Použitie spracovaných záznamov reči pacientov pre určenie štádia Parkinsonovej choroby

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

michal.vadovsky@tuke.sk, jan.paralic@tuke.sk

Abstrakt. Lekársky postup diagnostikovania určitej choroby u pacientov je časovo zdĺhavý a veľmi náročný. Metódy dolovania v dátach môžu tento proces urýchliť a pomôcť tak lekárom pri rozhodovaní v zložitých situáciách. V prípade Parkinsonovej choroby (PCH) je najväčším problémom diagnostika prvotného štádia, pretože symptómy nie sú tak jednoznačné a ľahko pozorovateľné. Preto sme sa v tomto článku zamerali na určenie štádia PCH z dát zaznamenávajúcich rečové signály pacientov pomocou rozhodovacích stromov (C4.5, C5.0, CART). S cieľom zlepšenia klasifikačných modelov sme použili aj metódy RandomForest, Bagging a Boosting. Odhad presnosti modelov bol realizovaný použitím k-násobnej krížovej validácie a validácie s vynechaním jedného záznamu (Leave-one-out). Okrem toho sme vykonali aj experimenty s odstránením kolinearity v dátach vypočítaním inflačného faktoru rozptylu (VIF) za účelom zvýšenia presností modelov.

Kľúčové slová: štádium Parkinsonovej choroby, reč, dolovanie v dátach

1 Úvod

Parkinsonova choroba (PCH) [1] je veľmi vážne neurologické ochorenie, na ktoré dodnes neexistuje žiaden liek. Hlavnou príčinou vzniku ochorenia je odumieranie nervových buniek, ktoré produkujú v mozgu dôležitú chemickú látku s názvom dopamín [2]. Medzi prvotné príznaky u ľudí trpiacich PCH patrí stuhnutie svalstva, problémy s rečou (dysfónia), pohybom alebo písaním (dysgrafia).

Pre meranie štádia ochorenia sa vytvorila jednotná škála hodnotenia PCH s názvom UPDRS (Unified Parkinson´s Disease Rating Scale) [3], ktorej celkové skóre je získané vyhodnotením dotazníka skladajúceho sa zo štyroch častí: I. myslenie, správanie a nálada; II. aktivity bežného života, III. vyšetrovanie hybnosti; IV. komplikácie liečby v poslednom týždni. Na základe výsledkov UPDRS je možné pomocou modifikovanej stupnice štádia (podľa Hoehnonovej a Yahra) rozdeliť pacientov do 8 štádií PCH: 0 (bez príznakov ochorenia), 1 (jednostranné príznaky), 1.5 (jednostranné a axiálne postihnutie), 2 (obojstranné postihnutie bez poruchy rovnováhy), 2.5 (obojstranné postihnutie s miernou poruchou rovnováhy, schopnosť vyrov-

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 215-220.

nať postoj), 3 (mierne až stredné obojstranné postihnutie, sebestačný), 4 (ťažká nespôsobilosť, schopný chodiť alebo stáť bez pomoci), 5 (odkázaný na vozík alebo pripútaný k lôžku, vstávanie s pomocou).

V našom článku sme sa zamerali na určenie štádia PCH pomocou transformovaných dát zo zvukových záznamov reči. Diagnostika PCH v skorom štádiu je veľmi náročná, pretože príznaky sú nejednoznačné a ťažšie rozpoznateľné. Preto sme sa v tomto článku snažili zistiť, akú najvyššiu presnosť dokážeme získať pomocou modelov klasifikujúcich záznamy pacientov do 8 rôznych tried. V prípade získania vysokých presností by mohli byť tieto modely implementované do systémov na podporu diagnostiky PCH pre lekárov.

2 Prehľad súčasného stavu

Keďže je PCH veľmi časté ochorenie a stále na ňu neexistuje liek, tak sa množstvo výskumníkov zameriava na diagnostiku tohto ochorenia priamo z prvotných symptómov, napr. reč alebo písmo. V publikácii [4] sa G. Yadav s kol. zameriavali na reč pacientov, kde pre vytvorenie modelov pre klasifikáciu do dvoch tried (1 – pacient s PCH, 0 – zdravý pacient) využil 3 metódy dolovania v dátach. Pre ich porovnanie a vyhodnotenie použili 10-násobnú krížovú validáciu, pričom najlepšie výsledky dosiahli pomocou SVM (76%), nasledovala metóda rozhodovacích stromov (75%) a logistická regresia (64%). A. Tsanas s kol. [5] sa venovali predikovaniu numerickej hodnoty UPDRS (0-176). Zozbierané dáta obsahovali taktiež rečové signály pacientov a okrem celkovej hodnoty UPDRS sa zamerali aj na predikciu škály hybných (motorických) funkcií pacienta (0-108). Použili pritom metódy troch lineárnych regresií a jednej nelineárnej regresie. Podľa dosiahnutých výsledkov dokázali predikovať motorické hodnoty UPDRS približne v rozmedzí 6 bodov a celkové UPDRS v rozsahu 7.5 bodov. Tieto výsledky odrážajú najlepší odhad chyby predikcie pri 1000 spusteniach 10 násobnej krížovej validácie. Konečné predpovede hodnôt pomocou modelov sú veľmi blízke lekárskym pozorovaniam na klinike.

3 Pochopenie a príprava dát

Dáta, s ktorými sme pracovali sú voľne dostupné na UCI Machine Learning Repository [6] a skladajú sa z celej rady biomedicínskych hlasových meraní od 31 osôb (z toho 23 s PCH). Spolu bolo k dispozícií 195 záznamov (riadkov), pretože každý pacient mal v dátach viacero záznamov, ktoré boli brané nezávisle od seba. Reč pacienta bola transformovaná do 23 atribútov, ako napríklad: priemerná, maximálna a minimálna vokálna frekvencia, miera variability vo frekvencii (atribúty skupiny Jitter), miera variability v amplitúde (atribúty skupiny Shimmer), merania pomeru hluku a tónových zložiek v hlase (NHR, HNR) a množstvo ďalších.

K danému dátovému setu sme následne pripojili ďalšie atribúty, ktoré sme našli vo vedeckom článku [7]. Obsahovali dodatočné informácie o pohlaví, veku, štádiu PCH (8 úrovní) a počte rokov pacienta od kedy mu bola táto choroba diagnostikovaná. Po pridaní atribútu *Stage* sme odstránili atribút *status*, ktorý podával informáciu o tom, či

pacient trpí PCH (1) alebo nie (0). Dáta o pacientoch s najhorším 5. štádiom sme nemali k dispozícii. Po celkovej úprave dát (spojenie datasetov, odstránenie chýbajúcich hodnôt) sme pracovali so 189 riadkami a 25 stĺpcami.

4 Modelovanie

Pre vytváranie klasifikačných modelov sme si hlavne kvôli jednoduchej interpretácie pre lekárov zvolili iba metódu rozhodovacích stromov a jej algoritmy C4.5, C5.0 a CART, ktoré dosiahli z viacerých algoritmov najvyššie presnosti. Pre vyhodnotenie modelov sme použili metódy 10-násobnej krížovej validácie (**10 – CV**) a validácie kde sa vynecháva jeden záznam (*Leave One Out* – **LOO**). Pri LOO sa jedná o podobný spôsob vyhodnocovania ako v prípade *k*-násobnej krížovej validácie, ale klasifikátor sa buduje na *n*-1 záznamoch v dátovej množine a testuje sa len na 1 zázname. Tento proces sa následne opakuje *n*-krát. Pre vytvorenie modelov sme najprv vybrali všetky atribúty a neskôr sme očistili dáta od kolinearity (2 alebo viac atribútov sú navzájom silne závislé), ktorá môže zhoršiť presnosť modelov [8]. Namiesto vytvorenia korelačnej matice je lepším spôsobom pre posúdenie kolinearity vypočítať inflačný faktor rozptylu (**VIF**) pre každý atribút. Najmenšia možná hodnota VIF je 1, čo predstavuje úplnú absenciu kolinearity. Ako pravidlo platí, že ak hodnota VIF presahuje 5 alebo 10, tak hovoríme o problematickom množstve kolinearity. Odstránili sme preto atribúty s hodnotou VIF väčšou ako 5 a ich počet sa zredukoval z 25 na 13.

Tab. 1. Výsledky model při 10 – CV a LOO – CV

Výber atribútov	**10 - CV**			**LOO - CV**		
	CART	*C4.5*	*C5.0*	*CART*	*C4.5*	*C5.0*
Všetky atribúty	69,71%	79,88%	83,54%	71,96%	80,42%	81,48%
VIF < 5	67,66%	85,15%	**86,2%**	72,47%	80,95%	82,01%

Z výsledkov v Tab. 1 si môžeme všimnúť, že odstránenie atribútov s vysokou kolinearitou nám zabezpečilo vyššie presnosti modelov skoro vo všetkých prípadoch. Výnimkou bol algoritmus CART, pri ktorom sa presnosť zmenšila zo 69,71% na 67,66%, avšak tieto presnosti sú v porovnaní s algoritmami C4.5 a C5.0 aj tak podstatne nižšie. Najvyššie presnosti vo všetkých prípadoch dosiahol algoritmus C5.0, pričom pri odstránení atribútov s vysokou kolinearitou a použití 10 – CV sme dosiahli presnosť na úrovni **86,2%** pre klasifikáciu pacientov do 7 tried.

S cieľom vylepšiť presnosti vytvorených modelov sme sa rozhodli použiť metódy **RandomForest**, **Bagging** a **Boosting**, ktoré používajú stromy ako stavebné bloky na vytvorenie silnejších predikčných modelov. RandomForest vytvára viacero rozhodovacích stromov, kde v každom strome pri výbere testovacieho atribútu je braných do úvahy *m* náhodne vybraných atribútov z ich celkového počtu *p*. Výsledná klasifikácia do triedy je zvolená hlasovaním všetkých vygenerovaných stromov. Ak sa pri danom uzle berú do úvahy všetky atribúty *p,* vtedy hovoríme o baggingu. Podobným spôsobom funguje aj boosting, avšak každý rozhodovací strom berie do úvahy aj informá-

ciu z predchádzajúceho stromu [9]. Záznamom, ktoré boli v predchádzajúcom strome klasifikované nesprávne je v ďalšej iterácii priradená väčšia váha, vďaka čomu bude pri ďalšej iterácii kladený na tieto záznamy väčší dôraz. V publikácii [10] autori uvádzajú, že s rastúcim počtom vygenerovaných stromov sa zvyšuje už len výpočtová záťaž a rozdiely v presnostiach sú už veľmi malé. Ich analýza 29 datasetov ukázala, že pri vygenerovaní 128 stromov už nie je významný rozdiel v presnosti ako pri 256, 512, 1024, 2048 a 4096 stromoch. Preto sme počet rozhodovacích stromov nastavili na 50, 100 a 150 a pre výpočet presností sme si zvolili opäť 10 – násobnú krížovú validáciu. Pri pokuse orezať vygenerované rozhodovacie stromy sa úspešnosti zmenšili. Rovnaký problém nastal aj pri obmedzení maximálnej hĺbky rozhodovacieho stromu, preto sme tento vstupný parameter nemenili a nechali sme ho nastavený na predvolenej (defaultnej) hodnote.

Tab. 2. Výsledky metód RandomForest, Bagging a Boosting

Počet stromov	RandomForest	Bagging	Boosting
m = 50	87,25%	77,78%	93,65%
m = 100	86,73%	77,78%	95,24%
m = 150	86,73%	78,31%	**95,77%**

Vo výsledkoch v Tab. 2 vidíme, že najvyššie presnosti dosiahla jednoznačne metóda Boosting a to pri vygenerovaní 150 stromov (95,77%). S rastúcim počtom stromov presnosť klasifikácie rástla pri metódach Bagging aj Boosting, no naopak pre RandomForest trošku klesla. Algoritmus C5.0 pri 10 – CV dosiahol lepšie presnosti ako Bagging v tomto prípade a porovnateľné s metódou RandomForest.

Pre metódu Boosting, ktorá dosiahla najvyššiu presnosť na úrovni 95.77% sme si zobrazili na Obr.1 aj kontingenčnú tabuľku, ktorá porovnáva modelom predikované hodnoty atribútu Stage s tými, ktoré boli poznané a dané v testovacej množine.

	Observed Class						
Predicted Class	0	1	1.5	2	2.5	3	4
0	47	0	0	2	1	0	0
1	0	15	0	0	0	1	0
1.5	0	0	19	0	0	0	0
2	1	1	0	28	0	0	0
2.5	0	2	0	0	42	0	0
3	0	0	0	0	0	23	0
4	0	0	0	0	0	0	7

Obr. 1. Kontingenčná tabuľka pre metódu Boosting

Keďže pre vyhodnotenie modelov sme použili 10-násobnú krížovú validáciu, testovanie prebiehalo na 10 rôznych testovacích množinách. Každý prvok v kontingenčnej tabuľke na Obr. 1 je vypočítaný ako súčet zo všetkých získaných kontingenčných tabuliek, pomocou ktorých je jasné vidieť, pri akej predikcií došlo najčastejšie k chybe. Hlavným cieľom je maximalizovať hodnoty v matici na hlavnej diagonále, čo predstavuje správnu predikciu daného štádia Parkinsonovej choroby. Z celkového

počtu 189 záznamov dokázal model správne predikovať v 181 prípadoch, čo predstavuje 95.77% presnosť. Pre jednotlivé štádia sme dosiahli tieto presnosti (v zátvorke je vyjadrený pomer správne klasifikovaných záznamov ku všetkým záznamov pre dané štádium PCH): štádium 0 (47/48) – 97.92%, štádium 1 (15/18) = 83.33%, štádium 1.5 (19/19) = 100%, štádium 2 (28/30) = 93.33%, štádium 2.5 (42/43) = 97.67%, štádium 3 (23/24) = 95.83%, štádium 4 (7/7) = 100%.

Môžeme si všimnúť, že so 100% presnosťou dokázal model predikovať štádium 1.5 a 4. Naopak najnižšiu presnosť na úrovni 83.33% dosiahlo prvé štádium PCH, kde model vedel v 15tich záznamoch určiť správne štádium a pri 3 záznamoch došlo k chybe (v 1 prípade predikoval štádium 2 a v dvoch prípadoch štádium 2.5). V končenom dôsledku dosiahol model skoro pri všetkých štádiách úspešnosť nad 93%, jedinou výnimkou bolo štádium 1, kedy model dosiahol iba 83.33%.

5 Záver a budúca práca

V tomto článku sme sa zamerali na určenie štádia pacientov s PCH z ich reči pomocou metód dolovania v dátach. Už pri prvom experimente a odstránení kolinearity v dátach sme pomocou rozhodovacieho stromu a algoritmu C5.0 dosiahli presnosť na úrovni 86,2% (pred odstránením kolinearity – 83,54%). Použitím metódy Boosting, ktorá vytvorí viacero rozhodovacích stromov, sme našu presnosť dokázali zvýšiť až na 95,77% (pri $m = 150$), čo je vzhľadom na klasifikáciu záznamov až do 7 tried vysoká úspešnosť. Napr. v publikácii [4] pri binárnej klasifikácii (1 – pacient s PCH, 0 – zdravý pacient) s rovnakými dátami dosiahli autori najvyššiu presnosť len 76% použitím SVM. Rovnako binárnu klasifikáciu pacientov sme robili aj v našej predchádzajúcej publikácii [11] a najlepší výsledok na úrovni 91,43% sme dosiahli použitím algoritmu C4.5. Aj keď sa jednalo o zložitejšiu klasifikáciu pacientov (7 tried) v porovnaní s binárnou klasifikáciou (2 triedy), napriek tomu sme dosiahli vyššiu presnosť modelu pomocou metódy Boosting.

V budúcej práci by sme sa chceli zamerať na spracovanie hovorenej reči do rovnakých atribútov, aby bolo možno vytvoriť aplikáciu, kde si ľudia nahrajú svoju reč a dokážu sa testovať. Ďalej by sme sa chceli zamerať taktiež na dáta, ktoré sme získali od spoločnosti *mPower: Mobile Parkinson Disease Study*. Zaznamenávajú demografické údaje pacientov, ale aj údaje o ich hlase, chôdzi, pamäti a klikaní na obrazovku mobilu. Vďaka týmto dátam by sme mohli rozšíriť náš výskum na viaceré oblasti a príznaky PCH.

Poďakovanie. Táto publikácia vznikla vďaka podpore Vedeckej grantovej agentúry MŠVVaŠ SR a SAV projekt č. 1/0493/16 a Kultúrnou a edukačnou grantovou agentúrou MŠVVaŠ SR, projekt č. 025TUKE-4/2015.

Literatúra

1. De Lau, L. M., Bretler, M. M.: Epidemiology of Parkinson's disease, In: The Lancet Neurology, vol. 5, no. 6 (2006), pp. 525 – 535.

2. Cnockaert, L., et al.: Low-frequency vocal modulations in vowels produced by Parkinsonian subjects. In: Speech Comunication, vol. 50, no. 4 (2008), pp. 288-300.
3. Fans, S., et al.: Members of the UPDRS Development Committee Unified Parkinson's Disease Rating Scale. In: Recent developments in Parkinson's disease, vol. 2 (1987), pp. 153-163.
4. Yadav, G., et al.: Predication of Parkinson's disease using data mining methods: a comparative analysis of tree, statistical, and support vector machine classifiers. In: Indian Journal of Medical Sciences, vol. 65, no. 6 (2011), pp. 231-242.
5. Tsanas, A., et al.: Accurate Telemonitoring of Parkinson's Disease Progression by Noninvasive Speech Tests. In: IEEE Transactions on Biomedical Engineering, vol. 57, no. 4 (2010), pp. 884-893.
6. UCI Machine Learning repository: Center for Machine Learning and Intelligent Systems – Parkinsons Data Set. Available at: https://archive.ics.uci.edu/ml/datasets/Parkinsons.
7. Little, M. A., et al.: Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease. In: IEEE Transactions on Biomedical Engineering, vol. 56, no. 4 (2009), pp. 1015-1022.
8. James, G., et al.: An Introduction to Statistical Learning. Springer-Verlag New York, 2013.
9. Schapire, R., et al.: A decision-theoretic generalization of on-line learning and an application to boosting. In: Proceedings of the Second European Conference on Computational Learning Theory, vol. 904 (1995), pp. 23-37.
10. Oshiro, T. M., et al.: How Many Trees in a Random Forest? In: Machine Learning and Data Mining in Pattern Recognition, vol. 7376 (2012), pp. 154-168.
11. Vadovský, M., Paralič, J.: Predikcia Parkinsonovej choroby pomocou signálov reči použitím metód dolovania v dátach. In: WIKT & DaZ 2016, Bratislava: STU, 2016. s. 329-333. ISBN: 978-80-227-4619-9.

Annotation:

Utilizing processed records of patients' speech in determining the stage of Parkinson´s disease

The medical procedures for disease diagnostics are significantly demanding and time-consuming. Data mining methods can accelerate this process and assist doctors in making decisions in complex situations. In case of Parkinson´s disease (PCH), the diagnostics of the initial disease stage is the primary issue, since the symptoms are not so unambiguous and easily observable. Therefore, this article is focused on determining the actual stage of PCH based on the data recording signals of patient´s speech using decision trees (C4.5, C5.0 and CART). Methods such as RandomForest, Bagging and Boosting were also employed to improve the existing classification models. Estimation of model accuracy was achieved by using k-fold cross-validation and validation with omission of one record (Leave-one-out). In addition, experiments were also performed to remove collinearity in data by computing the Variance inflation factor (VIF) in order to increase the accuracy of the models.