VÝBĚR RELEVANTNÍCH PRAVIDEL PRO PODPORU KLINICKÉHO ROZHODOVÁNÍ

Jan Kalina, Jana Zvárová

Anotace

Systémy pro podporu klinického rozhodování jsou důležitými telemedicínskými nástroji se schopností pomáhat lékařům při procesu rozhodování při stanovení diagnózy, terapie či prognózy pacientů. Navrhli a implementovali jsme prototyp systému pro podporu diagnostického rozhodování, který má podobu internetové klasifikační služby. Specifikem tohoto systému je sofistikovaná statistická komponenta, která umožňuje pracovat i s velkým počtem příznaků. Optimalizuje totiž výběr těch příznaků, které jsou nejdůležitější pro určení diagnózy. Její chování jsme ověřili při analýze dat genových expresí z kardiovaskulární genetické studie. Článek diskutuje principy mnohorozměrného statistického uvažování a ukazuje obtíže analýzy vysoce dimenzionálních dat, kdy počet pozorovaných proměnných (příznaků) převyšuje počet pozorování (pacientů).

Klíčová slova

podpora rozhodování, mnohorozměrná statistika, extrakce pravidel, klasifikační analýza, redukce dimenzionality

1 Systémy pro podporu rozhodování

Klinické rozhodování chápeme jako proces výběru aktivity nebo posloupnosti aktivit mezi několika alternativami, přičemž se bere v úvahu i neurčitost jako jeden z aspektů ovlivňujících výsledek. Systémy pro podporu rozhodování pak lze charakterizovat jako velmi komplikované systémy vytvořené s cílem pomáhat lékaři při procesu klinického rozhodování.

Data a znalosti představují pro systémy pro podporu rozhodování hlavní zdroje pro získání informací, jsou schopné řešit řadu komplexních úloh, analyzovat různé informační komponenty, získat informaci různého typu a odvodit z nich závěry relevantní pro diagnózu, terapii nebo prognózu [8]. Porovnávají různé alternativy podle míry jejich rizika. V dnešní době již získaly své místo v řadě klinických oborů, přičemž existují specializované systémy v jednotlivých dílčích oborech medicíny nebo i systémy specializované na podporu při předepisování léků [7].

Vstupní komponenty systému pro podporu rozhodování obvykle obsahují rodinnou a osobní anamnézu, přehled symptomů, výsledky klinických a laboratorních vyšetření v různých podobách (včetně měření genových expresí), obrazovou informaci a signály i teoretické znalosti o nemocech a lécích. Už sama různost formátů představuje komplikaci pro následné statistické analýzy. Každopádně znalosti používané systémem pro podporu rozhodování v různých podobách mohou být převzaty z expertního systému vytvořeného nejlepšími experty z daného oboru, ale vždy by měly být současně validovány na reálných datech.

2 Statistická klasifikační úloha

Ze statistického úhlu pohledu je cílem systému pro podporu rozhodování naučit se na trénovacích datech (např. z klinické studie) klasifikační pravidlo, které umožní zařadit i nového pacienta do jedné z daných skupin. Například může jít o pacienta, který je vyšetřen na dálku a není součástí trénovací klinické studie. Určení diagnózy pak znamená zařadit daného pacienta do takové skupiny pacientů, kteří mají společnou konkrétní diagnózu. Ve většině případů se klasifikační pravidlo učí supervidovaným způsobem, tj. s využitím známé diagnózy u pacientů z trénovacích dat.

V medicíně se v poslední době čím dál více setkáváme s vysoce dimenzionálními daty, kdy je počet veličin (příznaků, proměnných) větší než počet pozorování [1], a to i výrazně větší. Takové situace jsou velmi časté v genetických studiích, protože počet genů v řádu desítek tisíc výrazně převyšuje počet pacientů, který ve většině takových studiích bývá jen v řádu desítek či stovek. Jiným příkladem může být analýza magnetické rezonance mozku [5].

Systémy pro podporu rozhodování by měly být schopny umožnit konstruovat klasifikační pravidla i z vysoce dimenzionálních dat. Tím umožní lékařům i jejich analýzu, kterou by jinak museli provádět v jiném softwaru se všemi komplikacemi, které s tím souvisí, zejména s opětovným zadáním dat nebo jejich přenosem. Při analýze dat v rámci systému pro podporu rozhodování není třeba, aby lékař jako uživatel systému jako celku rozuměl principům jednotlivých statistických metod. Jednotlivé kroky analýzy dat by ale těžko lékař prováděl bez specializovaného softwaru, protože kupř. analýza genových expresí nového pacienta musí vždy začít posloupností transformací dat, které musejí zahrnovat vhodnou normalizaci nebo detekci odlehlých pozorování.

Analýza vysoce dimenzionálních dat je složitá i v situaci, že všechny pozorované proměnné jsou číselné. Standardní metody mnohorozměrné statistiky se totiž nehodí pro analýzu vysoce dimenzionálních dat, protože trpí tzv. prokletím dimenzionality. Možné postupy pro jejich analýzu zahrnují (1) redukci dimenzionality pomocí analýzy hlavních komponent (PCA) nebo jiných metod, které hledají vhodné lineární kombinace proměnných [2], (2) redukci dimenzionality pomocí selekce nejdůležitějších proměnných, a (3) analýzu pomocí speciálních statistických metod ušitých na míru pro vysoce dimenzionální data [1].

Statistická analýza vysoce dimenzionálních dat nicméně zůstává komplikovaná i z toho důvodu, že je obtížné ověřit předpoklady, na kterých jsou jednotlivé metody vybudovány. Někdy jde o značně technické předpoklady. Ověření statistických předpokladů zůstává imperativem i pro metody strojového učení (umělé neuronové sítě nebo metoda SVM), přestože se toto v literatuře většinou nepřiznává a v praxi je zvykem to ignorovat. Oproti metodám strojového učení vyžadují statistické metody obvykle silnější předpoklady nebo více specifické určení modelu, což můžeme

považovat za konkrétnější specifikaci apriorních znalostí [2]. Dodejme ještě, že zejména pro komplikované nebo ohrožené případy pacientů se nehodí metody strojového učení pro absenci jasné interpretace, proč je pacient klasifikován právě do určité skupiny.

Další komplikaci statistické analýzy představuje tendence běžných klasifikátorů k přeučení, kdy míry kvality modelu (reliabilita, validita) vedou k falešně optimistickým závěrům. Zejména pro malé počty pozorování se mohou i úplně náhodné modely zdát prediktivní se schopností vysvětlit diagnózu pacientů. Řešením jsou různé verze křížové validace, které odhadují chování klasifikačního pravidla na nezávislých datech.

Rámec tohoto článku přesahují tzv. robustní statistické metody, kterým se v poslední době věnuje značná pozornost, zejména při analýze mnohorozměrných a vysoce dimenzionálních dat. Jde o metody pro řešení různých statistických úloh, které jsou odolné vůči přítomnosti odlehlých (příp. chybně naměřených) hodnot v datech. Do tohoto kontextu zapadají práce [4,5] o regularizované lineární diskriminační analýze, která je oproti klasické lineární diskriminační analýze (LDA) spolehlivá i pro vysoce dimenzionální data a současně nabízí i jednoduchou interpretaci výsledků.

3 Systém SIR

Podíleli jsme se na návrhu a implementaci prototypu systému pro podporu klinického rozhodování. Tento prototyp jsme označili jako SIR zkratkou z anglického názvu System for selecting relevant Information for decision suppoRt.

SIR představuje snadno použitelnou webovou službou k podpoře rozhodování. Současně má důmyslně propracovanou komponentu ke sběru dat. SIR dokáže importovat celý soubor dat z klinické studie automaticky spolu s datovým modelem. Data následně představují tréninkovou množinu systému, na níž se nejprve provádí selekce proměnných a následně se konstruuje optimální klasifikační pravidlo pro řešení uvažovaného problému. SIR je navržen k použití zejména mezi praktickými lékaři v primární péči, ale je schopen zpracovat data z libovolné oblasti medicíny. Podrobněji jsme SIR z informatického hlediska popsali v [6]. Nyní se už soustřeďme jen na jeho sofistikovanou statistickou komponentu.

Prvním krokem statistické analýzy uvnitř systému SIR je čištění dat, které využívá např. kontroly, zda hodnoty importovaných kvantitativních proměnných nepřekročily hranice dané datovým modelem. Dalším krokem analýzy dat z klinické studie je redukce dimenzionality. Konkrétně jde o selekci nejdůležitějších proměnných z celé sady měřených symptomů nebo laboratorních měření. Tento krok, kdy se pro další analýzu vybere jen malá sada relevantních příznaků, je nutný zejména pro vysoce dimenzionální data naměřená v genetických studiích. Provádí se dopřednou procedurou, která postupně vybírá jednotlivé příznaky způsobem optimálním pro výsledné rozhodovací kritérium. Tu nyní popíšeme.

Přínos každé (spojité i kategorizované) proměnné k vysvětlení variability odezvy (tj. k odlišení jednotlivých skupin pacientů) je vyčíslen pomocí Shannonovy informace. Konkrétně je jako první proměnná vybrána ta, která má největší hodnotu Shannonovy informace mezi všemi proměnnými. Je tudíž nejvíc relevantní z hlediska klasifikační úlohy, která je cílem analýzy. Následně metoda postupně vybírá do množiny již vybraných proměnných tu nejvíce relevantní. Tuto statistickou závislost měříme pomocí podmíněné verze Shannonovy informace, která je vyčíslena jako přínos statistické informace podmíněně vzhledem k informaci v dosud již vybraných proměnných. Nakonec se pro následné analýzy vyberou jen ty nejdůležitější proměnné, jejichž celkový přínos k vysvětlení variability odezvy dosahuje alespoň 90 %.

Proces učení klasifikačního pravidla v SIRu má schopnost automaticky zvolit některou z možných klasifikačních metod. Kritérium optimality je adaptivně vybíráno tak, aby se minimalizovalo riziko chybné klasifikace v důsledku speciálních vlastností dat a velikosti vzorku. Použité metody zahrnují lineární diskriminačníanalýzu (LDA), což je metoda mnohorozměrné statistiky, která odděluje skupiny pomocí lineární funkce, přičemž v každé skupině je předpokládána stejná kovarianční struktura. Jiným přístupem implementovaným v systému SIR je empirický bayesovský mechanismus, který minimalizuje aposteriorní bayesovské riziko přes všechny skupiny vzorků.

Systém současně umožňuje i kvantifikovat vliv přidané další proměnné, tj. přidanou hodnotu dalšího vyšetření, na diagnostické rozhodnutí. Kromě toho může procedura pro redukci dimenzionality přihlížet i k finančním nákladům na získání každého jednotlivého klinického nebo laboratorního měření. Další specialitou systému SIR je možnost kombinovat při konstrukci klasifikačního pravidla data a medicínské znalosti. Konkrétně může lékař zasahovat ručně do systému s cílem vložit do něj dodatečné odborné znalosti založené na vzdělání, zkušenosti nebo intuici; může také např. pro specifickou kombinaci symptomů a znaků odstranit některou diagnózu, pokud jejich společný výskyt má nulovou pravděpodobnost.

Když je klasifikační pravidlo hotové, k systému se může připojit lékař, který má za cíl stanovit diagnózu nového pacienta, jenž není součástí klinické studie a jehož vyšetření mohlo proběhnout na vzdáleném místě. Do systému SIR zadá lékař všechny proměnné, které byly získány procedurou výběru proměnných. Vstup do systému může být proveden prostřednictvím automaticky generovaného rozhraní z elektronického zdravotního záznamu nebo zdravotního informačního systému, nicméně ruční zadání vstupních dat je také možné. Je výhodou, že lékař jako uživatel systému nemusí rozumět pozadí metod, ale stačí mu výsledek ve formě nejpravděpodobnější diagnózy pro daného pacienta, příp. doplněný i pravděpodobností správnosti takového výsledku.

4 Validace systému SIR na kardiovaskulární genetické studii

Chování prototypu systému SIR jsme ověřovali na datech naměřených v rámci kardiovaskulární genetické studie Centra biomedicínské informatiky v Praze. Cílem této studie z let 2006–2011 bylo najít sadu genů, které souvisejí se vznikem kardiovaskulárních onemocnění v české populaci. Mikročipová technologie byla využita na změření genových expresí všech genů, což zahrnovalo 20 701 transkriptů napříč celým genomem. Celkem jsme pracovali s 59 pacienty s infarktem, 45 pacienty s cévní

mozkovou příhodou (CMP) a 77 kontrolními (zdravými) jedinci. Jde tedy o klasifikační úlohu do tří skupin.

Systém SIR vybral z celkové množiny genů 300 nejdůležitějších a sestavil klasifikační pravidlo, které využívá pouze tyto vybrané geny. Klasifikační správnost pak dosahuje 0,82. To znamená, že když se naučené klasifikační pravidlo použilo pro zařazení všech 181 jedinců z dané studie, pak pro 82 % dává správný klasifikační výsledek. Při vybrání pouhých 10 nejvíc relevantních genů se klasifikační správnost sníží na 0,65.

Je třeba si uvědomit, že sestavení klasifikačního pravidla na daných datech je obtížná úloha z různých důvodů. Například nelze ani spočítat klasickou LDA kvůli příliš velkému počtu pacientů. Dalším důvodem je, že klasifikace do tří skupin může být i výrazně obtížnější než řešení klasifikačních úloh do dvou skupin. Konečně je důvodem nutnost založit klasifikační pravidlo na poměrně velkém počtu genů v řádu stovek (rozhodně ne desítek).

Pro porovnání jsme použili i redukci dimenzionality přes PCA. Zde bylo potřeba dát pozor na to, že ne všechny její implementace jsou vhodné pro vysoce dimenzionální data [3]. V první řadě jsme zjistili, že skupina několika prvních hlavních komponent není schopna výrazně pomoci při budování spolehlivého klasifikačního pravidla pro určení diagnózy pacientů z trénovací množiny dat. Proto analýza na 10 hlavních komponentách dává klasifikační správnost 0,52 a na 300 hlavních komponentách pak 0,85, ovšem za cenu obtížné interpretace i potřeby pozorovat exprese na všech 20 701 transkriptech.

Celkově systém SIR umožňuje podporu diagnostického rozhodování pro jedince, kteří zatím neprodělali infarkt ani CMP, ale jsou jejich manifestací ohroženi v nejbližší době. Tak mohou být pacienti se zvýšeným rizikem dalšího infarktu monitorováni nebo efektivněji a bezpečněji léčeni nebo začleněni do preventivního programu.

5 Závěry

Článek je věnován systémům pro podporu klinického rozhodování, a to zejména jejich statistické úloze konstruovat klasifikační pravidlo pro zařazení nového pacienta do jedné z předem daných skupin. Vysvětluje, jaké obtíže vyvstávají při analýze vysoce dimenzionálních dat, na něž se v medicíně naráží čím dál častěji. Navrhli a implementovali jsme prototyp systému SIR, který v tomto článku rozebíráme systém SIR ze statistického hlediska, aniž bychom se soustředili na popis jeho technických parametrů.

Systém SIR je vybaven procedurou pro výběr nejvíce relevantních proměnných (tj. relevantní informace) z naměřených dat. Ty jsou vybírány způsobem, který je optimální pro konstrukci klasifikačního pravidla. Systém jsme následně ověřili na reálných datech z kardiovaskulární genetické studie. Selekce proměnných nabízí srozumitelnou interpretaci. Celý článek tak slouží jako ilustrace toho, jaké úlohy se řeší v mnohorozměrné (bio)statistice a jaké jsou principy statistického uvažování při analýze vysoce dimenzionálních medicínských dat.

Poděkování

Práce vznikla s podporou grantu 17-01251S "Metaučení pro extrakci pravidel s numerickými konsekventy" Grantové agentury České republiky.

Literatura

[1.] Kalina, J. (2014). Classification methods for high-dimensional genetic data. Biocybernetics and Biomedical Engineering, vol. 34, 10–18.

[2.] Kalina, J. (2015). Statistical challenges of big data analysis in medicine. International Journal on Biomedicine and Healthcare, vol. 3, 24–27.

[3.] Kalina, J., Duintjer Tebbens, J. (2014). Metody pro redukci dimenze v mnohorozměrné statistice a jejich výpočet. Informační bulletin České statistické společnosti, vol. 25, 13–29.

[4.] Kalina, J., Hlinka, J. (2016). On coupling robust estimation with regularization for high-dimensional data. Studies in Classification, Data Analysis and Knowledge Organization. Přijato.

[5.] Kalina, J., Hlinka, J. (2016). Implicitly weighted robust classification applied to brain activity research. Biomedical Engineering Systems and Technologies, Communications in Computer and Information Science. Přijato.

[6.] Kalina, J., Seidl, L., Zvára, K., Grünfeldová, H., Slovák, D., Zvárová, J. (2013). System for selecting relevant information for decision support. Studies in Health Technology and Informatics, vol. 186, 83–87.

[7.] Kalina, J., Zvárová, J. (2016). Decision support for mental health: Towards the information-based psychiatry. In Psychology and Mental Health: Concepts, Methodologies, Tools, and Applications, 1–14. IGI Global, Hershey.

[8.] Zvárová, J., Zvára, K. (2011). e3 Health: Three main features of modern healthcare. In Moumtzoglou, A., Kastania, A. (ed.), E-health systems quality and reliability: Models and Standards, 18–27. IGI Global, Hershey.

Kontakty

RNDr. Jan Kalina, Ph.D.,
(1) Ústav informatiky AV ČR
Pod Vodárenskou věží 4, 182 08 Praha 8
(2) Ústav teorie informace a automatizace AV ČR
Pod Vodárenskou věží 4, 182 08 Praha 8
tel: 266 053 099
e-mail: *kalina@cs.cas.cz*
http://www2.cs.cas.cz/~kalina

Prof. RNDr. Jana Zvárová, DrSc.,
Ústav informatiky
AV ČR, Pod Vodárenskou věží 2
182 07 Praha 8
e-mail: *zvarova@cs.cas.cz*