STATISTICKÁ KLASIFIKACE DAT IMPULSNÍ OSCILOMETRIE POMOCÍ MATLABU

Petr Prášek

ČVUT FEL, K331 Katedra teorie obvodů

Abstrakt

Tento příspěvek popisuje použití bayesovské klasifikace při diagnostice plicních poruch. Analyzovaná data byla naměřena pomocí impulsní oscilometrie (IOS). Cílem popisovaného experimentu bylo nalezení vhodné reprezentace IOS dat a identifikace lidí trpících některou z poruch plicních funkcí. Řešení této úlohy není z důvodu velké variability lidské populace jednoduché. Proto bylo pro testování detektoru vytvořeno pět skupin lidí lišících se stupněm zdravotních problémů. Výsledky bayesovské klasifikace byly srovnány s klasifikací založenou na vyhodnocení euklidovské vzdálenosti.

Úvod

Klasifikačních metod existuje celá řada, lze je rozdělit na deterministické a statistické. Mezi ty druhé patří i bayesovské techniky. Celý proces má dvě fáze. V prvním kroku se vytvoří modely jednotlivých tříd, ve druhém kroku se provede vlastní klasifikace. Vzhledem k tomu, že IOS je metoda pro měření impedance dýchacích cest, je přirozené použít hodnoty impedance jako parametry pro klasifikaci. Dýchací systém lze modelovat pomocí elektrického obvodu a jeho prvky lze také zahrnout do klasifikace.

Bayesovská klasifikace

Nejdříve odvodíme Bayesův vzorec. Známý vztah pro násobení pravděpodobností a podmíněných pravděpodobností dvou jevů A a B lze zapsat jako

$$P(B) \cdot P(A|B) = P(A) \cdot P(B|A), \tag{1}$$

kde $P(A)$ a $P(B)$ jsou pravděpodobnosti jevů A a B, $P(A|B)$ a $P(B|A)$ jsou podmíněné pravděpodobnosti jevů A a B za podmínky uskutečnění jevů B a A. Předpokládejme nyní, že máme několik tříd popsaných jejich modely a vzorek naměřených dat. Jev A nyní bude představovat hypotézu h_i „vzorek patří do i-té třídy“. Jev B nahradíme vektorem naměřených dat $\mathbf{d}$. Místo pravděpodobností nyní budou ve vztahu vystupovat hustoty pravděpodobnosti p. Nyní můžeme přepsat rovnici (1) na tzv. Bayesův vztah

$$p(h_i|\mathbf{d}) = \frac{p(\mathbf{d}|h_i) \cdot P(h_i)}{p(\mathbf{d})} = \frac{p(\mathbf{d}|h_i) \cdot P(h_i)}{\sum_{i=1}^{S} p(\mathbf{d}|h_i) \cdot P(h_i)}, \tag{2}$$

kde $p(h_i|\mathbf{d})$ ukazuje pravděpodobnost hypotézy, že naměřená data (reprezentovaná parametry $\mathbf{d}$) náleží do i-té třídy. $P(h_i)$ je apriorní pravděpodobnost hypotéz h_i, $p(\mathbf{d}|h_i)$ je pravděpodobnostní mírou hypotézy h_i, $p(\mathbf{d})$ je úplnou pravděpodobností a S je počet tříd.

Modely tříd

Ze vhodně vybraných dat vytvoříme modely tříd. Pro každou z nich vypočítáme vektor středních hodnot všech parametrů použitého modelu

$$\mu_i[k] = \frac{1}{N}\sum_{j=1}^{N} d_j[k], \tag{3}$$

kde $i = 1, 2, \ldots, S$, $k = 1, 2, \ldots, M$ a N je počet měření v „trénovací“ skupině pro danou třídu; a kovarianční matici

$$\mathbf{C}_i = \begin{bmatrix} \sigma_1^2 & \ldots & \sigma_{1M} \\ \vdots & \ddots & \vdots \\ \sigma_{M1} & \ldots & \sigma_M^2 \end{bmatrix}, \tag{4}$$

kde σ_i^2 jsou rozptyly a σ_{ij} korelační koeficienty.

Vyhodnocení klasifikace

Hypotézy jsou vzájemně disjunktní a pro testovanému vzorku je přiřazena hypotéza s největší pravděpodobností. Pro vyhodnocení je zavedena vícerozměrná diskriminační funkce, používající vícerozměrné Gaussovské rozdělení [1], [7]

$$p(\mathbf{d}|h_i) = (2\pi)^{-\frac{M}{2}} |\mathbf{C}_i|^{-\frac{1}{2}} e^{-\frac{(\mathbf{d}-\mu_i)^T \mathbf{C}_i^{-1}(\mathbf{d}-\mu_i)}{2}}, \tag{5}$$

kde $|\mathbf{C}_i|$ je determinant kovarianční matice, μ_i je vektor středních hodnot parametrů i-té třídy a $\mathbf{d}$ je vektor naměřených dat. Za předpokladu rovnoměrného apriorního rozdělení lze po logaritmování výrazu a zanedbání konstant psát diskriminační funkci v následujícím tvaru

$$g_i(\mathbf{d}) = ln(p(\mathbf{d}|h_i)) = -ln|\mathbf{C}_i| - (\mathbf{d} - \mu_i)^T \mathbf{C}_i^{-1}(\mathbf{d} - \mu_i). \tag{6}$$

Diskriminační funkce je vyčíslena pro každou třídu a vzorek je přiřazen do třídy s největší hodnotou g_i.

Databáze signálů

Naměřená data použitá v experimentu pochází z IOS databáze vytvořené ve spolupráci s Nemocnicí Na Homolce v Praze. Celkem je v databázi přes 400 záznamů od 70 osob ve věku od 18 do 89 let. Vzhledem k tomu, že několik osob se účastnilo testu opakovatelnosti IOS měření, jsou v databázi desítky jejich záznamů. Navíc pro řadu pacientů se provádí měření opakovaně a často se provede i více testů v jednom dni. Z těchto důvodů je pro korektní vytvoření modelů tříd nutné vybrat jen některé záznamy. Proto byly do úvahy vzaty nejvýše první dva záznamy od jedné osoby z jednoho dne. Data z dalších dnů byly vypuštěny.

Z důvodu velké variability lidské populace bylo pro vytvoření modelu tříd vybráno celkem pět skupin lidí. Výběr by měl vyjadřovat rozdílnou úroveň zdravotních problémů. K rozdělení osob do skupin byla využita teoretická hodnota impedance spočítaná pro každou osobu před IOS testem. Výpočet je založen na pacientově věku, pohlaví, výšce

a hmotnosti [6]. Teoretická hodnota $Z(f)$ je orientační údaj. U lidí s hodnotou R_{5Hz} větší než 160 % teoretické se předpokládá některá z plicních dysfunkcí. Naopak, osoby s hodnotami menšími než 100 %, 120 %, atd. se považují za zdravé. Tímto způsobem byly vytvořeny první čtyři skupiny lidí (viz tab. 1). V těchto případech je mezi kategoriemi zdravý/nemocný pásmo neurčitosti. U poslední skupiny je hranice pouze jedna, a to 150 % teoretické hodnoty.

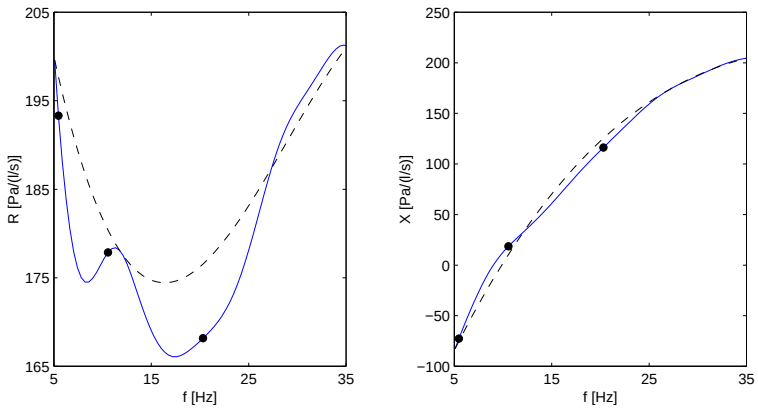
skupina	N	zdravý	nemocný
1	74	$< 100\,\%$	$160\,\% >$
2	92	$< 120\,\%$	$160\,\% >$
3	104	$< 130\,\%$	$160\,\% >$
4	112	$< 140\,\%$	$160\,\% >$
5	124	$< 150\,\%$	$150\,\% >$

Tabulka 1: Trénovací skupiny. N představuje počet záznamů v trénovací skupině.

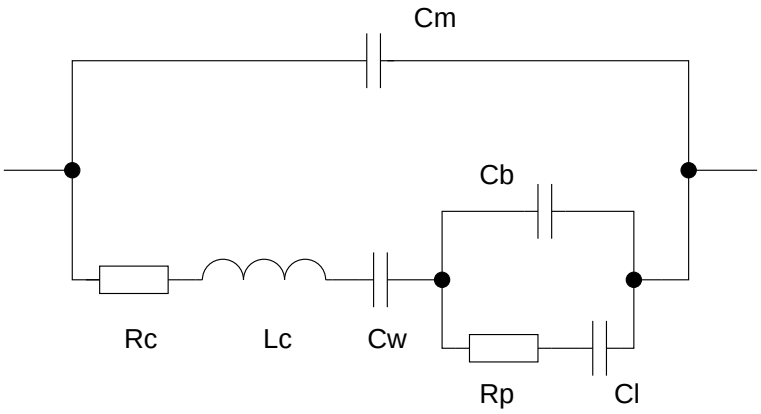
Reprezentace klasifikovaných dat

Pro klasifikaci bylo použito celkem pět modelů. Jejich parametry představovaly

- hodnoty $R(f)$ a $X(f)$, kde $f \in \{5, 10, 20\}$ Hz (obr. 1),
- hodnoty sedmi R, L, C prvků modelu plic (obr. 2, podrobnosti lze najít v [6]),
- kombinace předchozích dvou modelů,
- pouze jeden parametr, hodnota R_{5Hz},
- koeficienty a_i dvou polynomů třetího stupně $b(f) = a_1 f^3 + a_2 f^2 + a_3 f + a_4$ aproximujících impedanční charakteristiku (b představuje R, X).



Obrázek 1: Průběhy $R(f), X(f)$. Tečky vyznačují parametry prvního modelu, čárkovaně je naznačena aproximace polynomem.



Obrázek 2: Model plic.

Realizace a výsledky

Všechny potřebné výpočty byly provedeny s použitím MATLABU verze 5.3 a funkcí Statistics Toolboxu. Výsledky klasifikace byly srovnány s deterministickým přístupem, založeným na výpočtu euklidovské vzdálenosti v

$$v_i = \sqrt{\sum_{m=1}^{M} (p_i[m] - q_i[m])^2}, \tag{7}$$

kde $p[m]$ a $q[m]$ jsou vektory M parametrů modelů jednotlivých tříd a vzorek dat.

Výsledky klasifikace jsou v tab. 2. Algoritmus byl testován na stejných datech, která byla použita pro vytvoření modelů tříd, jde tedy o tzv. *uzavřený test*. Příklad dat je na obr. 3 (skupina 1 a model $R(f), X(f)$).

skupina	1		2		3		4		5	
model	e_{Eu} [%]	e_{Ba} [%]	e_{Eu} [%]	e_{Ba} [%]	e_{Eu} [%]	e_{Ba} [%]	e_{Eu} [%]	e_{Ba} [%]	e_{Eu} [%]	e_{Ba} [%]
R, X	14.90	5.40	11.97	7.60	10.58	9.62	11.61	9.82	12.10	12.10
$RLC + R, X$	14.90	2.70	4.34	11.97	7.69	10.58	11.61	8.04	12.10	8.06
RLC	16.22	9.46	14.13	9.78	13.46	9.62	13.39	13.39	14.52	12.90
R_{5Hz}	9.46	12.16	7.61	10.87	9.62	9.62	9.82	11.61	11.29	12.90
R_{pol}, X_{pol}	25.68	4.05	20.65	5.43	19.23	9.62	19.64	10.71	19.35	11.29

Tabulka 2: Chybovost euklidovské (e_{Eu}) a bayesovské (e_{Ba}) klasifikace v uzavřeném testu.

Závěr

Nejlepšího výsledku je dosaženo bayesovským přístupem pro kombinovaný model (impedanční charakteristika a R, L, C prvky). Dobré výsledky dávají také modely tvořené impedanční charakteristikou a její aproximací. Naopak, nejhorší výsledky jsou dosaženy při použití jediného parametru R_{5Hz}. Při Bayesovské klasifikaci je chybovost menší než u Euklidovské metody, s výjimkou R_{5Hz}.

Oznámení

Příspěvek byl zpracován v rámci výzkumného záměru číslo MSM210000012. Autor by rád poděkoval panu Hans-Jürgen Smithovi (Jaeger Company) za jeho pomoc při získání signálů.

Literatura

[1] Čmejla, R., Sovka, P. *Úvod do bayesovské klasifikace dat.* Akustické listy, 8(2), 2003

[2] Havránek, T. *Statistika pro biologické a lékařské vědy.* Academia, Praha, 1993

[3] Nucci, G., Polese, G., Rossi, A., Cobelli, C. *On-line estimation of respiratory parameters of lung mechanics in different pathologies.* Proceedings IEEE-EMBS, Chicago, 1997

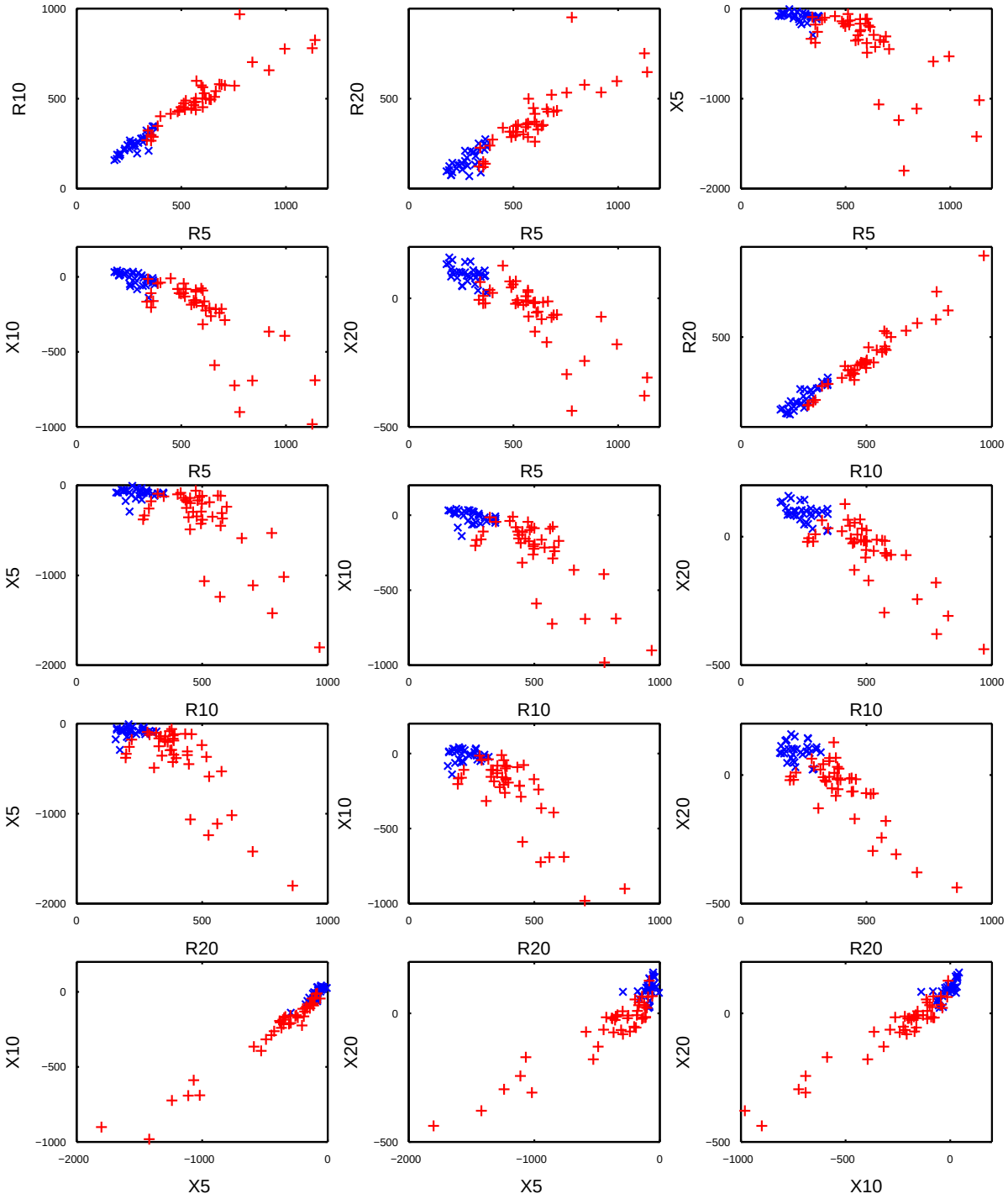
[4] Prášek, P. *Impulsní oscilometrie: spektrální vzdálenosti.* Analýza a zpracování signálů III, Vydavatelství ČVUT Praha, 2003

[5] Prášek, P. *Lung Disease Diagnostics Using Bayesian Classification Technique.* Proceedings of the Polish-Hungarian-Czech Workshop on Circuit Theory, Signal Processing, and Applications. CTU Prague, 2003

[6] Vogel, J., Smidt, U.: *Impulse oscillometry: analysis of lung mechanics in general practise and clinic, epidemiological and experimental research.* pmi Verlagsgruppe GmbH, Frankfurt am Main, 1994

[7] Zvára, K. *Biostatistika.* Karolinum, Praha, 1998

Kontaktní adresa: Ing. Petr Prášek
České vysoké učení technické v Praze, Elektrotechnická fakulta, Katedra teorie obvodů
Technická 2, 166 27 Praha 6, Tel.: +420 224 352 286
E-mail: xprasek@feld.cvut.cz



Obrázek 3: Hodnoty R, X, všechny kombinace dvojic. Nemocní jsou vyznačeni znaky $+$.