DETECTION OF PATHOLOGICAL VERTEBRAE IN SPINAL CT UTILISED BY MACHINE LEARNING METHODS

Bohdan Tyshchenko

Master Degree Programme (2), FEEC BUT

E-mail: xtyshc00@stud.feec.vutbr.cz

Supervised by: Jiří Chmelík

E-mail: chmelikj@feec.vutbr.cz

Abstract: This paper presents a computer aided detection system to identify pathological vertebrae and to classify a type of pathology. Designed classification system is based on using neural network (NN), which performs classification step and on principal component analysis (PCA), which is used to reducing the original number of observation features.

Keywords: Neural network, Classification, CT, Machine Learning, Pathologies of spine, Principal Component Analysis, Vertebra.

1 ÚVOD

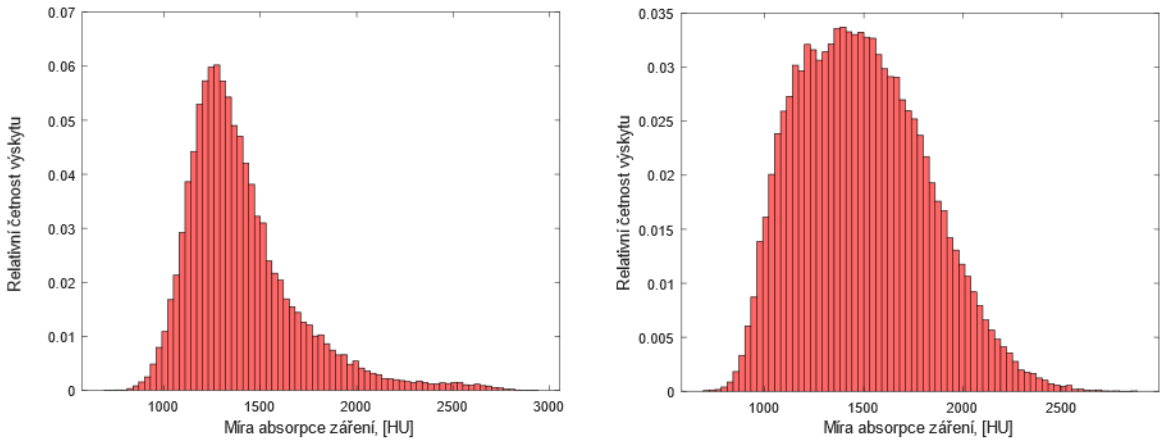
Automatizovaný systém detekce patologií je velmi důležitým nástrojem v lékařské diagnostice. Jelikož rychlost a přesnost stanovení diagnózy je často rozhodujícím faktorem, v dnešní době se zvyšuje tendence k použití umělé inteligence.

Cílem práce je implementovat techniku strojového učení, která by umožnila klasifikaci zdravotního stavu obratlového těla. Základními patologiemi obratle jsou zlomeniny, nádorová onemocnění a jejich kombinace. Charakter projevu těch patologií je velmi různorodý a popsat ho obecnými pravidly buď není možné anebo je moc náročné. A proto je nutnost použit strojové učení, které ze znalostní báze vyvodí pravidla, která popisuje každou klasifikační skupinu.

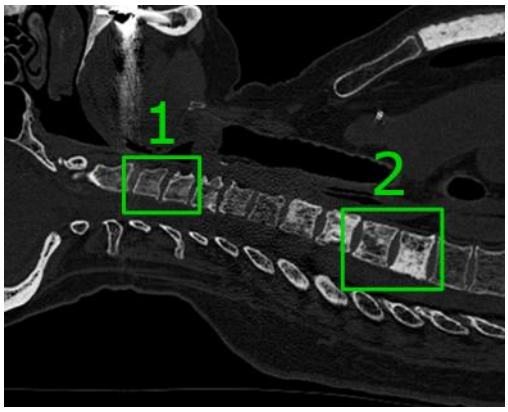
2 NÁVRH ALGORITMU

2.1 PŘÍZNAKY A ODŮVODNĚNÍ JEJICH VOLBY

Celá sada zvolených příznaků se skládá z pěti skupin. První čtyři jsou založeny na deskriptivní statistice. První skupina odhaluje změny v histogramu obratle způsobenými patologiemi. Na Obr. 1 lze vidět jak se mění tvar histogramu v případě nádorového onemocnění. Výskyt tohoto onemocnění v CT datech pozorujeme na Obr. 2. Statistické příznaky jako medián, modus, střední hodnota, koeficient špičatosti, šikmosti a jiné dokáží tyto změny zaznamenat. Druhá skupina je zaměřená na detekci strmých změn absorpce záření ve spongiózní části obratle, která je u zdravé tkáně relativně homogenní. Proto se nejprve provádí eroze binární masky, která odstraňuje kortikální část, a dále se statistické příznaky počítají z gradientního obrazu spongiózní části obratlového těla. V případě neprovedení eroze nebylo možné rozeznat, které vysoké hodnoty v gradientním obrazu jsou způsobeny změnou hustoty v kortikální části, a které nádory. Třetí a čtvrtá skupina příznaků jsou podobné dvěma prvním skupinám, ale na rozdíl od nich jsou vstupem do funkce výpočtu statistických znaků parametrické prostory lokálních směrodatných odchylek s různou velikosti okna. Tento přístup umožnuje odhalit drobné patologické jevy. Pátá skupina zkoumá geometrické vlastnosti obratlů a přispívá k detekci zlomenin a deformací. Zahrnuje příznaky jako délky hlavních os elipsoidu fitovaného na binární masku obratlového těla, výpočet těžiště a jiné.



Obrázek 1: Histogram hodnot míry absorpce záření zdravého (vlevo) obratle a postiženého metastázemi (vpravo) obratle.



Obrázek 2: Výskyt patologií v CT snímcích (1 – zdravý obratel, 2 – postižený metastázemi).

2.2 PŘÍPRAVA DAT

Pro vytvoření znalostní báze byla využita sada CT snímků páteře od 11 pacientů, zdravých a nemocných, která byla získána ústavem Biomedicinského inženýrství v rámci spolupráce s jinými vědeckými institucemi. Tato data byla předzpracována a obsahovala 217 vysegmentovaných obratlových těl. Každé obratlové tělo bylo předem posouzeno lékařským expertem a byl určen jeho patologický stav. Dále byla učební množina rozdělena do 4 tříd: zdravý obratel (1), deformovaný (2), s metastázemi (3), deformovaný a zároveň postižený metastázemi (4).

Etapa přípravy dat je představená řetězcem kroků zobrazeným na Obr. 3. Nejprve se provádí extrakce voxelů, které reprezentují jednotlivá obratlová těla, z CT snímků páteře a rozšíření (augmentace) učební množiny přidáním šumu, který byl vygenerován jako matice náhodných hodnot normálního rozložení s nulovou střední hodnotou a rozptylem 50 HU. Dále následuje výpočet příznaků, validace dat a příprava pro další využití neuronovou sítí. Validace dat zahrnuje kontrolu na přítomnost chybějících hodnot, abnormálních hodnot, objasnění jejich vzniku a případná úprava. Jelikož učební množina obsahuje kategorickou proměnnou (typ patologie), používá se One Hot Encoding pro převod této proměnné na binární formu, která umožní učení modelu. Konečně, výchozí učební množina obsahuje 434 případů a 60 příznaků.

Obrázek 3: Postup přípravy učební množiny.

2.3 POČÁTEČNÍ MODEL

Počáteční model využívá všechny příznaky. Jako klasifikační algoritmus slouží dopředná neuronová síť. Současně se navrhují dvě sítě s různými topologiemi: první je založena na důsledku z Kolmogorovy věty (topologie č. 1 – 60-60-121-4-4, kdy první a poslední čísla reprezentují počet neuronů ve vstupní a výstupní vrstvě, a čísla mezi nimi – počet neuronů ve vnitřních vrstvách), a druhá má heuristický charakter (topologie č.2 – 60-90-60-16-4-4). V obou modelech se jako přenosová funkce používá hyperbolická tangenciální sigmoidální funkce, optimalizační algoritmus – Scaled Conjugate Gradient. Před učením sítí se učební množina dělí do dvou množin (70% – trénovací, 30% – testovací) takovým způsobem, aby byl zachován stejný poměr výskytu klasifikačních tříd.

Návrh modelů byl realizován v prostředí Matlab pomocí Deep Learning Toolbox. 100×krát byla provedena inicializace a učení každé topologie. Kvůli náhodnému nastavení počátečních hodnot vah a prahů, probíhalo učení různě s různými dosaženými výsledky. Podle úspěšnosti klasifikace byla zvolena nejlepší síť. Kritériem úspěšnosti je poměr počtu správně klasifikovaných obratlů k celkovému počtu testovací množiny. Mezivýsledky jsou uvedeny v Tab. 1.

Název sítě	Topologie	Úspěšnost klasifikace nejlepší sítě, [%]	Průměr úspešnosti ze všech 100 sítí, [%]	Směrodatná odchylka úspešnosti ze všech 100 sítí, [%]
Síť č. 1	60-60-121-4-4	97,69	91,25	3,05
Síť č. 2	60-90-60-16-4-4	98,46	91,35	3,67

Tabulka 1: Porovnání navržených počátečních neuronových sítí.

2.4 ZLEPŠENÍ MODELU

Tento krok má za cíl snížení výpočetní náročnosti identifikováním nevýznamných příznaků a jejich následujícím vyloučením za podmínky zachování dosažené síly klasifikace. Pro tento účel je využita Principal Component Analysis (PCA), díky které jsou vyloučeny příznaky s minimální informační silou. Je důležité vést v patrnosti to, že přístup vychází z předpokladu, že nejvíce se měnící příznaky jsou klíčové, což nemusí vždy platit.

Nejprve se provádí standardizace příznaků učební množiny, dále následuje PCA (viz Obr. 4). Určovaní nejvýznamnějších příznaků se děje podle takto definovaného postupu: po standardizaci dat se zvolí tolik hlavních komponent s maximálními hodnotami variabilit, aby jejich součet byl více než 90%. Absolutní hodnoty koeficientů (hodnoty z vlastních vektorů), které se používají pro transformaci příznaku do PCA prostoru, se sčítají a podléhají standardizaci. Takto vzniká vektor hodnot přínosu každého příznaku. Všechny prediktory s hodnotou přínosu pod 0,2 vyloučeny (hodnota prahu byla zjištěna empiricky). Dále se provádí korelační analýza zbývajících příznaků. Pomocí Spearmanova korelačního koeficientu zjišťujeme páry korelovaných příznaků a vylučujeme jeden z nich. Ve výsledku zůstává 15 prediktorů.

Obrázek 4: Aplikace PCA.

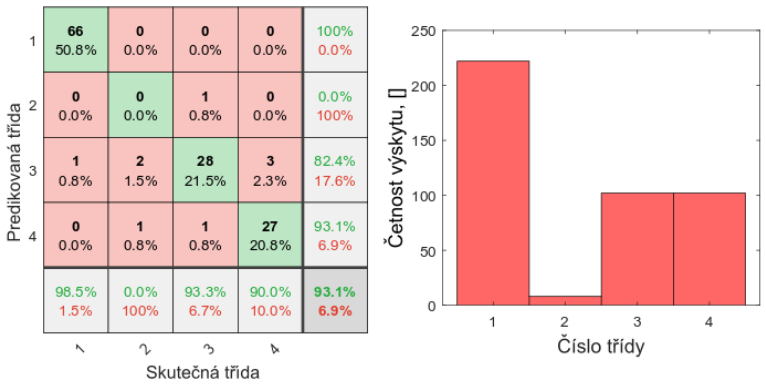
2.5 FINÁLNÍ KLASIFIKAČNÍ MODEL

Obdobným způsobem, jako v Kapitole 2.3, navrhujeme 2 sítě. Pro učební množinu s menším počtem příznaků upravujeme topologie sítí zmenšením počtu neuronů ve vrstvách.

Název sítě	Topologie	Úspěšnost klasifikace nejlepší sítě, [%]	Průměr úspěšnosti ze všech 100 sítí, [%]	Směrodatná odchylka úspěšnosti ze všech 100 sítí, [%]
Síť č. 1	15-15-31-4-4	93,08	88,31	2,47
Síť č. 2	15-23-15-8-4-4	92,30	87,08	5,8

Tabulka 2: Porovnání navržených neuronových sítí.

Lepší predikční vlastnosti má síť č. 1, avšak rozdíl není významný. Je vidět, že se po vynechání 75 % příznaku povedlo zachovat vysokou hladinu úspěšnosti klasifikace. Analýza přesnosti klasifikace jednotlivých tříd (Obrázek 4 [vlevo], poslední sloupec, hodnoty zvýrazněny zelenou barvou) ukazuje na souvislost mezi přesností a počtem dostupných případů v učební množině. Extrémem je třída číslo 2 – „pouze deformace“. K dispozici bylo jen 8 vzorků této třídy (Obrázek 4 [vpravo]), což je nedostatečné, aby se síť naučila.



Obrázek 5: Matice úspěšnosti klasifikátoru (vlevo), histogram diagnóz (vpravo).

3 ZÁVĚR

Cílem této práce je implementovat metodu strojového učení pro určování zdravotního stavu obratlového těla v CT snímcích páteře. Jako hlavní nástroj byla použita dopředná neuronová síť. Bylo dosaženo vysoké úspěšnosti klasifikace – 93.08 %. Pomocí PCA byl zmenšen počet příznaků, což snižuje celkovou výpočetní náročnost, za cenu ztráty 5 % přesnosti. Je důležité poznamenat, že základní míru chybovosti způsobila nevyváženost učební množiny. Za podmínky rozšíření učebních dat s přibližně stejným zastoupením výskytu jednotlivých tříd lze očekávat zvýšení úspěšnosti klasifikace.

REFERENCE

[1] HOLČÍK, J., KOMENDA, M. *Matematická biologie: e-learningová učebnice [online]*. Brno: Masarykova univerzita, 2015. ISBN 978-80-210-8095-9.

[2] JAN, Jiří. *Číslicová filtrace, analýza a restaurace signálů.* 2. upr. a rozš. vyd., Brno: VUTIUM, 2002. ISBN 80-214-1558-4.

[3] THEODORIDIS, Sergios a KOUTROUMBAS Konstantinos. *Pattern recognition.* 4th ed. Burlington, Mass.: Academic Press, 2009, xvii, 961 s. ISBN 978-1-59749-272-0.