NEUROEDA – INTERAKTIVNÍ WEBOVÁ APLIKACE PRO HODNOCENÍ NEUROLOGICKÝCH DAT

Ondřej Klempíř, Laura Shala, Radim Krupička

Anotace

Využití metod průzkumové analýzy dat (exploratory data analysis, EDA) je při hodnocení klinických dat v medicíně klíčovou fází. Vizualizační principy, modely poukazující na trendy vývoje či např. znázornění potenciálních závislostí, pomáhají k lepší interpretaci měření a v diagnostickém rozhodování. Počet dostupných moderních EDA balíků pro vývojáře v posledních letech roste v souvislosti s rozvojem oboru Data Science.

NeuroEDA je interaktivní webová aplikace pro hodnocení biomedicínských dat. Aplikace byla naprogramována ve statistickém jazyce R, v rámci reaktivního paradigmatu frameworku Shiny. Je dále rozvíjena a využívána Katedrou biomedicínské informatiky FBMI ČVUT ve spolupráci s Neurologickou klinikou 1. LF UK a VFN v Praze, především pro hodnocení pacientů s dystoniemi a Parkinsonovou nemocí. Zpracování uživatelských dat v tabulkové formě (.csv, excel) probíhá v serverové části.

Kromě základních popisných statistik, průzkumových grafů a shlukové analýzy, které jsou vhodné i pro hodnocení velkých dat, nabízí aplikace metody pro robustní a neparametrickou analýzu. Ty jsou v neurologii obzvlášť vhodné. Typicky z důvodu malých počtů a vlivných pozorování. Dále kvůli častému nesplnění dalších statistických předpokladů.

Mezi její výhody patří snadná rozšiřitelnost o nové R balíky a rychlá odezva ve webových prohlížečích. Uživatelské interaktivní prostředí umožňuje práci s funkcemi jazyka R bez znalosti skriptování.

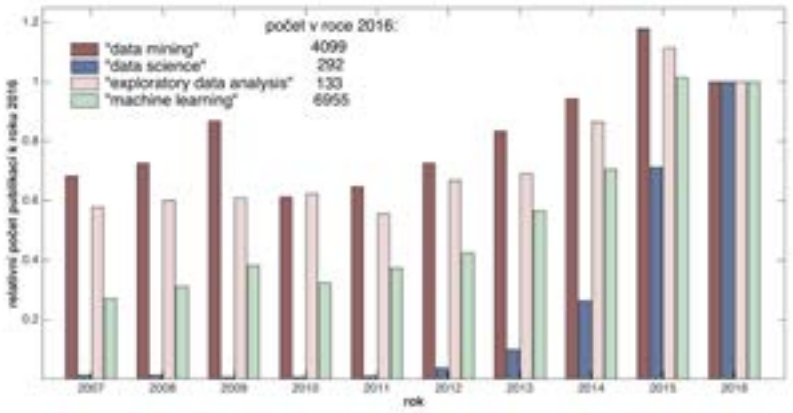
Klíčová slova

průzkumová analýza dat, analýza biomedicínských dat, neuroinformatika, jazyk R, Shiny

1 Úvod

Průzkumová analýza dat (EDA, exploratory data analysis) je jako pojem systematicky studována od doby statistika Johna W. Tukeyho. V jeho knize (1977) byla EDA definována jako soubor statistických metod a postupů pro hledání zajímavých hypotéz a vztahů v datech [1]. Jednalo se tehdy především o grafické techniky reprezentace dat: krabicové grafy, histogramy, bodové grafy, případně ručně vypočtenou analýzu hlavních komponent aj. Mnoho základních popisných i pokročilých počítačových technik patřících pod EDA bylo v průběhu času adaptováno do data miningu a analytiky velkých dat. Důkladná analýza stavu vyplněnosti a rozptylů hodnot, korelačních vztahů a např. diskriminativnosti skupin v datech hraje naprosto klíčovou roli pro předzpracování a následnou tvorbu popisných a predikčních modelů. V medicíně je význam EDA ještě vyšší, typicky z důvodu malých počtů či vlivných pozorování.

Výzkum a aplikace metod EDA se neustále rozvíjí a tento rozvoj se odráží i v prudkém nárůstu důležitosti oblastí spojených s počítačovou analýzou dat a datovou vědou



Obrázek 1 – Roční vývoj počtů vědeckých článků ve Web of Science souvisejících s frázemi z oblasti Analýzy a zpracování dat za posledních 10 let (relativně k roku 2016) (vlastní zpracování dle WoS)

obecně (Obr. 1). Počet vědeckých článků napříč všemi oblastmi vědy v citační službě Web of Science každým rokem roste. Z grafu je patrné, že s tímto růstem souvisí i zvyšování počtů článků přímo spojených s tématikou analýzy dat za posledních 10 let. Jako medicínské příklady rozvoje EDA v poslední době lze uvést aplikaci pro elektronické medicínské záznamy [2], text mining v porodnictví [3], nebo jako metodiku v neurovědách [4].

I přes matematickou povahu a doporučované postupy je moderní EDA jistou formou umění a kreativity autora/analytika [5]. Kreativitu autorů a jejich přínos v podobě různých balíčků, nástrojů a knihoven lze zakomponovat do prací programátorů biomedicínského softwaru. Vytvoření přehledného, snadno interpretovatelného grafu/dashboardu je v mnoha případech užitečnější než induktivní analýza založena na uvádění p-hodnot. Současné statistické studie poukazují na nadbytečné či zautomatizované problémové používání statistické významnosti dle p-hodnot v oblasti medicíny a psychologie [6, 7, 8, 9]. Důkladná vizuální analýza (visual mining) se tudíž zdá být dobrou alternativou k induktivní statistice [10].

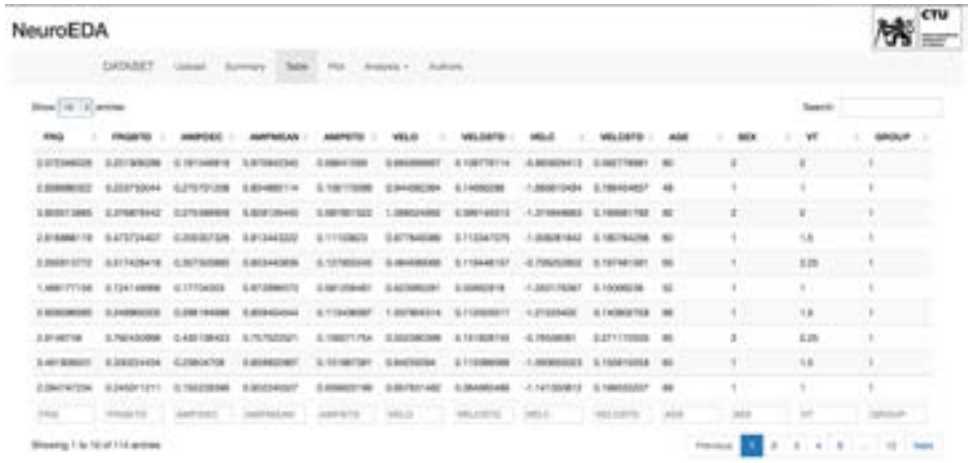
Příspěvek představuje interaktivní webovou aplikaci NeuroEDA, která moderní balíčky pro EDA a tvorbu modelů implementuje a využívá je pro hodnocení neurologických dat. Katedra biomedicínské informatiky FBMI ČVUT při analýze heterogenních neurologických dat dlouhodobě spolupracuje s Neurologickou klinikou 1. LF UK a VFN v Praze (např. data z měření blízkou infračervenou spektroskopií (NIRS), transkraniální magnetickou stimulací (TMS), kamerovými systémy či mikroelektrodových záznamů (MER)).

2 Aplikace NeuroEDA

Aplikace vznikla z důvodu potřeby integrujícího prostředí, s novými statistickými metodami, pro hodnocení biomedicínských dat KBI FBMI ČVUT. Byla naprogramována v open source programovacím jazyce R, který je významným zástupcem na poli statistických výpočtů. Jádro tvoří framework Shiny, který je založen na paradigmatu reaktivního programování [11, 12]. Reaktivní programování bylo především navrženo jako způsob, jak zjednodušit tvorbu interaktivních uživatelských rozhraní. Aplikace Shiny se skládá ze dvou základních částí, uživatelské (ui.R) v podobě webové stránky, a serverové (server.R). V Shiny je reaktivita zprostředkována reaktivními vstupy a výstupy. Typickým vstupem je uživatelský požadavek ve webovém rozhraní. Například výběr z několika možností formuláře, vyplnění hodnoty textového pole nebo kliknutí na tlačítko. Tyto akce nastaví hodnoty, na které aplikace okamžitě reaguje v podobě výstupu (zobrazení grafu, operace s tabulkou aj.). Lze ji spustit na lokálním serveru a pracovat ve webovém prohlížeči. Uživatel aplikaci ovládá pomocí uživatelského rozhraní a tím dává požadavky serverové části, která provádí výpočty a aktualizuje zobrazení výsledků.

2.1 Načítání dat a základní operace s datasetem

Aplikace umožňuje importovat data uživatele ve formátu .csv. Výběr souboru probíhá ze souborového systému. Lze volit mezi typem oddělovače (středník/čárka/tabulátor) a zobrazením hlavičky. Po nahrání datasetu se reaktivně zobrazí základní informace o souboru. Uživatel se může přepnout do zobrazení shrnujících statistik o jednotlivých atributech, případně se zaměřit na data jako tabulku s možnostmi stránkování, řazení, filtrování dat a vyhledávání (Obr. 2).



Obrázek 2 – Pohled na prostředí aplikace NeuroEDA s načtenými daty a jejich tabulkovým zobrazením

2.2 Průzkumové metody

Pro průzkumovou analýzu datasetů bylo implementováno několik metod:

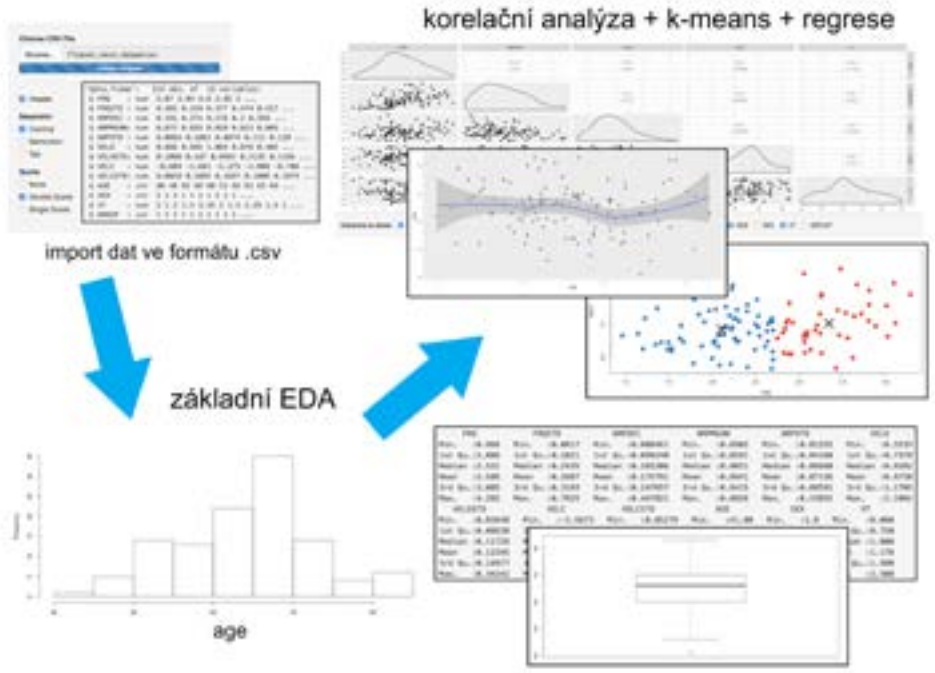
- oddíl grafů prostřednictvím balíku „ggplot2“: krabicový graf (boxplot), histogram, bodový graf (scatter plot),
- interaktivní korelační analýza + odhad jádrové hustoty rozdělení prostřednictvím balíku „ggally“,
- k-means algoritmus pro hledání přirozených kompaktních shluků v datech,
- jednoduchá regresní analýza s vysvětlující a vysvětlovanou proměnnou dle metody nejmenších čtverců,
- lokálně váhovaná vyhlazovací regrese (LOESS) prostřednictvím balíku „ggplot2“ – dokáže zachytit nelineární trend, vhodné pro detekci rychlý poklesů či jiných intervencí,
- robustní regrese prostřednictvím balíku „robust“ – menší citlivost na odlehlá pozorování, lineární modely s menším počtem pozorování reprezentuje lépe než odhad metodou nejmenších čtverců.

3 Klinické využití

Aplikace byla testována na několika veřejně dostupných datasetech různých rozměrů (např. iris, mtcars z balíku „datasets“). Dále byla použita pro průzkumovou analýzu klinického neurologického datasetu z měření kamerovým systémem. Jedná se záznamy parametrů periodického pohybu ruky, neboli finger tappingu (FT). Jde o opakované spojení palce a ukazováku s následným maximálním oddálením, nejrychleji jak subjekt dokáže. Měřeno bylo ve skupině zdravých (N = 59) a nemocných s Parkinsonovou nemocí (N = 55). Přehled vybraných zaznamenaných parametrů je uveden v tabulce (viz Tab.1).

název parametru	význam
GROUP	příslušnost do skupiny: = 1 (nemocný)
SEX	pohlaví: = 1 (žena)
VT	hodnota expertního posouzení
FRQ	průměrná frekvence kmitání prstů [Hz]
FRQSTD	směrodatná odchylka FRQ
AMPDEC	pokles amplitudy oddálení prstů
AMPMEAN	průměrná amplituda oddálení prstů
AMPSTD	směrodatná odchylka amplitudy
VELO	rychlost otvírání prstů

Tabulka 1 – Název a popis vybraných parametrů z měření finger tappingu



Obrázek 3 – Procesní mapa aktuální verze aplikace NeuroEDA

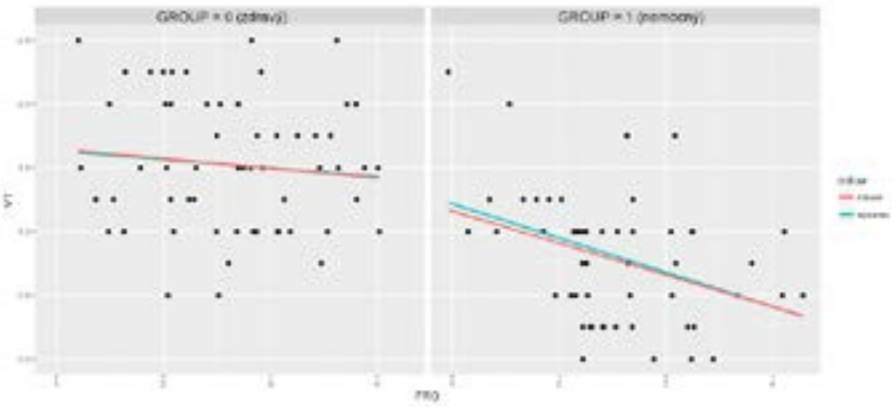
Možnosti práce s neurologickým datasetem v aplikaci NeuroEDA jsou znázorněny procesní mapou. Z obrázku je např. dobře patrné, že na základě parametrů FRQ a VELO lze poměrně spolehlivě automaticky odlišit zdravé a nemocné algoritmem k-means (viz Obr. 3).

Praktický význam má příklad vizuálně nalezené hypotézy regresními metodami. Bylo zjištěno, že existuje rozdíl ve skupině zdravých vs. nemocných při uvážení následujícího regresního modelu (Obr. 4):

vysvětlovaná proměnná (y): VT

vysvětlující proměnná (x): FRQ

Parametr frekvence tappingu (FRQ) má ve skupině nemocných podstatně větší vliv na to, jak rater (expert – hodnotitel) ohodnotí stav pacienta. Jinými slovy, u nemocných se rater mnohem více zaměřuje na pousouzení frekvence finger tappingu než ve skupině zdravých (u zdravých hodnotí pravděpodobně podle jiného parametru než FRQ). Z lineární a robustní regrese je zřejmé, že u nemocných má odhad směrnice modelové přímky zápornou hodnotu (tzn. existuje nějaký negativní vztah). Rater tudíž ohodnotí větší číslem, pokud je FRQ menší. Ve skupině zdravých je však směrnice přímky prakticky nulová.



Obrázek 4 – Parametr FRQ má na hodnocení raterem větší význam ve skupině nemocných

4 Další vývoj

Aplikace je stále ve fázi alfa verze a v současné době se pracuje na implementaci dalších funkcí a jejich testování. Kromě načítání dat ze souborového systému, uvažujeme o přímém napojení do databáze uchovávající data o různorodých neurofyziologických vyšetřeních, která je rovněž na KBI FBMI ČVUT vyvíjena. K modulu analýzy tabulkových dat připravujeme modul pro analýzu časových řad. Mezi hlavní funkce bude patřit výpočet frekvenčního spektra na pohyblivém okně, stanovení informačních příznaků a dalších funkcí dle potřeb nově naměřených signálů a lékaři specifikovaných hypotéz. Aplikace bude nasazena na dostupný webový server, ke kterému se bude možné připojit pomocí webového prohlížeče. Díky tomu lze pracovat s aplikací nejen v laboratorním počítači, ale i na tabletech či mobilních zařízeních bez nutnosti instalace na lokální stanici a v případě povolených bezpečnostních politik pracovat i vzdáleně mimo laboratoř.

5 Závěr

V práci jsme představili možnost tvorby biomedicínského softwaru pro klinické využití. Byla vytvořena webová aplikace, která implementuje celou řadu metod pro explorační analýzu dat. Velkou výhodou je rozšiřitelnost dle dostupných balíků pro skriptovací jazyk R. Aplikace disponuje rychlou odezvou i v případě většího datového souboru. Testována byla také funkčnost v aktuálních verzích webových prohlížečů (Google Chrome, Safari, Mozilla Firefox, Internet Explorer). Hlavní předností aplikace je robustní regrese, která standardně nebývá zahrnuta v dostupných komerčních statistických programech a v medicíně je velice potřebná. Uživatelské interaktivní prostředí umožňuje práci s funkcemi jazyka R bez znalosti skriptování, nabízí se tudíž využití netechnickými zaměstnanci klinik. Aplikace je v současné době aktivně využívána pro hodnocení heterogenních neurologických dat.

Poděkování

Tato práce byla podpořena grantem AZV č. 16-28119A: Analýza pohybových poruch pro studium mechanismů postižení u extrapyramidových onemocnění pomocí „motion capture" kamerových systémů.28119A

Literatura

[1.] TUKEY, John Wilder. Exploratory data analysis. Reading, Mass.: Addison-Wesley Pub. Co., c1977. ISBN 0201076160.

[2.] HUANG, Chih-Wei, Richard LU, Usman IQBAL, et al. A richly interactive exploratory data analysis and visualization tool using electronic medical records. BMC Medical Informatics and Decision Making. 2015, 15(1). DOI: 10.1186/s12911-015-0218-7. ISSN 1472-6947.

[3.] TAGAWA, Miki, Yoshio MATSUDA, Tomoko MANAKA, Makiko KOBAYASHI, Michitaka OHWADA a Shigeki MATSUBARA. Exploratory analysis of textual data from the Mother and Child Handbook using a text mining method (II): Monthly changes in the words recorded by mothers. Journal of Obstetrics and Gynaecology Research. 2017, 43(1), 100–105. DOI: 10.1111/jog.13178. ISSN 13418076.

[4.] MORI, Etsuro, Manabu IKEDA, Kenya NAKAI, Hideaki MIYAGISHI, Masaki NAKAGAWA a Kenji KOSAKA. Increased plasma donepezil concentration improves cognitive function in patients with dementia with Lewy bodies: An exploratory pharmacokinetic/pharmacodynamic analysis in a phase 3 randomized controlled trial. Journal of the Neurological Sciences. 2016, 366, 184–190. DOI: 10.1016/j.jns.2016.05.001. ISSN 0022510x.

[5.] PENG, Roger. The Art of Data Science. lulu.com, 2016. ISBN 978-1-365-06146-2.

[6.] Any Forward Progress on p-Values? [online]. 2017 [cit. 2017-02-11]. Dostupné z: https://www.r-bloggers.com/any-forward-progress-on-p-values/

[7.] VIDGEN, Bertie a Taha YASSERI. P-Values: Misunderstood and Misused. Frontiers in Physics. 2016, 4. DOI: 10.3389/fphy.2016.00006. ISSN 2296-424x.

[8.] The problem with p-values. AEON [online]. 2016 [cit. 2017-02-11]. Dostupné z: https://aeon.co/essays/it-s-time-for-science-to-abandon-the-term-statistically-significant

[9.] Experts issue warning on problems with P values. Sciencenews.org [online]. 2016 [cit. 2017-02-11]. Dostupné z: https://www.sciencenews.org/blog/context/experts-issue-warning-problems-p-values

[10.] WU, Chunshan, Yongqiang WENG, Qiaowei JIANG, Wenming GUO a Cong WANG. Applied research on visual mining technology in medical data. In: 2016 4th International Conference on Cloud Computing and Intelligence Systems (CCIS). IEEE, 2016, s. 229–233. DOI: 10.1109/CCIS.2016.7790259. ISBN 978-1-5090-1256-5.

[11.] Garrett Grolemund, The Shiny Cheat sheet, 2014, http://shiny.rstudio.com/articles/cheatsheet.html

[12.] Shiny – Reactivity paradigma. Shiny – R Studio [online]. 2016 [cit. 2017-02-12]. Dostupné z: http://shiny.rstudio.com/articles/reactivity-overview.html

Kontakt

Ondřej Klempíř

Laura Shala

Radim Krupička
Katedra biomedicínské informatiky
Fakulta biomedicínského inženýrství ČVUT nám. Sítná 3105
27201 Kladno
e-mail: *{ondrej.klempir, laura.shala, radim.krupicka}@fbmi.cvut.cz*