5. ročník
Žilina 2.- 4. júla 2013
Otvorený softvér vo vzdelávaní, výskume a v IT riešeniach
ossconf.soit.sk
OSSConf

OSSConf 2013: 17–24

PDFLaTeX A TVORBA NESTATICKÝCH ELEKTRONICKÝCH DOKUMENTŮ

ROBERT MAŘÍK (CZ)

Abstrakt. V příspěvku shrneme současné dostupné možnosti a nástroje zařazení interaktivních prvků do PDF dokumentů. Vhodné používání těchto prvků vede k tvorbě atraktivnějších a obsahově bohatších souborů. Důraz bude kladen na tvorbu učebních materiálů pomocí pdfLaTeXu.

Klíčová slova. LaTeX, PDF.

PDFLaTeX IN CREATION OF NONSTATIC ELECTRONIC PUBLICATIONS

Abstract. We summarize existing tools and methods for inserting interactive objects into PDF documents.The main stress is given on tools and techniques suitable and capable to create eye candy and easy to use teaching materials with pdfLaTeX.

Key words and phrases. LaTeX, PDF.

Úvod

Prakticky všechny dokumenty vytvářené vědeckými a pedagogickými pracovníky na vysokých školách jsou dokumenty nadčasové, s dlouhou životností. Pokud obsahují nějaké komplikovanější formátování než běžné kancelářské dokumenty je vhodné pořizovat jinak, než v sice běžných, ale často akutalizovaných programech a kancelářských balících s problematickou zpětnou kompatibilitou. Pokud se navíc jedná o texty matematického nebo technického charakteru, jsou TeX nebo LaTeX téměř jasnou volbou. Vzhledem k početnosti uživatelů se změříme spíše na LaTeX. Je vhodné si uvědomit, že LaTeX je mnohem více než systém pro snadnou sazbu a archivaci matematických a dalších textů. LaTeX je programovací jazyk, který umožňuje provádět s textem operace, které u WYSIWYG programů nemají obdobu.

V tomto příspěvku si ukážeme některé dostupné nástroje, které umožní naše texty obohatit o nestatický obsah a vytvořit tak nestatický dokument. Nestatickým dokumentem budeme v tomto příspěvku rozumět dokument obsahující prvky reagující na činnost uživatele. To si vyžaduje zobrazení dokumentu v prohlížeči, který tyto interaktivní prvky podporuje. Pokud se jedná o PDF dokument

pracující s Javascripty, je nutné použít pro práci s takovým dokumnetem zpravidla pouze program Adobe Reader, částečná podpora Javascriptů je i v programu FoxitReader.

Příspěvek je členěn následovně. V úvodní části zrekapitulujeme některé LaTeXové balíčky umožňující tvořit nestatické PDF dokumenty. Poté si ujasníme, z jakých stavebních prvků může být tento nestatický dokument sestaven a na závěr odkážeme na nástroje, metody a zdroje informací pro práci s těmito stavebními prvky.

1. Ukázky nestatických PDF dokumentů

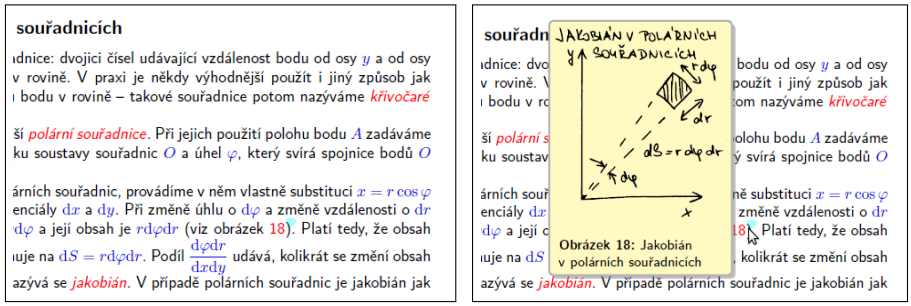
1.1. Balíčky `ocg-p` a `ocgtools`

Balíčky `ocg-p` a `ocgtools` pracují s vrstvami v PDF dokumentech (viz dále). Balíček `ocg-p` umožňuje vkládat do PDF dokumentu materiál, který je viditelný nebo tisknutelný pouze za určitých podmínek. Přímo v balíčku `ocg-p` jsou například makra pro sazbu tabulek, které se samy seřadí podle sloupce na jehož záhlaví klikne čtenář myší. Trik spočívá v tom, že se tabulka vysází v několika různě seřazených vrstvách na jedno místo a pro zobrazení a tisk se použije vždy jen jedna z těchto vrstev.

Balíček `ocgtools` obsahuje makra pro vkládání vrstev a aktivních oblastí dokumentu. Aktivní oblasti při kliknutí nebo najetí myší přepnou u odpovídajících vrstev viditelnost. Příkladem použití je vložení náhledu obrázku, který se při kliknutí zvětší na celou stranu PDF.

1.2. Balíček `fancytooltips` a skript `fancy-preview`

Balíček `fancytooltips` funguje z hlediska čtenáře podobně jako `ocgtools` – v dokumentu máme k dispozici aktivní oblasti, které umožňují zobrazit dodatečný materiál. Na rozdíl od balíčku `ocgtools` se však nejedná o kus našeho dokumentu, ale o stránku externího PDF souboru. S balíčkem `fancytooltips` spolupracuje skript `fancy-preview`. Tento skript nejprve projde náš dokument a extrahuje do pomocného dokumentu prostředí, která nás zajímají, například číslované rovnice, obrázky, tabulky a odkazy na literaturu. Poté skript provede kompilaci pdfLaTeXem s tím, že předefinuje příkazy `\ref` a `\eqref` tak, aby kromě čísla odkazovaného objektu a hypertextového odkazu vkládaly i tlačítko, které odkazovaný objekt zobrazí vložený do aktuální strany na místě ve kterém je kurzor. Možné aplikace zahrnují tvorbu učebních textů a vědeckých článků, kde se při najetí za odkaz na rovnici, obrázek nebo položku seznamu literatury zobrazí tento odkazovaný objekt v aktuální stránce. K odskoku na stránku s objektem dojde až při kliknutí. Výhodou je, že pro rychlé zorientování stačí náhled a odpadá tím nutnost návratu na stranu, ze které vedl odkaz.



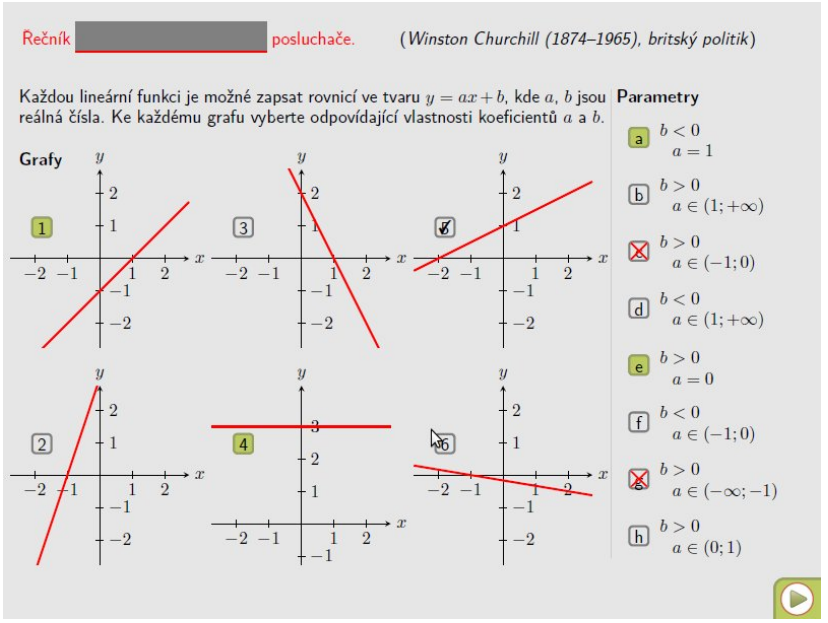
Obrázek 1. `Fancy-preview`: čistý text a text s aktivovaným náhledem odkazovaného obrázku.

1.3. Balíček `dps` – párovací hra Das Puzzle Spiel

Balíček `dps` z programového balíku AcroTEX slouží pro tvorbu výukových her ve formě dokumentů obsahujících zaškrtávací políčka rozdělená do dvou skupin, například otázky a odpovědi. Jedna odpověď patří vždy k jedné otázce. Úkolem čtenáře je spárovat otázky a odpovědi. Při každém vytvoření správného páru se objeví políčko tajenky. Hra je vhodná například pro oživení vyučovacích hodin. Snímek obrazovky na obrázku 2 je příkladem takové hry. Jedná se o dokument vytvořený pro projekt Matematika s radostí ([1], `http://msr.vsb.cz`) s použitím vlastní modifikace balíčku `dps`. Tato modifikace zahrnuje vložení tajenky do vrstev místo do textového pole (umožní mít v tajence nejenom čistý text, ale libovolný objekt vysázený TEXem) a náhodné permutování otázek a odpovědí při otevření PDF dokumentu (s využitím stejného triku jako u výše popsaných tabulek balíčku `ocg-p`, které se řadí podle různých kritérií).

1.4. Systém AcroTEX

Systém AcroTEX (`http://www.acrotex.net`) je systém pro tvorbu elektronických testů. V jeho možnostech je tvorba testů s testovými otázkami, které jsou uzavřené (volba jedné nebo několika odpovědí z nabízených variant) nebo otevřené (odpovědí je textový nebo matematický výraz, v případě matematického výrazu je správnost odpovědi vyhodnocována numerickým výpočtem a procedury pro vyhodnocení správnosti si poradí i s případy, kdy je správná odpověď dána jednoznačně až na aditivní nebo multiplikativní konstantu). Správnost odpovědí může být vyhodnocena buď okamžitě nebo až po zodpovězení všech otázek. Podle nastavení může být řešitel testu informován buď jenom o počtu správných odpovědí a počtu získaných bodů, nebo mohou být do testu i vyznačeny opravy a správné odpovědi. Pro projekt Matematika s radostí jsme styly AcroTEXu rozšířili tak, aby při otevření PDF souboru došlo k náhodnému permutování variant jednotlivých odpovědí podobně, jako je tomu u párovacích her.



Obrázek 2. Párovací hra.

2. Typy nestatických objektů v PDF

V této podkapitole si sesumarizujeme, které nejdůležitější objekty zajišťují interaktivitu v PDF souborech.

2.1. Javascripty a formulářové prvky

PDF dokumenty mohou obsahovat formulářové prvky, jako jsou zaškrtávací políčka, rozbalovací menu, textová pole, tlačítka a přepínače. Všechny tyto prvky mohou být viditelné nebo skryté, editovatelné (a klikací) nebo pouze ke čtení. Možnosti formátování nejsou příliš bohaté (je možné nastavení barvy objektu, barvy okraje, tloušťky okraje, jsou omezení na písmo pro popisek a obdélníkový tvar hranice). Pokud však v praxi nedostačují, je možné omezení plynoucí z dostupných možností PDF formátu obejít tak, že objekt vysázíme s nulovými rozměry průhledný a bez popisku a pod něj vysázíme TEXem cokoliv potřebujeme. K těmto objektům mohou být připojeny Javascripty spouštěné při různých událostech (onblur, mouseover, mouseexit, enter atd.). Platí i částečný opak, že vlastnosti těchto objektů mohou být měněny pomocí Javascriptů. Je tedy možné programově kterýkoliv z těchto objektů učinit skrytým nebo needitovatelým, vložit do něj určitý popisek nebo symbol, nastavit barvu textu, symbolu nebo okraje a podobně.

Další Javascripty je možné spouštět při otevření PDF souboru (do této části zpravidla také zapisujeme počáteční hodnoty společných proměnných a definice funkcí) nebo při otevření určité strany.

2.2. Optional Content Groups (vrstvy)

Části PDF dokumentu mohou být označeny jako Optional Content Groups (OCG, zkráceně layers, tj. vrstvy). U těchto vrstev je možno nastavit, zda se budou či nebudou zobrazovat a zda se budou či nebudou tisknout. Viditelnost vrstev je podobně jako u formulářových prvků možné při smluvených akcích a událostech přepínat pomocí Javascriptů. Ve vrstvě může být libovolný text nebo objekt vysázený TeXem.

Využití vrstev jsme viděli v balíčcích `ocg-p` a `ocgtools`. Další využití zahrnuje například zatím experimentální možnost balíčku `hyperref` vkládat hypertextové odkazy do PDF dokumentu tak, že se na obrazovce zobrazují barevně a jsou výrazné, ale tisknou se černě a neruší okolní text na papíře barvou nebo šedým odstínem.

2.3. Formulářové prvky podruhé

Pomocí vrstev je možné nechat na stránce objevit nebo zmizet jakýkoliv text, který je na této stránce zapsán. Existuje ještě jeden mechanismus, umožňující zobrazit ve stránce nějaký text či objekt. Tato možnost je založena na tom, že tlačítka kromě popisku mohou obsahovat i ikony – stránky jiného PDF souboru. Tyto ikony je možné pomocí Javascriptů měnit a je také možné nastavit jak se mají ikony chovat a kam umisťovat, pokud velikost ikony neodpovídá velikosti tlačítka. Je tedy možné kamkoliv do hlavního dokumentu vložit neviditelná tlačítka obsahující jako ikony stránky pomocného PDF dokumentu a tlačítka, na která budeme tyto ikony umisťovat dle potřeby. Takto je možno Javascriptem kteroukoliv ikonu zkopírovat ze skrytého tlačítka a zobrazit ji na kterémkoliv jiném tlačítku. Pokud je toto tlačítko průhledné, bez okraje a jen ke čtení, je efekt stejný jako kdybychom v tomto místě odkryli skrytou vrstvu s textem. Výhoda je v tom, že informace v tomto místě nemusela být zapsána autorem dokumentu, ale přenesli jsme ji na cílové místo z úložiště ve formě skrytého tlačítka. Tato technika byla pravděpodobně poprvé použita v balíčku `animfig` pro vkládání animací do PDF souborů, je popsána v [2] a je použita v balíčku `fancytooltips`.

3. Metody vkládání nestatických objektů do PDF

Tato kapitola slouží jako jistý rozcestník pro autory, kteří mají zájem vkládat do PDF dokumentů tvořených pdfLaTeXem interaktivní prvky. Do PDF souboru je možné vkládat příslušné příkazy přímo pomocí primitivů programu pdfLaTeX, častější a pohodlnější je však využití balíčků připravených pro tuto práci. Přesné možnosti, názvy a syntaxe příkazů jsou zpravidla součástí dokumentace, proto je

uvádět nebudeme. Místo toho se pokusíme upozornit na některá úskalí čekající na uživatele.

Není-li řečeno jinak, jsou stabilní verze balíčků buď přímo v použitém systému (nejčastěji MikTEX a TEX Live) nebo v archivu CTAN (`http://www.ctan.org`).

Při výměně verzí PDF souboru se spoluautory nebo při výměně rukopisů a opravených rukopisů s redakcí časopisů je běžné používat anotace. Tyto anotace se používají při čtení hotového PDF souboru v příslušném GUI. Kromě toho je možné vkládat anotace již při tvorbě PDF souboru, což umožňuje balíček `pdfcomment`.

Nejužitečnějšími balíčky jsou zřejmě balíčky systému AcroTEX. Tyto balíčky nejsou na TEX Live a je nutné si je v případě použití s TEX Live nainstalovat ručně. Primárním zdrojem k šíření není CTAN, ale autorovy vlastní stránky `http://www.acrotex.net` a `http://www.math.uakron.edu/~dpstory/webeq.html`, verze na CTAN nemusí být aktuální. V praxi se mi dobře osvědčil volně šiřitelný AcroTEX eDucation Bundle, díky podpoře pdfLATEXu. Na výběr jsou i balíčky AcroTEX eDucation Bundle Pro a AcroTEX Presentation Bundle, které obsahují dodatečné funkce nebo šablony prezentací a vyžadují finální zpracování PDF dokumentu v programu Adobe Professional. V AcroTEXu jsou přítomny například balíček `insdljs` pro vkládání Javascriptů do PDF dokumentů, `eforms` pro vkládání formulářových prvků, `exerquiz` pro tvorbu interaktivních testů. Balíčky `eforms` a `insdljs` jsou vyžadovány i při použití `ocgtools` a `fancytoltips`.

Na dvě úskalí při použití balíčku AcroTEX je vhodné upozornit. První z nich se týká pouze slovenských uživatelů a spočívá v tom, že jeden z regulárních výrazů použitých v Javascriptech dokumentu je v konfliktu s aktivními znaky balíku `babel` při použití jazyka `slovak`. Často pomůže načíst nejdříve balíčky AcroTEXu a teprve poté `babel` a slovenštinu. Další úskalí na tvůrce interaktivních testů spočívá v tom, že aby byly testy schopny vyhodnocovat matematické výrazy, je v Javascriptech těchto dokumentů použita funkce `eval`. Kvůli této funkci označuje Google testy tvořené AcroTEXem jako nebezpečné a chová se k nim jako k virům – neumožní je posílat jako přílohy a nedoručí je na poštovní účet služby Gmail. Jedná se o nepříjemnost, která je známa a postihuje více projektů. Není jasné, zda je možné doufat, že se v tomto směru něco změní, proto je vhodné mít tuto skutečnost buď na paměti nebo v ideálním případě vhodně ošetřenu, abychom nevylučovali velmi početnou skupinu uživatelů služby Gmail. Například v testech projektu Matematika s radostí [1] tato nepříjemnost odpadá, protože jsme díky nepoužívání otevřených otázek v testech mohli funkci `eval` vymazat.

Z pera stejného autora a ze stejného zdroje jako AcroTEX jsou balíčky `dps` a `jj_game` pro tvorbu párovacích her a hry Jeopardy. Kromě toho pro tvorbu hry Jeopardy existuje na CTAN ještě i balíček `jeopardy`.

Vkládání vrstev do PDF souboru řeší například balíčky `ocg` a `ocg-p`. Balíček `ocg` je starší neudržovaný, je omezen pouze na pdfLATEX a vykazuje nepříjemné chování při použití velkého množství vrstev (projeví se však až u řádově stovky

vrstev). Proto je vhodnější využívat `ocg-p`. Balíček `ocgtools` pro schovávání materiálu do vrstev a vyvolávání tohoto materiálu pomocí tlačítek byl navržen pro starší verzi `ocg`, měl by však pracovat i s `ocg-p`.

Pokud máme na interaktivní dokument vlastní požadavky, které nejsou řešeny přímo některým LaTeXovským stylem, je vhodné si uvědomit, že možnosti ladění Javascriptů v PDF dokumentu jsou (zejména bez použití programu Adobe Professional) velmi omezené. Proto je velmi vhodné naplánovat práci dokumentu jako sled jednotlivých malých kroků, každý krok u kterého to je možné si odladit samostatně a teprve poté vše spojit do jednoho dokumentu. Výhodou tohoto postupu je, že při řešení malých kroků, které se vzpírají našemu TeXařskému úsilí je relativně snadné zformulovat prosbu o pomoc na některém z diskusních fór věnovaných LaTeXu a doufat v pomoc komunity.

Bohatou studnicí nápadů a zdrojem odpovědí na naše otázky je server `http://tex.stackexchange.com`. Zde jsou k dispozici například makra, která dokáží ohraničit úsek textu tak, že při kliknutí na tento text se změní zoom při prohlížení dokumentu tak, aby byl zobrazen právě tento označený úsek[1]. V projektu Matematika s radostí jsme například z tohoto serveru využili nápad s makrem, které umí rozřezat obrázek na kousky[2]. Stačilo tyto části vkládat do vrstev odkrývaných při odhalování tajenky párovací hry a rázem jsme měli k dispozici možnost tvořit párovací hry, kde tajenkou není text, ale postupně odkrývaný obrázek.

Při tvorbě interaktivních PDF dokumentů kombinujících použití několika výše uvedených balíčků je možno očekávat problémy, proto je nutno vše nejprve odladit na krátkém dokumentu. Zejména mohou nastat problémy v dokumentech, kde je použit registr `pdfpageattr`, který není zvykem předefinovávat a proto s jeho nenulovým obsahem nemuseli autoři balíčku počítat. Někteří uživatelé jej však předefinovávají, aby předešli zkreslení barev pod Linuxem[3].

4. Závěrem

Vzhledem k rozvoji HTML5 a úpadku některých technologií vytvořených firmou Adobe se někdy věští i brzká stagnace a zánik formátu PDF. Přirozenou otázkou tedy je, zda tvorba PDF dokumentů není slepou uličkou.

Dle mého názoru podpora PDF dokumentů v operačních systémech ohrožena není, protože PDF je v podstatě jediný formát používaný pro tvorbu a archivaci vědeckých článků, nepočítáme-li export do html, což je zatím vzhledem ke kvalitě zobrazování matematiky možné brát spíše jenom jako orientační náhled.

Mé přesvědčení, že se od formátu PDF a programu pdfLaTeX neodvrátí běžní uživatelé typu autor vysokoškolských učebních textů se upevňuje pokaždé, když

[1] `http://tex.stackexchange.com/questions/12290/`

[2] `http://tex.stackexchange.com/questions/70458`

[3] viz například `http://tex.stackexchange.com/questions/35868` a `http://tug.org/pipermail/pdftex/2007-December/007482.html`

kolegové píšící matematické texty v programu Microsoft Word upgradují na novou verzi a musí ve svých dokumentech kontrolovat a přepisovat matematické výrazy.

Uživatel pdfLaTeXu se však nemusí cítit ohrožen ani v případě, že by úpadek formátu PDF opravdu nastal. Stačí dodržovat jistá pravidla při pořizování dokumentů, zejména dbát na oddělení obsahu dokumentu od definice vizuálního stylu. Například v projektu Matematika s radostí [1] sestavuje tým přibližně třiceti pracovníků interaktivní testy a hry v PDF formátu. Bylo by jistě dobré mít tuto práci uchovanou tak, aby se dala využít i v budoucnu a aby nebyla ohrožena podporou či nepodporou programů pro práci s PDF. Proto k zápisu otázek nepoužíváme přímo příkazy AcroTeXu, jak je dokumentováno v manuálu AcroTeXu nebo v [3], ale vytváříme mezi naším dokumentem a AcroTeXem další vrstvu. Ve stylu máme pro zapsání otázky definován příkaz s jedním nepovinným a dvěma povinnými parametry. Nepovinným parametrem může být klíčové slovo nebo slova ovlivňující formátování otázky, prvním povinným parametrem je text otázky a druhým povinným parametrem jsou varianty odpovědí, kde odpovědi jsou odděleny čárkou a správná odpověď je buď uvedena jako první, nebo začíná hvězdičkou. V otázkách mohou být použita vlastní makra definovaná ve speciální sekci v hlavičce dokumentu. Tento formát byl zpočátku navržen pouze pro snadný a rychlý zápis otázek a odpovědí. Časem se však ukázala vhodnost této volby, když od grafika přišly poznámky ke grafickému vzhledu (například vizuální oddělení otázek a odpovědí) nebo když došlo na rozšíření funkcionality (na rozdíl od běžných testů tvořených AcroTeXem, testy projektu Matematika s radostí automaticky permutují varianty odpovědí). Předefinováním použitých příkazů dokážeme vyexportovat otázky a odpovědi ve zkompilovatelných tvarech, ať už jako úryvek TeXovského dokumentu nebo PNG náhled otázky náhledy jednotlivých odpovědí. V případě, že bude dostupná vhodná nová technologie, neměl by pro zkušeného programátora být velký problém přepnout testy na tuto technologii a vdechnout tak AcroTeXovým testům v PDF nový život.

Reference

[1] *Matematika s radostí*, projekt OPVK, reg. č. CZ.1.07/1.1.00/26.0042, `http://msr.vsb.cz`.

[2] MAŘÍK, R.: *Vkládání JavaScriptů pdfLaTeXem prakticky*, Zpravodaj CSTUG. 2007. sv. 2, č. 2, s. 72–83. ISSN 1211-6661.

[3] MAŘÍK, R. – PLCH, R. – ŠARMANOVÁ, P.: *Tvorba interaktivních testů pomocí systému AcroTeX*, Zpravodaj CSTUG. 2010., sv. 20, č. 4, s. 266–299. ISSN 1211-6661.

Kontaktní adresa

doc. Mgr. Robert Mařík, PhD., Ústav matematiky, Fakulta lesnická dřevařská, Mendelova univerzita, Zemědělská 1, 613 00 Brno, Česká republika,
E-mailová adresa: `marik@mendelu.cz`, `http://user.mendelu.cz/marik`

OTVORENÝ SOFTVÉR VO VEDE A VÝSKUME

Otvorený softvér je často považovaný len za jednu z ciest ako sa dostať k pomerne kvalitnému alternatívnemu softvéru zadarmo. Je to tak aj vo vede a výskume? Moje doterajšie skúsenosti pri vedení diplomantov a doktorandov ale aj vlastnom výskume hovoria, že otvorený softvér ponúka viac.
Mladým začínajúcim vedeckým pracovníkom umožňuje dostať sa k takým sofvérovým nástrojom, ktoré sú pre nich inak nedostupné. Ale nie je to úplne zadarmo. Čaká ich nejedno rozhodnutie, od výsledku ktorého závisí, či budú s výberom spokojní. Nehovoriac už o prekážkach, ktoré treba prekonať pri inštalácii a overovaní kvality takýchto produktov. Skúsenosť so zdolávaním takýchto prekážok však prispieva k ich odbornému rast. Otvorený kód je aj príležitosťou pre zvedavých programátorov nahliadnúť do sveta tvorcov kvalitných programov, zoznámiť sa s ich technikami, pokúsiť sa opravu prípadných „bugov“ či tvorbu vlastných doplnkov. K nezaplateniu je aj rýchly neformálny kontakt s „prirodzene otvorenými“ autormi programov.
Čo ponúka nám skôr narodeným, viac či menej poučených z nenaplnených očakávaní? Pre mňa je to možnosť vyskúšať v pomerne krátkom čase nástroje s novými optimalizačnými solvérmi. V prípade recenzií a expertných prác tak mám príležitost ponúknuť alternatívu k jednostrannej orientácii na drahé komerčné riešenia. Pri vlastnom výskume sa zase môžem rýchlejšie rozhodnúť, či sa oplatí vyvíjať nové modely a metódy riešenia alebo postačuje vhodná kombinácia dostupných OSS riešení. Oceňujem tiež uživatelské pohodlie pri nastavovaní parametrov väčšiny OSS nástrojov na požadovanú mieru. Iste nie na poslednom mieste je to i studnica nápadov „ako na to“.

Rád by som upozornil na niekoľko tohoročných príspevkov, ktoré ma poučili. Dozvedel som sa ako použiť nástroj Weka pre výpočet odhadu rizika u pacientov so srdcovým zlyhaním, ako sa možno pohrať s jazykom Ruby a miniatúrnym akcelerometrom pri meraní dynamických vlastností šípu i ako netradične využiť logický diferenciálny počet v teórii spoľahlivosti. Ostáva mi zaželať, aby sa príspevky v tejto sekcii stali podnetom pre ďalších autorov, ktorí by sa radi podelili nielen o svoje pozitívne ale i negatívne skúsenosti s otvoreným softvérom a jeho tvorbou pri riešení svojich vedeckých projektov.

Štefan Peško

5. ročník
Žilina 2.- 4. júla 2013
Otvorený softvér vo vzdelávaní, výskume a v IT riešeniach
ossconf.soit.sk
OSSConf 2013: 27–32

RISK ESTIMATION OF HEART FAILURE PATIENTS USING WEKA

JÁN BOHÁČIK (UK, SK), C. KAMBHAMPATI (UK), DARRYL N. DAVIS (UK) AND MIROSLAV BENEDIKOVIČ (SK)

Abstract. Two to three percent of the adult population suffer from heart failure and half of all patients with this diagnosis die within four years. To minimize life-threatening situations and costs of treatment, it is interesting to predict if a patient with known heart failure could die soon. For these reasons risk estimation of heart failure patients on the basis of collected data is described in this paper. The risk estimation makes use of some data mining techniques which are implemented in open source software tool Weka - Waikato Environment for Knowledge Analysis. A description of used risk estimation models based on data mining techniques and experimental results showing the performance of these models are also given.

Key words and phrases. Data mining, risk estimation, heart failure.

ODHADOVANIE RIZIKA PRI PACIENTOCH SO SRDCOVÝM ZLYHANÍM POMOCOU NÁSTROJA WEKA

Abstrakt. Dva až tri percentá dospelej populácie trpí zlyhaním srdca a polovica všetkých pacientov, ktorí majú túto diagnózu zomrú do štyroch rokov. Za účelom minimalizácie životu ohrozujúcich situácií a nákladov na liečbu je zaujímavé odhadovať či by pacient so známym zlyhaním srdca mohol zakrátko zomrieť. Práve v tomto článku je popísané odhadovanie risku u pacientov so zlyhaním srdca. Pri odhadovaní risku sa využívajú techniky dolovania z údajov, ktoré sú implementované v open-source softvérovom nástroji Weka — Waikato Environment for Knowledge Analysis. K dispozícii je daný aj popis použitých modelov na odhadovanie risku založených na technikách pre dolovanie z údajov a experimentálne výsledky ukazujúce výkonnosť jednotlivých modelov.

Kľúčové slová. Dolovanie z údajov, odhadovanie risku, zlyhanie srdca.

Introduction

According to [5], heart failure makes up an important medical, social, and economic problem. Patients with heart failure suffer disabling symptoms, the most common of which are fatigue and dyspnea, while in terms of disability, the end stage of the disease is comparable to the end stage of terminal cancer. Although there are problems with reliable estimates in many countries, the prevalence of heart failure is about 2%–3% of the adult population and it increases with age. The cost of medical care for heart failure is measured in billions of euros per year.

Within the costs, the most powerful contributing factor is repeated hospitalization. The prevalence of heart failure progressively increased from the early 1950s. It is likely a new increase will be observed in the future mainly because of the aging of the population and because of the trend showing an increasing prevalence of major heart risk factors, including obesity and diabetes.

The long-term prognosis associated with heart failure is not optimistic at all. Half of all patients diagnosed with heart failure die within four years. It is very interesting to predict if a patient dies soon so that an effective prevention can be employed. The goal is to minimize life-threatening situations and to minimize the costs of treatment. Due to massive explosion of data collected about patients in hospitals and home telemonitoring systems, data mining techniques such as a Bayes network classifier and a decision tree classifier can be used. In this paper we concentrate on data mining techniques implemented in open source software tool Weka - Waikato Environment for Knowledge Analysis [4] and we employ it on our heart failure dataset from Hull LifeLab [2].

The paper is organized as follows. Our data about patients with heart failure are described in Section 1. In Section 2 open source software tool Weka and used data mining techniques are discussed. The performance of the techniques on our data is given in Section 3. Section 4 concludes this paper.

1. Heart Failure Dataset

As a heart failure dataset, a group of 2032 patients ($\boldsymbol{V}$) classified into two levels of patient status (C) and described by 9 attributes ($\boldsymbol{A}$) as queries about clinical findings and physiological measurements is used. Patients and data about them are from Hull LifeLab which is large, epidemiologically representative, information-rich clinical data [2]. Its purpose is studying patients with heart failure so that its definition and diagnosis, its natural history, its mechanisms and markers of progression, the associated costs to health services and society, and the delivery of proven treatment to patients are improved.

The description of attributes and their summary can be seen in Table 1. Describing attributes ($\boldsymbol{A}$) are defined as $\boldsymbol{A} = \{A_1; \ldots; A_k; \ldots; A_9\}$. If A_k is a categorical attribute, $A_k = \{a_{k,1}; \ldots; a_{k,l}, \ldots; a_{k,l_k}\}$. Class attribute C is used to classify patients into two possible categorical values c_1 and c_2 (*alive* and *dead*). It expresses if a patient is alive or dead within a period of time after the risk estimation. In the heart failure dataset there are 1512 patients (74.4094%) classified as *alive* and 520 patients (25.5906%) classified as *dead*.

2. Weka and Data Mining Techniques

Weka (Waikato Environment for Knowledge Learning) is an open source data mining system implemented in Java [4]. Releasing Weka as an open source software

Attribute	Data type	Value range
Age (A_1)	Numerical	27 – 96
Blood Creatinine Level (A_2)	Numerical	37 – 1262
Blood Sodium Level (A_3)	Numerical	123 – 148
Blood Uric Acid Level (A_4)	Numerical	0.11 – 1.06
Height (A_5)	Numerical	1.2 – 1.96
NT-proBNP Level (A_6)	Numerical	0.89 – 18236
Pulse Rate (A_7)	Numerical	38 – 150
Sex (A_8)	Categorical	*female* ($a_{8,1}$)
		male ($a_{8,2}$)
Weight (A_9)	Numerical	29.8 – 193.8
Patient Status (C)	Categorical	*alive* (c_1)
		dead (c_2)

Table 1. Heart failure dataset.

and implementing it in Java are two factors which ensure that it remains maintainable and modifiable irrespective of the commitment or health of any particular institution or company.

Its aim is to provide a comprehensive collection of machine learning algorithms and data preprocessing tools to researchers and practitioners alike. It can cope with preprocessing and data analysis, classification models, association models, and evaluation metrics. There are three modes of Weka operation:

a) GUI,

b) command-line,

c) Java API.

Java API allows to make computer programs for solving classification tasks. In the weka.classifiers package, the most important class is *Classifier*. It is a general scheme for any classification model in Weka. *Classifier* contains two significant methods, *buildClassifier()* and *classifyInstance()*. The former is for building the classification model, the latter is for determining the value of class attribute C when the values of all attributes in $\boldsymbol{A}$ are known (i.e. classification).

We used two principally different data mining techniques for classification: a Bayes network classifier (BNC) and a decision tree classifier (DTC). They can be described as follows. A BNC is based on a Bayesian network which represents a joint probability distribution over a set of categorical attributes [1]. Since it is

a distribution over a set of categorical attributes, numerical attributes in $\boldsymbol{A}$ are discretized and transformed into categorical.

It consists of $\langle G; \Theta \rangle$, a directed acyclic graph G consisting of nodes and arcs and conditional probability tables $\Theta = (\theta_{A_1}; \ldots; \theta_{A_k})$. The nodes represent attributes in $\boldsymbol{A}$ and attribute C whereas the arcs indicate direct dependencies. The Bayesian network allows the computation of the (joint) posterior probability distribution of any subset of unobserved assignments of values to attributes in $\boldsymbol{A}$, which makes it possible to use for determination of $c_j \in C$. A DTC consists of a decision tree which is generated on the basis of instances in $\boldsymbol{V}$ [3].

There are two types of nodes in the decision tree:

a) the root and internal nodes (associated with an attribute $A_k \in \boldsymbol{A}$);

b) leaf nodes (associated with a $c_j \in C$).

Basically, each non-leaf node has an outgoing branch for each possible value $a_{k,l} \in A_k$, $A_k \in \boldsymbol{A}$ is an attribute associated with the node. Numerical attributes $A_k \in \boldsymbol{A}$ are discretized. Value $c_j \in C$ is determined for a new instance using a decision tree, beginning with the root, successive internal nodes are visited until a leaf node is reached. At the root node and at each internal node, a test is applied. The outcome of the test determines the branch traversed, and the next node visited. Value $c_j \in C$ for the instance is simply c_j of the final leaf node.

3. Experimental Results

The main purpose of the experiments is to compare the performance of a Bayes network classifier and a decision tree classifier on our heart failure dataset. The classifiers are implemented in Weka [4]. The performance of the classifiers is measured with sensitivity $\frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fn}}$ and specificity $\frac{\mathrm{tn}}{\mathrm{tn}+\mathrm{fp}}$.

In the formulas, tp / fp / fn / tn is the number of true positives / false positives / false negatives / true negatives. "*C is alive*" is considered negative, "*C is dead*" is considered positive. Values tp, fp, fn and tn are computed during 10-fold cross-validation. In 10-fold cross-validation, the (fuzzified) data is partitioned into 10 folds of patients. The partition is random, but all folds contain roughly the same proportions of alive and dead patients. Of the 10 folds, a single fold is retained as the testing data for evaluation of discovered knowledge, and the remaining 9 folds are used as the learning data. The learning data is analyzed by the algorithm for the purpose of discovering the knowledge. The cross-validation process is repeated 10 times, with each of the 10 folds used exactly once as the testing data.

The results of our experiments are given in Table 2. Bayes denotes a Bayesian network classiffier implemented in Weka as class BayesNet. DTC is a decision tree classifier implemented in Weka as class J48. Sensitivity is sensitivity in percentages, Specificity is specificity in percentages and Sum is the sum of Sensitivity

Classifier	Sensitivity	Specificity	Sum
BNC	24.038	95.37	119.408
DTC	40.96	87.90	128.86

Table 2. Experimental results.

and Specificity. It is very important to avoid classification of dead patients as alive as it would lead to life-threatening situations. Many alive patients classified as dead ones increase the costs considerably. The classifier based on decision tree (DTC) with sensitivity 40.96 and specificity 87.90 outperforms the other one.

4. Conclusions

Open source machine-learning Java software tool Weka was described with focus on classifiers such as a Bayes network classifier and a decision tree classifier. It was employed on our heart failure dataset from Hull LifeLab. For evaluation purposes, two different criteria were considered:

a) minimization of life-threatening costs;

b) minimization of treatment costs.

Life-threating situations appear when patients who are at risk of death soon are considered alive, which is measured by sensitivity. These situations should be minimized and so sensitivity should be maximized. Costs are increased when patients with a low risk of death are treated as if they could die soon, which is measured by specificity. For the costs to be minimized, specificity should be maximized. For our Hull LifeLab data about 2032 patients, a decision tree classifier with sensitivity 40.96% and specificity 87.90% was found, which is significantly better than the other tested classifier (a Bayesian network classifier). In the future it can be useful to consider other classifiers such as a group of fuzzy rules since it would allow us to address vagueness and ambiguity present in the dataset.

Acknowledgment. This work was supported by a HEIF-5 funded project. The authors would like to thank the University of Hull, UK for its support.

References

[1] BAESENS, B., EGMONT-PETERSEN, M., CASTELO, R., VANTHIENEN, J.: *Learning Bayesian network classifiers for credit scoring using Markov Chain Monte Carlo search*, Proceedings of International Congress on Pattern Recognition, Montreal, Canada, IEEE Computer Society, 2010, pp. 49-52,

[2] CLINICAL EFFECTIVENESS AND EVALUATION UNIT OF THE ROYAL COLLEGE OF PHYSICIANS: *Managing Chronic Heart Failure: Learning from Best Practice*. The Lavenham Press Ltd, Sudbury, Suffolk, UK, 2005,

[3] GAROFALAKIS, M., HYUN, D., RASTOGI, R., SHIM, K.: *Building decision trees with constraints*, Data Mining and Knowledge Discovery, Volume: 7, Number: 2, Pages: 187-214, Year: 2003, ISSN: 1384-5810,

[4] HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P., WITTEN, I. H.: *The WEKA data mining software: An update*, ACM SIGKDD Explorations Newsletter, Volume: 11, Issue: 1, Year: 2009, Pages: 10-18, ISSN: 1931-0145,

[5] LÓPEZ-SENDÓN, J.: *The heart failure epidemic*, Medicographia, Volume: 33, Number: 4, 2011, pp. 363-369,

Contact addresses

Ing. Ján Boháčik, PhD., EUR ING, Department of Computer Science, Faculty of Science, University of Hull, Cottingham Road, Hull, HU6 7RX, United Kingdom and Department of Informatics, Faculty of Management Science and Informatics, Univerzitná 8215/1, 010 26 Žilina, Slovak Republic,
E-mail address: J.Bohacik@hull.ac.uk

Dr. C. Kambhampati, Department of Computer Science, Faculty of Science, University of Hull, Cottingham Road, Hull, HU6 7RX, United Kingdom,
E-mail address: C.Kambhampati@hull.ac.uk

Dr. Darryl N. Davis, Department of Computer Science, Faculty of Science, University of Hull, Cottingham Road, Hull, HU6 7RX, United Kingdom,
E-mail address: D.N.Davis@hull.ac.uk

RNDr. Miroslav Benedikovič, Department of Informatics, Faculty of Management Science and Informatics, Univerzitná 8215/1, 010 26 Žilina, Slovak Republic,
E-mail address: Miroslav.Benedikovic@fri.uniza.sk