. Veličina Ω_c má povahu charakteristické chyby dat hodnocené z hlediska kvantifikačního procesu, zatímco ω_c odpovídá chybě dat měřené v estimační metrice. Lze ukázat, že při slabé neurčitosti dat (malé relativní chyby všech dat souboru) se všechny charakteristiky Ω_c, $-\omega_c$, $\bar{h}_q$ i $\bar{h}_e$ liší od průměrné relativní chyby dat pouze o veličinu $O(\varepsilon^2)$, kde ε je největší absolutní hodnota relativní chyby dat. Uvedené čtyři veličiny tedy představují různá gnostická zobecnění relativních chyb dat platná i pro případy silné neurčitosti. Jejich užitečnost je v tom, že se liší stupněm robustnosti vůči odlehlým pozorováním.

Ve vzorcích /19/ vystupují i dvě další důležité veličiny, průměrná nevěrnost $\overline{f^{-1}}$ a průměrná věrnost $\bar{f}$. Při výpočtu normalizačních vah podle /18/ se čtverce těchto veličin mohou rozepsat pomocí veličin $\overline{f^{-2}}$ a $\overline{f^2}$ a pomocí průměrů součinů irrelevancí. Poslední z těchto veličin nabízí možnost zobecnění korelačních koeficientů, zatímco odchylky prvních čtyř veličin od 1 představují gnostická zobecnění výběrových rozptylů. Lze se přesvědčit, že všechny čtyři veličiny $(\overline{f^{-1}}-1)/2$, $(1-\bar{f})/2$, $(\overline{f^{-2}}-1)/4$ a $(1-\overline{f^2})/4$ se při omezení všech relativních chyb hodnotou ε liší od relativní hodnoty výběrového rozptylu o veličinu $O(\varepsilon^3)$. Všechny čtyři veličiny tedy jsou gnostickými zobecněními výběrového rozptylu. Lze ukázat, že se vzájemně liší stupněm robustnosti vůči odlehlým pozorováním.

Gnostická teorie datových souborů nám tedy nabízí řadu charakteristik různé robustnosti. Závislost těchto veličin na neznámé ideální hodnotě z_o umožňuje formulovat a řešit problém odhadování ideální hodnoty jako úlohu extremalizace vybrané charakteristiky. V dalším příspěvku /4/ je uveden souhrnné řešení takového problému.

O KONSISTENCI GNOSTICKÉ TEORIE

Za konsistentní považujeme teorii odvozenou pomocí nějakého jazyka či kalkulu formálně správně z nerosporného a úplného systému axiomů. Takovou vlastnost "vnitřní" či vertikální konsistence zatím můžeme gnostické teorii přisoudit (dokud nebude dokázán opak). U teorie, která má poskytovat model nějakého výseku skutečnosti je však důležitá i jiná, "vnější" či "horizontální" konsistence: vzhledem k "nyní" nepochybným teoriím "sousedních" výseků skutečnosti. Lze ukázat /3/, že gnostická teorie je v souladu s teorií měření, s relativistickou fyzikou a s klasickou termodynamikou a že za určitých zvláštních podmínek, které umožňuje stanovit dává gnostická teorie výsledky konvergující k výsledkům klasické (nerobustní) statistické teorie a teorie informace. Lze rovněž ukázat, že tak rozsáhlou návaznost na jiné teorie existující statistická teorie neprokazuje.

<u>Literatura</u>:

/1/ Kovanic P., Gnostical theory of individual data, v tisku

/2/ Kovanic P., Gnostical theory of small samples of real data, v tisku

/3/ Kovanic P., On relations between information and physics, v tisku (články vyjdou v časopise Problems of control and information theory, č. 4, 5 a 6, r. 1984)

/4/ Kovanic P., Gnostické algoritmy zpracování dat, Sborník ROBUST ´84, MFF-UK (1984)

GNOSTICKÉ ALGORITMY ZPRACOVÁNÍ DAT

Pavel Kovanic

SOUHRN

Příspěvek podává stručnou informaci o současném stavu algoritmizace robustních estimačních metod založených na gnostické teorii dat. Jde zejména o odhadování parametrů polohy jednorozměrových datových souborů, jejich měřítka, distribuční funkce a hustoty pravděpodobnosti. Souhrnně je popsán systém pro interaktivní gnostickou analýzu datových souborů na osobním počítači.

GNOSTICKÉ PARAMETRY POLOHY A MĚŘÍTKA

Gnostická teorie stručně vyložená v jiném příspěvku v tomto sborníku /1/ nepoužívá statistický model. Přesto však nabízí charakteristiky datových souborů blížící se svým smyslem tomu, k čemu statistika používá některé parametry svých modelů, k vystižení polohy a variability datového souboru. Gnostický model dat doplněný oproti /1/ o parametr měřítka má tvar

$$z_i = z_o \exp(s\,\Omega_i) \qquad /1/$$

kde z_i je i-tá položka datového souboru $Z(n, z_o, s)$ obsahujícího n výsledků kvantifikace (tj. měření nebo čítání) kvantity, která by se při ideálních podmínkách kvantifikace zobrazila jako číslo z_o, tzv. ideální hodnota. Veličina Ω_i charakterizuje vliv neurčitosti na i-tou položku dat a veličina "s" měřítka. Především tyto dvě veličiny budeme odhadovat při zpracování datových souborů.

GNOSTICKÁ OPTIMALITA ODHADŮ POLOHY

V předchozím příspěvku /1/ je uvedena dvojice gnostických charakteristik $\bar{E}_q$ a $\bar{E}_e$, které mají smysl zobecněné "střední" chyby dat. Jsou to kvantifikační a estimační irrelevance, složky tensorů nepodobnosti. Tamtéž jsou uvedeny čtyři gnostická zobecnění poměrného výběrového rozptylu určená odchylkami průměrné věrnosti $\bar{f}$, průměrné nevěrnosti $\overline{f^{-1}}$ a průměrů čtverců těchto veličin od jedničky. Za nejlepší gnostický odhad J-tého typu můžeme přijmout veličinu z_J extremalisující některou ze šesti uvedených charakteristik datového souboru $Z(n,z_o,s)$. Každý z takto získaných odhadů má důležitou interpretaci jak je vidět z Tab. 1:

Typ odhadu	Podmínka	Interpretace optimality odhadu
z_{qI}	$d\overline{f^{-2}}/dz_o=0$	Minimalizace intenzity ztráty informace při kvantifikaci
z_{qF}	$d\overline{f^{-1}}/dz_o=0$	Minimální zvýšení termodynamické entropie při kvantifikaci
z_{qS}	$\overline{h}_q=0$	1) Minimalizace celkové kvantifikační ztráty informace 2) Kvantifikační symetrie dat
z_{eS}	$\overline{h}_e=0$	1) Maximalizace celkového estimačního zvýšení informace 2) Estimační symetrie dat
z_{eF}	$d\overline{f}/dz_o=0$	Maximální snížení entropie při estimaci
z_{eI}	$d\overline{f^2}/dz_o=0$	1) Maximalizace intenzity zvyšování informace při estimaci 2) Maximalizace hustoty pravděpodobnosti

Tab. 1: Přehled optimalizačních podmínek pro stanovení gnostických odhadů parametru polohy a jejich teoretická interpretace. Symbol $\overline{q}$ značí aritmetický průměr n-tice veličin $q_1,\ldots,q_n$.

ROVNICE PRO ODHAD GNOSTICKÉHO PARAMETRU POLOHY

Dosazením do podmínek optimality dostáváme pro všech 6 případů z Tab. 1 jedinou rovnici

$$\sum_{i=1}^{n}\left((z_e/z_i)^{2/s}-(z_i/z_e)^{2/s}\right)/\left((z_o/z_i)^{2/s}+(z_i/z_e)^{2/s}\right)^M = 0 \qquad /2/$$

kde exponent M je určen Tab. 2:

Odhad	z_{qI}	z_{qF}	z_{qS}	z_{eS}	z_{eF}	z_{eI}
M	-1	0	0	1	2	3

Tab. 2: Exponent M pro rovnici /1/ odhadu parametru polohy z_o

Důsledkem volby exponentu M je stupeň a druh robustnosti odhadu (blíže viz /2/). Pro M=3 je odhad maximálně robustní a pro M=-1 citlivý k odlehlým pozorováním.

ODHAD MĚŘÍTKA A HUSTOTY DATOVÉHO SOUBORU

Odhad parametru měřítka je spjat s hustotou dat interpretovatelnou jako gnostický odhad hustoty pravděpodobnosti datového souboru. Ten je dán vzorcem

$$\frac{dP}{d} = \frac{1}{n}\sum_{i}^{n} 4/\left((z_i/z)^{2/s}+(z/z_i)^{2/s}\right)^2 \qquad /3/$$

kde P má smysl průměrné pravděpodobnosti, že neznámá veličina z_o je větší než z, podmíněné tím, že data mají hodnoty $z_1,\ldots,z_n$. Známe-li parametr měřítka, můžeme tedy z dat sestrojit odhad hustoty pravděpodobnosti, aniž bychom potřebovali statistický model. Vzorec /3/ můžeme však také použít k získání algoritmu pro odhadování parametru měřítka. Teoretický prostředek k tomu najdeme opět v gnostické teorii: za nejlepší odhad parametru měřítka přijmeme tu jeho hodnotu, která zaručuje shodu mezi průměrnou nepodobností dat a mezi střední nepodobností vypočtenou pomocí odhadnuté hustoty pravděpodobnosti jako funkci parametru měřítka.

SYSTÉM PROGRAMŮ PRO GNOSTICKOU ANALÝZU DAT

V současné době již existuje programová realizace algoritmů vyplývajících z gnostické teorie, umožňujících detailní analýzu jednorozměrových datových souborů. Je vytvořena jako interaktivní systém programů v jazyce BASIC pro osobní počítač SINCLAIR ZX-81 s pamětí 16K. Má následující funkce:

1/ Vstup dat z klávesnice. Data mohou odpovídat buď multiplikativnímu nebo aditivnímu modelu vlivu neurčitosti.

2/ Výběr typu operace na datech

3/ Odhadování parametru polohy pro volitelnou robustnost (parametr M, viz Tab. 2)

4/ Odhadování parametru měřítka s volitelnou robustností

5/ Odhadování hranic datového souboru

6/ Odhadování pravděpodobnosti $P(z_0 > z)$

7/ Odhadování hustoty pravděpodobnosti dP/dz

8/ Testování hypotéz o homogenitě datového souboru a určování počtu shluků

9/ Určování vah jednotlivých dat

10/ Uspořádávání dat

11/ Odhadování základních statistických charakteristik datového souboru

12/ Numerický i grafický výstup výsledků

Tento programový systém byl vyzkoušen na různých aplikacích a prokázal výhodné vlastnosti gnostických algoritmů. Výsledky dávají naději, že touto cestou získají vědecká pracoviště účinný a flexibilní nástroj pro zkoumání vlastností malých datových souborů.

Literatura:

/1/ Kovanic P., Základy gnostické teorie dat, Sborník ROBUST ´84, MFF-UK (1984).