Proč většina publikovaných vědeckých tvrzení neplatí?

Jiří Kléma

Katedra kybernetiky,
FEL, ČVUT v Praze

IDA
Intelligent Data Analysis
RESEARCH GROUP

http://ida.felk.cvut.cz/moodle/

Osnova

- úskalí statistického testování hypotéz
  - přímo zejména v medicíně a bioinformatice,
  - dopady ale i na jakýkoli jiný způsob analýzy dat,
- objektivní příčiny
  - nízká **apriorní pst** existence testované závislosti
  - frekventistické testy s ní nijak nepracují
- subjektivní příčiny
  - **zaujetí** jako důsledek "publish or perish" principu
  - **opakované testování** a "file drawer" problém
    - "Čím je výsledek statisticky významnější, tím je důležitější."
- proč jsou statistici skeptičtí k dolování dat?

---

* **Na co si dát pozor při čtení i psaní vědeckého článku?**

Jak rozpoznat podezřelé závěry? Jak mohu "nechtíc" publikovat zaujatou hypotézu?

Příklady běžných hypotéz

- Medicína
  - zjištění účinnosti nového léku N ve srovnání se stávajícím lékem S,
  - $H_0 : \mu_N = \mu_S$ (střední doba rekonvalescence je pro oba léky identická), $H_a : \mu_N < \mu_S$
  - data ideálně vznikají randomizovaným dvojitě slepým experimentem.
- Bioinformatika
  - nalezení všech diferenciálně exprimovaných genů v microarray datech,
  - dif. exprese: střední exprese genu ve vzorcích různých tříd se liší,
    (střední hodnotu lze nahradit mediánem, třídy mohou být normální a nemocné tkáně),
  - vícenásobném porovnávání, malý počet vzorků, velký počet genů, často šetření.
- Strojové učení
  - který z klasifikátorů C_A a C_B bude přesnější nad příštími příklady z dané domény?
  - $H_0 : \epsilon_A(x) = \epsilon_B(x)$, $H_a : \epsilon_A(x) < \epsilon_B(x)$
    ($\epsilon(x)$ – pst chyby klasifikace pro náhodně zvolený příklad x)
  - lze zobecnit na algoritmy učení L_A a L_B.

Spolehlivost a síla statistického testu

- α ... hladina významnosti (pst chyby I. druhu), $1-\alpha$ ... **hladina spolehlivosti**
- β ... pst chyby II. druhu, $1-\beta$ ... **síla testu**
- způsob testování: řídíme α, $\beta = f(\alpha \uparrow\downarrow, n \uparrow\downarrow, eff \uparrow\downarrow)$
  - n ... počet vzorků, eff ... velikost testovaného efektu (je testovaná závislost silná či slabá?)

	H_0 platí	H_0 neplatí
H_0 se nezamítá	$1-\alpha$	β
H_0 zamítnuta ve prospěch H_a	α	$1-\beta$

- obvyklá formulace statistického testu
  - H_0 ... pozorování je důsledkem náhody (studovaný efekt je nulový),
  - H_a ... pozorování mají i jinou než náhodnou příčinu (studovaný efekt existuje),
  - hledáme závislosti (efekty) $\longrightarrow$ pozitivním výsledkem je zamítnutí H_0 ve prospěch H_a
  - $1-\beta$ odpovídá **senzitivitě** testu (TPrate), $1-\alpha$ **specificitě** testu (TNrate)

  - jaká je pst, že nalezená závislost skutečně existuje?
  - jaká je šance správně nalezené závislosti?
    (identická otázka – poměr psti správně a falešně nalezené závislosti, přehlednější zápis)

Aposteriorní pravděpodobnost, že nalezená závislost je skutečná

- uvažujme četnosti: z n testů možných vztahů nechť v n_0 případech H_0 skutečně platí

	skutečná závislost	žádná závislost	celkem
nalezená závislost	$(1-\beta)(n-n_0)$	αn_0	$(1-\beta)(n-n_0)+\alpha n_0$
nenalezená závislost	$\beta(n-n_0)$	$(1-\alpha)n_0$	$\beta(n-n_0)+(1-\alpha)n_0$
celkem	$n-n_0$	n_0	n

- pst, že nalezená závislost je skutečná – positive predictive value (PPV), také **precision**

$$PPV = \frac{(1-\beta)(n-n_0)}{(1-\beta)(n-n_0)+\alpha n_0}$$

- může být většina nalezených závislostí klamná (tj. PPV < 0.5 nebo dokonce PPV $\to 0$)?
  - řízením α sice udržujeme nízký poměr falešně pozitivních poplachů,
  - při průměrné senzitivitě a nízkém poměru skutečně pozitivních tvrzení snadno PPV $\to 0$,
  - apriorní pst skutečné závislosti $\frac{n-n_0}{n}$ je vlastnost domény a hraje klíčovou roli,
  - apriorní pst skutečné závislosti není pozorovatelnou veličinou!

Aposteriorní šance, že nalezená závislost je skutečná

- bayesovský vztah pro šance:
  - posterior odds = prior odds × likelihood ratio → $\frac{P(H|E)}{P(\neg H|E)} = \frac{P(H)}{P(\neg H)} \times \frac{P(E|H)}{P(E|\neg H)}$
  - nová informace = původní informace × informace z testu
  - E je pozitivní výsledek testu, H je testované tvrzení o existenci závislosti
- pro testování hypotéz:
  - prior odds = R = $\frac{n-n_0}{n_0}$, likelihood ratio = LR = $\frac{1-\beta}{\alpha}$
  - aposteriorní šance může být malá pro tvrzení s nízkou apriorní pstí
  - aposteriorní šance může být malá pro testy se silou, která nejde k 1 (nijak neobvyklé pro slabší efekty a běžné velikosti množiny vzorků)
- příklad genové studie
  - nejsou neobvyklé hodnoty $R = 10^{-3}$, i pro slušné $1 - \beta = 0.6$ a běžné $\alpha = 0.05$ platí:

$$PO = R \times LR = 10^{-3} \times 12 = 0.012 \rightarrow PPV = \frac{PO}{1 + PO} \simeq 0.012$$

  - **vlivem R** je pouze 1% nalezených závislostí skutečných!

Studie mohou být zaujaté

- Zaujetí u
  - procento testů, které objektivně závislosti nenachází, ale jsou prezentovány opačně
  - mimo přirozenou variabilitu pokrytou běžným aparátem
  - zavádějící návrh experimentu, práce s daty nebo prezentace výsledku
    - pro zjednodušení uvažujme, že zaujetí nesouvisí se skutečnou existencí závislostí
    - počet nenalezených závislostí poklesne v poměru $(1-u)$
    - v horním řádku doplníme sloupcové sumy na 1

	skutečná závislost	žádná závislost
nalezená závislost	$1-\beta+\beta u$	$\alpha+(1-\alpha)u$
nenalezená závislost	$\beta(1-u)$	$(1-\alpha)(1-u)$
celkem	1	1

- Příčiny zaujetí (obecně "publish or perish")
  - selektivní publikace, absence korekce opakovaného testování, manipulace se vzorky
- Změní se věrohodnostní poměr LR

$$LR = \frac{1-\beta}{\alpha} \rightarrow \frac{1-\beta+u\beta}{\alpha+(1-\alpha)u}$$

  - pro $u = 0.05$ v genovém příkladu $LR : 12 \rightarrow 6.4$
  - **vlivem** u poklesne počet správných publikovaných tvrzení na polovinu!

Zaujetí ve strojovém učení

Data Set Selection

Doudou LaLoudouana* **and Mambobo Bonouliqui Tarare**
Lupano Tecallonou Center
Selacie, GUANA
doudoula3@hotmail.com, fuzzybear@yahoo.com

Abstract

We introduce the community to a new construction principle whose practical implications are very broad. Central to this research is the idea to improve the presentation of algorithms in the literature and to make them more appealing. We define a new notion of capacity for data sets and derive a methodology for selecting from them. The experiments show that even for not so good algorithms, you can show that they are significantly better than all the others. We give some experimental results, which are very promising.

File drawer problem (Rosenthal, 1979)

- Kdo o svém výzkumu napíše článek? Čí článek bude přijatý?

neznámá realita: jev X zvyšuje šanci výskytu nemoci N o 30%, LR=1.3
(je-li v celé populaci 20% nemocných, u lidí s X jich bude 25%)

$\Downarrow$

H_0: jev X neovlivňuje výskyt nemoci N

$\Downarrow$

Doc. Plašil	Dr. Anděl	Prof. Pešek
		
$\Downarrow$	$\Downarrow$	$\Downarrow$
$LR \sim 2, p < 0.001$	$LR \sim 1.3, p > 0.05$	$LR \sim 1, p \to 1$

File drawer problem (Rosenthal, 1979)

- Jen (především) zamítnutí nulové hypotézy je publikováno
  - negativní závěr je zřídka překvapivý
    - Cimrman J.: Vyfukováním tabákového kouře do vody zlato nevzniká.
  - i negativní výzkum je ale užitečný
    - existuje Journal of Negative Results
      - a speciální negativní časopisy pro ekologii, biomedicínu, přirozený jazyk,
    - prakticky jen občasné polemiky s dříve publikovanými pozitivními závěry,
      - pouze ve velmi renomovaných časopisech,
  - celé obory založené na zcela mylných předpokladech mohou přežívat léta
    - téměř jistě homeopatie,
    - ideální pro odhad zaujetí – všechny pozitivní závěry jsou falešné,
- File drawer problem postihuje i **meta-studie**
  - meta-studie shrnuje výsledky všech studií studujících stejný efekt/závislost
  - vynikají malým zaujetím a objektivitou (není u nich tlak na pozitivní závěr)
    - pracují ale často s hodně vychýleným vzorkem.

Některá tvrzení jsou prověřována opakovaně různými týmy

- Každá testovaná závislost má opakovanou příležitost být alespoň jednou potvrzena
  - předpokladem je nezávislost testů,
  - každý tým shromáždí vlastní měření, testuje vlastní pacienty apod.,
  - technicky nelze provést společnou korekci opakovaného testování,
  - uvažujeme m testů provedených m různými týmy,

	skutečná závislost	žádná závislost
alespoň 1 potvrzená závislost	$1-\beta^m$	$1-(1-\alpha)^m$
nenalezená závislost	β^m	$(1-\alpha)^m$
celkem	1	1

- Opět se změní věrohodnostní poměr LR

$$LR = \frac{1-\beta}{\alpha} \rightarrow \frac{1-\beta^m}{1-(1-\alpha)^m}$$

  - pro $m = 10$ v genovém příkladu $LR : 12 \rightarrow 2.5$
  - **vlivem** m poměr platných publikovaných závislostí poklesne téměř 5ti násobně!

Analýza microarray dat akceleruje vývoj statistických metod

- alternativy při vícenásobném porovnávání:
  - family-wise error rate (FWER)
    - pst, že 1 nebo více nulových hypotéz je zamítnuto falešně,
    - 0 chyb I. druhu je silný předpoklad $\longrightarrow$ testujeme s malou silou,
  - **false disovery rate** (FDR, Benjamini & Hochberg, 1995)
    - pst, že jednotlivá zamítnutá nulová hypotéza je zamítnuta falešně,
    - silnější test $\longrightarrow$ více zamítnutí než FWER se stejným prahem,
    - FDR=1-PPV . . . řídíme (odhadujeme) přímo i PPV,
- metoda SAM (**Siginificance Analysis of Microarrays**, Tusher et al., 2001)
- metoda EBAM (**Empirical Bayes Analysis of Microarrays**, Efron et al., 2001)
  - ilustrují nutnost odlišného přístupu k bioinfomatickým testům – opakované permutace tříd,
  - hlavní výhoda: žádné parametrické předpoklady ani předpoklad nezávislosti genů,
  - experimentálně odhadují apriorní pst platnosti dílčích hypotéz,
  - neřeší problém vícenásobného porovnání různými týmy – nelze použít v meta-analýze.

Shrnutí

- Testování hypotéz je dobrou vědeckou metodou
  - omezuje volnost v posuzování platnosti tvrzení,
  - platnost tvrzení váže exaktně na výsledek experimentu,
  - jinde je pravdivost závěrů ještě problematičtější,
  - **ALE!** Statistická významnost nesmí být jediným kritériem významu a platnosti závislosti.
- K vědeckým publikacím je třeba přistupovat velmi kriticky
  - objektivní příčiny chyb
    - $R \downarrow$ – velký počet testů apriorně nepravděpodobných tvrzení bez předvýběru,
    - $\beta \uparrow$ – malé počty vzorků a slabé efekty – závislosti na hraně obvyklých hladin významnosti,
    - $m \uparrow$ – horké téma prověřované mnoha lidmi,
    - $\beta \downarrow$ – extrémně rozsáhlé studie detekují i slabé závislosti s nulovým praktickým dopadem,
  - subjektivní příčiny chyb
    - $u \uparrow$ – publikace (a tedy pozitivní závěry) přináší uznání, doktoráty, habilitace, profesury,
    - $u \uparrow$ – pozitivní závěry často vyhovují výrobcům, výzkumníkům přináší peníze,
    - $u \uparrow$ – víra v platnost závislosti i α-level přenáší,
  - ptejme se po věcné významnosti závěrů, praktických ověřeních, ohlasech.

Kritika data miningu je často oprávněná

- Bagrování dat a lovení závislostí
  - $R \downarrow$ – extrémní počet testů,
  - $R \downarrow$ – nahodilé vztahy a tedy apriorně nepravděpodobná tvrzení,
  - $R \downarrow$ – předvýběr někdy na základě apriorní znalosti, většinou ale žádný.
- Speciálně bioinfomatika je velmi rizikovou doménou
  - většina výše uvedených komplikací pro ni platí,
  - microarray – mraky chybových dat,
  - explorace je chybovější než potvrzování potenciálně platných hypotéz,
  - zdlouhavé ověřování slabých pozorování je velmi pravděpodobně ztrátou času.

Zdroje přednášky

- Ioannidis, J.P.A.: *Why Most Published Research Findings Are False*, PLoS Medicine, 2005.
- Pauker, S.: *The Clinical Interpretation of Research*, 2005.
- Soukup, P.: *Statisticky významný neznamená důležitý*, Socioweb, 2007.
- Young, N.S. et al.: *Why Current Publication Practices May Distort Science*, PLoS Medicine, 2008.
- Ioannidis, J.P.A.: *Genetic Associations: False or True*, Trends in Molecular Medicine, 2003.
- Efron, B. et al.: *Empirical Bayes Analysis of a Microarray Experiment*, Journal of the American Statistical Association, 2001.
- Benjamini, Y., and Hochberg, Y.: *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*, Journal of the Royal Statistical Society, 1995.
- Tusher, V. G. et al.: *Significance Analysis of Microarrays Applied to the Ionizing Radiation Response*, Proceedings of the National Academy of Science, 2001.