Analýza dat v neurologii

XXXIV. Bayesovské sítě

L. Dušek, T. Pavlík,
J. Jarkovský, J. Koptíková

Institut biostatistiky a analýz
MU, Brno

✉
doc. RNDr. Ladislav Dušek, Dr.
Institut biostatistiky a analýz
MU, Brno
e-mail: dusek@cba.muni.cz

V tomto díle seriálu uzavřeme výklad principů bayesovské statistiky ukázkou aplikace, která má význam v teoretickém i aplikovaném klinickém výzkumu. Z prostorových důvodů se již nebudeme vracet k základním principům bayesovského hodnocení a čtenáře odkazujeme na předchozí díly seriálu číslo XXXI–XXXIII, kde jsme této problematice věnovali dostatečný prostor. Úvodem však k bayesovským sítím musíme připomenout hlavní výhody bayesovského usuzování, které bývá v literatuře označováno za zvláštní a ucelený koncept myšlení. Bayesovský přístup tedy:

- poskytuje nejen možnost odhadovat hodnoty parametrů studovaných rozdělení pravděpodobnosti, ale také pracovat s jejich neurčitostí pomocí jejich vlastního pravděpodobnostního chování, tedy velmi dobře interpretovatelným způsobem,
- umožňuje kombinovat exaktní (experimentální) vstupy s teoretickými předpoklady a odhady ve formě apriorních informací (pravděpodobností); oba typy vstupních informací přitom nejsou v rozporu, naopak se dobře doplňují.

V tomto díle krátce přiblížíme aplikaci nástroje pro modelování a zobrazování pravděpodobnostních vztahů náhodných veličin a jevů. Jde o nástroj užitečný např. při studiu vzájemných souvislostí mezi rizikovými faktory nějaké nemoci či mezi prediktory jejího vývoje. Tato problematika je velmi častá v medicínském výzkumu i praxi, jen málokdy můžeme pracovat s více prediktory, které by byly vzájemně zcela nezávislé. Přesto jsme pro jednoduchost v dosud vysvětlovaných aplikacích bayesovské statistiky předpokládali vzájemnou nezávislost použitých prediktorů a řešili jsme pouze jejich vztah s jevem, jehož pravděpodobnost jsme odhadovali. Například vztah charakteristik pacienta a nemoci k pravděpodobnosti výskytu toxicity při určité léčbě. Tento přístup ale neodpovídá realitě, kdy pravděpodobnostní vztahy mezi prediktory zcela běžně existují. Například v modelu pro predikci rizika cévní mozkové příhody bude věk pacientů ve vztahu s různými komorbiditami, jako je diabetes mellitus nebo hypertenze. S jednotlivými prediktory tak není možné automaticky pracovat jako se vzájemně nezávislými faktory. Z pohledu frekventistické statistiky jde o problém redundance prediktorů, který je jedním z nejvážnějších problémů při tvorbě vícerozměrných stochastických modelů. A také jednou z častých příčin odmítnutí takových modelů v prestižních vědeckých časopisech.

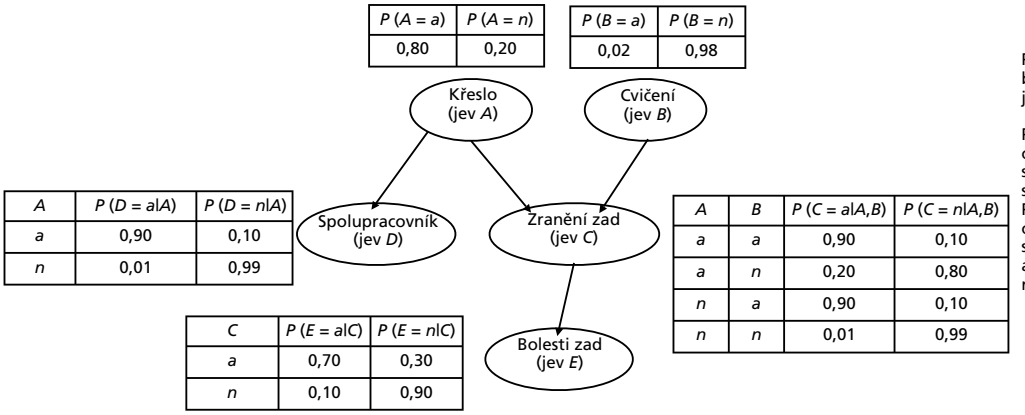
Jedním z nástrojů, které umožňují zohlednit vzájemné vztahy prediktorů, jsou tzv. *bayesovské sítě*. Tato metoda slouží k popisu pravděpodobnostní sítě vzájemných vazeb jak prediktorů, tak i cílových parametrů hodnocení. S pomocí bayesovských sítí je možné odhadnout pravděpodobnost nastání hodnoceného jevu tak, že bereme do úvahy vazby nejenom mezi prediktory a cílovým parametrem, ale i mezi prediktory navzájem. Jde tedy o modely užitečné i pro klinický výzkum, kde se jen málokdy potkáváme s čistě deterministickým rozhodováním bez neurčitostí. Přidanou hodnotou bayesovských sítí je také fakt, že jejich výstupy mohou být podpořeny přehlednou grafickou vizualizací.

Základním pojmem, se kterým bayesovské sítě pracují, je tzv. *podmíněná nezávislost náhodných jevů* či *náhodných veličin*. Náhodné jevy A a B jsou podmíněně nezávislé za podmínky C, pokud platí $P(A \cap B|C) = P(A|C)P(B|C)$. Musí tedy platit, že pravděpodobnost současného nastání jevů A a B za podmínky nastání jevu C je rovna součinu podmíněných pravděpodobností výskytu těchto jevů za podmínky C. Podmíněná nezávislost se někdy také značí $A \perp B|C$. Základním cílem bayesovských sítí je prezentovat znalosti o sledovaném náhodném jevu (cílový parametr hodnocení, *endpoint*) na pokladě nezávislých prediktorů a ty potom využít v dalším rozhodování či usuzování.

Příklad 1 (převzato z práce Ben-Gal, 2007) popisuje jednoduchou bayesovskou síť. Matematicky řečeno jde o tzv. acyklický orientovaný graf, kde jednotlivé hodnocené jevy (veličiny) jsou reprezentovány uzly grafu a ty jsou spojeny takzvanými hranami (šipky v grafu), které popisují jejich vzájemnou závislost (popisují závislost sledovaných veličin). Kromě tohoto typu sítě, kde hrany mají definovaný směr, existují ještě sítě s hranami bez definovaného směru, tzv. *markovské sítě*, jejichž výklad ale nyní přesahuje rámec tohoto článku. Oba typy sítí patří do skupiny metod nazývaných souhrnně *pravděpodobnostní grafické modely*.

Ke každému uzlu u v grafu (příklad 1) je přiřazena pravděpodobnost jeho nastání v závislosti na jeho tzv. *rodičovských uzlech*, $P(u|rodice(u))$, přičemž rodičovskými uzly jsou všechny předchozí uzly, z nichž vychází hrany (šipky v grafu) k danému uzlu (ten je nazýván také jako potomek). Topologie bayesovské sítě tak popisuje, jak „rodiče" z hlediska pravděpodobnosti ovlivňují své „potomky". Genealogická terminologie je používána také pro označení uzlů jako „předci", což je sada uzlů, ze kterých je hodnocený potomek dosažitelný přímou cestou v grafu nebo „následovníci", kteří mohou být v grafu přímo dosaženi, pokud vyjdeme z hodnoceného uzlu. Zároveň platí, že žádný uzel nemůže být sám sobě rodičem ani potomkem.

Zadání: Po zranění zad (jev *C*) může u hodnocené osoby dojít k rozvinutí bolesti zad (jev *E*). Ke vzniku zranění může dojít při nevhodném sportovním cvičení (jev *B*) nebo jako důsledek nevhodného sezení v zaměstnání (veličina křeslo – jev *A*). Pokud je zranění způsobeno nevhodným sezením, je pravděpodobné, že podobnými problémy budou trpět i spolupracovníci (jev *D*) hodnoceného pacienta. Všechny veličiny (*A, B, C, D, E*) jsou v tomto příkladu binární a mohou nabývat hodnot ano (a)/ne (n). Pro popis pravděpodobností systému konstruujeme bayesovskou síť o pěti uzlech.



Podmíněné pravděpodobnosti jsou buď počítány z primárních dat nebo jsou výsledkem externí znalosti.

Pro rodičovský uzel jde o pravděpodobnost nastání jeho stavu ano (*a*)/ne (*n*) v hodnoceném souboru / populaci. Pro potomky jde o pravděpodobnost nastání jejich stavu ano (*a*)/ne (*n*) podmíněnou aktuálním stavem jejich rodičovského uzlu(ů).

Z bayesovské sítě jsou patrné vztahy mezi uzly v pozici „rodičů" a „potomků", popsané pomocí tabulek podmíněných pravděpodobností v jednotlivých uzlech. S pomocí těchto dat a vzorce pro výpočet sdružených pravděpodobností:

$$P(u_1,....,u_n) = \Pi_{i=1}^{n} P\left(u_i \mid rodice(u_i)\right)$$

jsme schopni spočítat sdružené pravděpodobnosti ať již pro celou síť nebo pro libovolnou sadu jejích uzlů.

Spočítejme sdruženou pravděpodobnost sítě pro situaci, kdy je instalováno nové nepohodlné křeslo (*A* = *a*) a zároveň došlo u pacienta k bolestem zad (*E* = *a*), výsledná pravděpodobnost pak slouží pro podporu nebo vyvrácení hypotézy, že nepohodlné křeslo souvisí s bolestí zad. Celková pravděpodobnost, že po instalaci nepohodlného křesla dojde k bolestem zad, je po dosazení do vzorce 0,183, tedy cca 18 %.

Jak bylo zmíněno výše, sdružené pravděpodobnosti je možné počítat pro libovolnou sadu uzlů sítě. V jednoduchém příkladu, kdy nás bude zajímat pravděpodobnost nastání zranění zad (*C* = *a*) při cvičení (*B* = *a*), získáme po dosazení do vzorce p = 0,018, tedy z celé hodnocené populace je pravděpodobnost vzniku zranění zad při sportu pouze 1,8 %.

Převzato a upraveno z: Ben-Gal I. Bayesian Networks. In: Ruggeri F, Faltin F, Kenett R (eds). Encyclopedia of Statistics in Quality & Reliability. Hoboken: Wiley & Sons 2007.

Příklad 1. Aplikace bayesovské sítě pro modelový popis systému sledujícího příčiny bolesti zad.

Každému uzlu v síti přiřazujeme tabulku s rozdělením pravděpodobností jeho výskytu (příklad 1). V případě uzlů, které nemají rodiče, je to nepodmíněná pravděpodobnost, v opačných případech jde o podmíněné pravděpodobnosti.

Bayesovská síť je tedy pravděpodobnostní model, kterým můžeme popsat kauzální vazby mezi studovanými náhodnými veličinami v síti prezentovanými jako uzly. Hrana *A* –> *B* znamená, že *A* kauzálně ovlivňuje *B*, a tedy pozorování jevu *A* poskytuje kauzální podporu pro výskyt jevu *B*. Velkou výhodou bayesovských sítí je možnost vykreslit i velmi složitý systém uzlů a vzájemných vazeb graficky, což usnadňuje čtení i interpretaci. Grafická forma je také pro svou prostorovou úspornost dobře využitelná pro publikace. Bayesovskou síť nazýváme jednoduše souvislou, pokud mezi dvěma uzly existuje právě jedna neorientovaná cesta (hrana). Někdy je taková síť v literatuře označovaná jako les. Jeho zvláštní formou je strom, což je graf, kde každý uzel má jen jednoho rodiče.

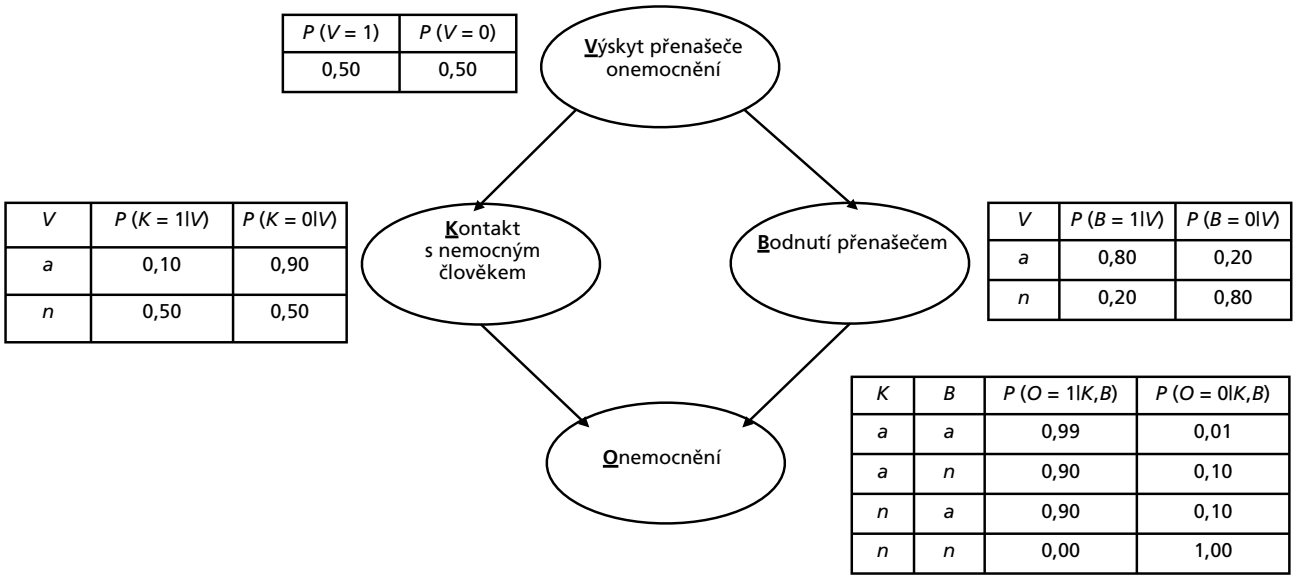
Pokud jednotlivé uzly sítě očíslujeme tak, že rodiče mají vždy nižší pořadové číslo než jejich potomci, pak platí, že každý jednotlivý uzel je podmíněně nezávislý na uzlech s nižším číslem s výjimkou svých rodičů. Tato vlastnost umožňuje výpočet tzv. *sdružené pravděpodobnosti* ať již pro celou síť nebo pro vybranou sadu jejích uzlů: $P(u_1,\ldots,u_n) = \Pi^n_{i=1} P(u_i|rodice(u_i))$. Jako sdruženou pravděpodobnost označujeme pravděpodobnost současného nastání sledovaných náhodných jevů, ve zde zavedené terminologii tedy uzlů. Dle definice platí, že sdružená pravděpodobnost nezávislých diskrétních veličin se rovná součinu marginálních pravděpodobností. To lze na příkladu dvou náhodných veličin *A* a *B* zapsat následovně:
$P(A\cap B) = P(A)P(B|A) = P(B)P(A|B)$. Jsou-li tyto veličiny zcela nezávislé, pak získáme známý vztah: $P(A\cap B) = P(A)P(B)$.

Ačkoliv šipky v síti označují příčinné vazby a jejich směr (příčinu a důsledek), je možné pohybovat se při výpočtu sdružených pravděpodobností v síti libovolným směrem.

V příkladu 2 využíváme výše popsané možnosti výpočtu sdružených pravděpodobností pro pravděpodobnostní usuzování (inferenci), kdy při znalosti struktury sítě a podmíněné pravděpodobnosti v jednotlivých uzlech můžeme vypočítat aposteriorní pravděpodobnost výskytu (nastání) libovolného uzlu sítě. Vzhledem k výpočetní náročnosti neuvádíme obecný postup výpočtu, který může být prováděn relativně širokou škálou dostupných algoritmů; v případě zájmu čtenáře odkazujeme na specializovanou literaturu a dostupný software, jako je např. knihovna *deal* pro statistický jazyk R (Boettcher and Dethlefsen, 2003).

V analýze pomocí bayesovských sítí můžeme postupovat od pravděpodobnosti výskytu jevu k jeho teoretickým

Zadání: V dané oblasti se může vyskytovat přenašeč onemocnění (vektor, *V*). Onemocnění může být přeneseno buď přímo bodnutím přenašečem (*B*), nebo kontaktem s nemocným člověkem (*K*). V případě kontaktu s nemocným nebo bodnutí přenašečem se u sledovaného pacienta může nebo také nemusí rozvinout dané onemocnění (*O*). Všechny veličiny (*V*, *K*, *B*, *O*) jsou v tomto příkladu binární a mohou nabývat hodnot ano (1)/ne (0).



P (*V* = 1)	*P* (*V* = 0)
0,50	0,50

V	*P* (*K* = 1|*V*)	*P* (*K* = 0|*V*)
a	0,10	0,90
n	0,50	0,50

V	*P* (*B* = 1|*V*)	*P* (*B* = 0|*V*)
a	0,80	0,20
n	0,20	0,80

K	*B*	*P* (*O* = 1|*K*,*B*)	*P* (*O* = 0|*K*,*B*)
a	*a*	0,99	0,01
a	*n*	0,90	0,10
n	*a*	0,90	0,10
n	*n*	0,00	1,00

Otázka: Při analýze celého systému nás zajímá, co je po zjištění onemocnění u pacienta jeho nejpravděpodobnější cestou přenosu. Vyhodnotíme tedy pravděpodobnost, zda v případě onemocnění pacienta (*O* = 1) nastal kontakt s nemocným člověkem *P* (*K* = 1|*O* = 1) nebo naopak bodnutí přenašečem *P* (*B* = 1|*O* = 1).

$$P(K=1\,|\,O=1)=\frac{P(K=1,O=1)}{P(O=1)}=\frac{\sum_{v,b}P(V=v,K=1,B=b,O=1)}{P(O=1)}=$$

$$=\frac{\sum_{v,b}(P(V=v)P(K=1\,|\,V=v)P(B=b\,|\,V=v)P(O=1\,|\,K=1,B=b))}{P(O=1)}=\frac{0{,}2781}{P(O=1)}=0{,}430$$

$$P(B=1\,|\,O=1)=\frac{P(B=1,O=1)}{P(O=1)}=\frac{\sum_{v,k}P(V=v,K=k,B=1,O=1)}{P(O=1)}=$$

$$=\frac{\sum_{v,k}(P(V=v)P(K=k\,|\,V=v)P(B=1\,|\,V=v)P(O=1\,|\,K=k,B=1))}{P(O=1)}=\frac{0{,}4581}{P(O=1)}=0{,}710$$

kde: $$P(O=1)=\sum_{v,b,k}(P(V=v)P(K=k\,|\,V=v)P(B=b\,|\,V=v)P(O=1\,|\,K=k,B=b))=0{,}6471$$

Spočtené sdružené pravděpodobnosti sumarizují celou cestu příčin a následků, jak je popsána v bayesovské síti. Kombinací rodičovských uzlů sítě předcházejících uzlu onemocnění pacienta (tedy od výskytu přenašeče přes možnost kontaktu s nemocným a možnost bodnutí přenašečem) tak zjišťujeme, že v případě onemocnění pacienta jde s pravděpodobností 0,71 o důsledek bodnutí přenašečem a s pravděpodobností 0,43 o důsledek kontaktu s nemocným člověkem.

Závěr: Maximální aposteriorní pravděpodobnost má příčina bodnutí přenašečem, tedy *P* (*B* = 1|*O* = 1) > *P* (*K* = 1|*O* = 1), pravděpodobnějším způsobem nákazy je tak bodnutí přenašečem.

Příklad 2. Aplikace bayesovské sítě pro statistické usuzování (inferenci) při diagnostice příčiny onemocnění.

příčinám (tzv. **diagnostická inference**), ale i naopak od znalosti pravděpodobnosti možných příčin (např. rizikových faktorů) k hodnocenému jevu (tzv. **kauzální inference**). Nicméně ani bayesovské sítě nejsou nástroj, který by mohl samostatně pracovat bez člověka, a zejména při hodnocení kauzality jevů si musíme uvědomovat riziko příslovečného „korelování hrušek s jablky". I když znalost kauzálních vztahů by měla být základem tvorby bayesovské sítě,

ne všechny vazby v síti musí mít nutně kauzální příčinu.

Možnost pracovat s topologií sítě představuje z výpočetního hlediska podstatnou výhodu ve smyslu redukce parametrů modelu. Představme si, že studujeme vliv pěti různých binárních znaků na určité riziko související s nemocí pacientů. Pokud bychom chtěli prověřit vzájemnou nezávislost všech těchto faktorů, museli bychom hodnotit pravděpodobnost výskytu všech jejich kombinací (po dvou, po třech, …) a celkem bychom tak hodnotili 2^5 komponent. Pravděpodobnost výskytu všech potenciálních interakcí faktorů (prediktorů) by většinou nebylo možné empiricky posoudit, nehledě na to, že by zásadně narostly požadavky na velikost vzorku v experimentu. Pravděpodobnostní model pracující na základě bayesovské sítě tak představuje velmi lákavou alternativu. Např. v příkladu 1 v síti s pěti uzly jsme hodnocení redukovali na 10 vztahů daných vzájemnými vazbami uzlů (veličin) v síti.

Bayesovské sítě představují grafický nástroj, který je intuitivně snadno uchopitelný, a pro interpretaci tak nabízí i využitelný pravděpodobnostní popis zobrazovaných vztahů. Popularita bayesovských sítí je v posledním desetiletí na vzestupu a používají se pro řadu aplikací ve vytěžování dat a textů, při vývoji postupů pro rozpoznávání řeči, v analýze signálů, předpovědi počasí a v neposlední řadě v medicínských aplikacích, zejména diagnostice. Obecně platná grafická forma sítí umožňuje začlenit jako uzly nejenom měřené náhodné veličiny (jevy), ale také hypotézy, očekávání a jiné teoretické faktory. Bayesovská síť je ideální nástroj pro kombinování apriorní (expertní) znalosti kauzality s výsledky vyplývajícími z analýzy reálných dat. Rozvoj počítačové techniky umožňuje aplikací bayesovských sítí popisovat i složité klinické problémy, kde počty sledovaných rozměrů jdou do stovek a tisíců. Po několik století budovaná teorie se tak stává základem pro studium skutečně reálných problémů. V brzké budoucnosti se bayesovské sítě stanou klíčovým nástrojem umělé inteligence a umožní její nasazování pro řešení velmi složitých systémů.

Literatura

Boettcher SG, Dethlefsen C. Deal: A Package for Learning Bayesian Networks. Journal of Statistical Software 2003; 8(20): 1–40.

Ben-Gal I. Bayesian Networks. In: Ruggeri F, Faltin F, Kenett R (eds). Encyclopedia of Statistics in Quality & Reliability. Hoboken: Wiley & Sons 2007.

Heckermann D. A tutorial on learning with Bayesian networks. In: Jordan M (ed). Learning in Graphical Models. Cambridge: MIT Press 1999.

Heckermann D. Bayesian Networks for Data Mining. Data Mining and Knowledge Discovery 1997; 1(1): 79–119.

Jiroušek R. Úvod do teorie bayesovských sítí. Praha: Vysoká škola ekonomická 1994.

Jensen FV. Bayesian Networks and Decision Graphs. New York: Springer-Verlag 2001.