Exploračná analýza medicínskych záznamov

František Babič, Michal Vadovský, Ján Paralič

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky, Technická univerzita v Košiciach, Letná 9/B, 042 00, Košice, Slovenská republika

{frantisek.babic, michal.vadovsky, jan.paralic}@tuke.sk

Abstrakt. Medicínska diagnostika predstavuje komplexný proces pozostávajúci z množstva vstupov a potenciálnych závislostí, ktoré môžu v konečnom dôsledku ovplyvniť správnosť výsledku a následnú liečbu. Dátová analytika a príbuzné domény ako štatistika alebo umelá inteligencia môžu byť v tomto smere nápomocné, ak sú k dispozícii medicínske záznamy v elektronickej podobe. V našom prípade sme sa zamerali na jedno z vážnych civilizačných ochorení s názvom Metabolický syndróm a jeho výskyt u hypertenzívnych žien v menopauze. V rámci tejto úlohy sme už realizovali viacero predikčných experimentov, ale v tomto článku popíšeme práve analýzu potenciálne existujúcej závislosti medzi cieľovou diagnózou a skupinami vstupných faktorov určenými spolupracujúcim medicínskym expertom. Na tento účel sme použili logistickú regresiu a dosiahnuté výsledky sme overili prostredníctvom experta a existujúcich prác.

Kľúčové slová: medicínske zoznamy, logistická regresia, interval spoľahlivosti

1 Úvod

Podpora medicínskej diagnostiky predstavuje jednu z hlavných výziev, ktorá momentálne stojí pred oblasťou dátovej analytiky. K dispozícii máme metódy strojového učenia, umelej inteligencie, štatistiky alebo exploračnej analýzy, ktoré je možné aplikovať na medicínske záznamy dostupné v elektronickej podobe.

Metabolický syndróm (MS) sa definuje ako nenáhodný spoločný výskyt porúch metabolizmu cukrov súvisiacich s inzulínovou rezistenciou, centrálnou obezitou, dyslipidémiou spojenou so zvýšením hladiny triacylglycerolov a znížením lipoproteínov s vyššou denzitou, artériovou hypertenziou a ďalších faktorov, ktoré sa podieľajú na zvýšenom riziku ischemickej choroby srdca a cukrovky 2. typu [2]. O dôležitosti tejto problematiky svedčí aj fakt, že diagnostické kritériá pre MS spĺňa 20 % slovenskej populácie, čo potvrdila nedávna multicentrická skríningová štúdia s názvom „Prevalencia diabetes mellitus a metabolického syndrómu na Slovensku“ [3].

V našich výskumných aktivitách sa venujeme analýze dostupných medicínskych vzoriek, či už s cieľom vytvoriť čo najpresnejšie predikčné modely alebo pochopiť vstupné vzorky prostredníctvom metód exploračnej analýzy alebo štatistiky. V článku na minuloročných Dáta a Znalosti 2016 sme sa venovali identifikácii kľúčových fak-

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 166-170.

torov pre diagnostiku ochorenia s názvom Mierne kognitívne zhoršenie prostredníctvom vybraných štatistických testov a identifikácii kľúčových hodnôt pomocou ROC krivky a Youdenovej metódy [1]. V tomto článku popisujeme podobne orientovanú analýzu, ale zamerali sme sa na MS; konkrétne sme na skúmanie potenciálnych vzťahov medzi vstupnými faktormi a cieľovou diagnózou použili logistickú regresiu. Realizované experimenty nadväzujú na už našu ďalšiu publikovanú prácu [6].

2 Pochopenie dát

Hypertenzívne ženy v menopauze predstavujú rizikovú skupinu pre vývoj MS alebo kardiovaskulárnych ochorení [4]. Vzorka dát obsahuje informácie o 200 pacientkach z klinickej praxe vo veku od 47 do 59 rokov, 69 zdravých a 131 s potvrdenou diagnózou MS na základe kritérií medzinárodnej federácie diabetikov [5]:

- Základné kritérium – abdominálna obezita (obvod pásu nad 80cm) a k tomu aspoň dve zo 4 ďalších kritérií:
  - Glykémia nalačno nad 5.6 mmol/l alebo predtým diagnostikovaný diabetes mellitus 2. typu.
  - Triacylglycerol > 1.7 mmol/l
  - HDL cholesterol < 1.3 mmol/l
  - Zvýšené hodnoty krvného tlaku > 130/85.

Každý pacient je zároveň charakterizovaný hodnotami 62 faktorov, ktoré predstavujú potenciálne dôležité vstupy pre diagnostiku MS. Vybrané atribúty sú popísané v Tab.1.

Tab. 1. Vybrané vstupné atribúty

Názov	Popis
Kedy bola diagnostikovaná hypertenzia (v rokoch)	{<5, <5,10>, >10}
Regulácia hypertenzie	{áno, nie}
Menopauza a jej trvanie (v rokoch)	{nie, <1, <1,3>, >3}
Diagnostika cukrovky	{áno, nie}
Komplikácie pri liečení cukrovky	{nie, mac, mic}
Liečba cukrovky (orálne antidiabetiká, orálne antidiabetiká + inzulín, inzulín)	{nie, alebo, a/alebo, a}
Trvanie cukrovky (v rokoch)	{<5, <5,10>, >10}
Užívanie statínov viac ako 3 mesiace	{áno, nie}
Užívanie beta-blokátorov viac ako 3 mesiace	{áno, nie}
Užívanie antikoagulačných liekov viac ako 3 mesiace	{áno, nie}
Užívanie analgetík viac ako 3 mesiace a v minulom roku	{áno, nie}
Užívanie antibiotík viac ako dvakrát za rok	{áno, nie}
Užívanie lieku Metformin na liečbu cukrovky 2.typu	{áno, nie}

Užívanie ACE-inhibítorov viac ako 3 mesiace	{áno, nie}
Metabolický syndróm	<0, 1>

3 Analýza dát

Ako sme už spomenuli vyššie, na analýzu sme použili logistickú regresiu. Vykonali sme niekoľko experimentov s dvoma vybranými množinami dát. V prvom prípade sme brali do úvahy atribúty reprezentujúce diagnostiku a následné liečenie hypertenzie a cukrovky. V druhom prípade sme analyzovali vplyv anamnézy pacientiek a paralelného výskytu viacerých ochorení, čo dokumentuje užívanie rôznych typov liekov.

Dosiahnuté výsledky sú popísané pomocou parametrov ako z-hodnota, pomer šancí (OR) a interval spoľahlivosti (CI 95%). OR vyjadruje pomer šance zaradenia objektu do 1. cieľovej skupiny ak sa hodnota danej premennej zvýši o 1, pričom hodnoty ostatných premenných v modeli zostanú nezmenené, k pôvodnej šanci jej zaradenia do tejto cieľovej skupiny. V prípade binárnych premenných je interpretácia jednoduchšia, t.j. OR predstavuje pomer šance zaradenia objektu do 1. cieľovej skupiny, ak hodnota premennej = 1 ku šanci jeho zaradenia, ak hodnota premennej = 0, pri rovnakých hodnotách ostatných premenných.

diagnóza/ premenná	áno	nie
áno	a	b
Nie	c	d

$$OR = \frac{(a * d)}{(b * c)} \tag{1}$$

Hodnota OR v intervale <0, 1> znamená menšiu šancu zaradenia objektu do 1. cieľovej skupiny, hodnota >1 znamená opačnú situáciu a OR =1 definuje rovnakú šancu zaradenia do prvej alebo druhej cieľovej skupiny, t.j. nezávislosť.

Takisto sme použili McFaddenov koeficient determinácie, ktorý slúži na určenie kvality modelu logistickej regresie. Model s najvyšším R^2 je podľa tohto kritériá tým najlepším [7].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{2}$$

V prvom prípade najlepší model dosiahol hodnotu koeficientu determinácie 0.28, čo znamená že vytvorený model má určitú predikčnú silu smerom k cieľovej diagnóze, ale hodnota je bližšia skôr k 0 ako k 1. Hodnota 0 znamená, že daný model logistickej regresie nemá predikčnú silu. Na úrovni spoľahlivosti 95% môžeme identifikovať ako významný parameter len *trvanie menopauzy <1,3>* (z-hodnota = 2.18, OR = 11.2, CI = <1.2, 69.2>). Ak by sme hranicu posunuli na 90%, tak by pribudli parametre *trvanie menopauzy >3* (1.73, 6.9, <1.1, 44.5>) a *regulácia hypertenzie = normálna* (-1.9, 0.3, <0.1, 0.8>).

V druhom prípade najlepší model dosiahol hodnotu R^2 0.41, t.j. má väčšiu predikčnú silu ako predchádzajúci. Ako významné parametre na úrovni spoľahlivosti sme identifikovali užívanie *statínu* (5.9, 15.8, <7.3, 34.2>), *Metforminu* (3.5, 44.5, <7.6, 259.5>) a *beta-blokátorov* (2.04, 2.7, <1.2, 5.9>).

4 Záver

V našej práci sme sa venovali analýze dostupnej vzorky dát pomocou logistickej regresie, ktorej výsledok predstavuje zoznam významných vstupných premenných pre cieľovú diagnostiku Metabolického syndrómu. Dosiahnuté výsledky v oboch skupinách experimentov sme konzultovali s expertom a overovali prostredníctvom existujúcich prác a štúdií. Tieto potvrdili vyššiu prevalenciu MS u žien po menopauze, ale ju samotnú nepovažovali za rizikový faktor. Skôr sa zamerali na kardiovaskulárne faktory, ktorých včasná diagnostika môže prispieť k správnej liečbe a prevencii.

Literatúra

1. Vadovský M,, Babič, F., Muchová, M.: Systém na podporu rozhodovania pomocou jednoduchého a efektívneho pochopenie medicínskych záznamov. In: WIKT and DaZ 2016, Bratislava: STU (2016) 89 93.
2. Eckel, R.A., Grundy, S.M., Zimmet, P.Z.: The metabolic syndrome. Lancet 365 (2005) 1415 1428.
3. Galajda, P.: Metabolický syndróm, kardiovaskulárne a metabolické riziká. Via practica 4 (2007) 5 9.
4. Carr M. C..: The emergency of the metabolic syndrome with menopause. The Journal of Clinical Endocrinology & Metabolism 88 (2003) 2404 2011.
5. Alberti K. G., Zimmet P., Shaw J.: Metabolic syndrome – a new world-wide definition. A Consensus Statement from the IDF. Diabet Med 23 9(2006) 469 480.
6. Babič F., Majnarić L., Lukáčová A., Paralič J., Holzinger A.: On Patient's Characteristics Extraction for Metabolic Syndrome Diagnosis: Predictive Modelling Based on Machine Learning. In: Information Technology in Bio-and Medical Informatics, Springer, LNCS 8649 (2014) 118 132.
7. McFadden D.: Conditional Logit Analysis of Qualitative Choice Behavior. In: P. Zarembka (ed.), Frontiers in Econometrics, New York: Academic Press (1974).

Poďakovanie: Táto publikácia vznikla vďaka čiastočnej podpore projektu č.1/0493/16 financovaného Vedeckou grantovou agentúrou MŠVVaŠ SR a SAV (VEGA), projektom č. 025TUKE-4/2015 a č. 005TUKE-4/2017 financovaných Kultúrnou a edukačnou grantovou agentúrou MŠVVaŠ SR (KEGA).

Annotation:

The medical diagnostic is a complex process consisting of inputs and potential dependencies affecting the correctness of the outcome and subsequent treatment. Data analytics and related domains, such as statistics or artificial intelligence, can be ap-

plied to electronic medical records supporting the doctor's decision process. We focused on one of the major civilization diseases called Metabolic Syndrome and its occurrence in a specific target group. We oriented our experiments to investigate potentially existing dependence between the target diagnosis and the input factor groups determined by the cooperating medical expert. We used logistic regression and the results were evaluated by the expert as interesting and usable in clinical practice.