2006 © JČMF

APLIKÁCIA KLASIFIKAČNÝCH METÓD NA ANALÝZU VYDYCHOVANÝCH PLYNOV PRE DETEKCIU PĽÚCNYCH CHORÔB.

Katarína Cimermanová
katarina.cimermanova@gmail.com
Ústav merania, Slovenská akadémia vied, Dúbravská cesta 9, 841 04 Bratislava, Slovensko

ZHRNUTIE

Horúcim tipom na detekciu niektorých typov rakoviny je analýza vydychovaných plynov. Vo vydychovanom plyne sa dajú detekovať chemické látky, ktoré poskytujú informáciu pri diagnóze rakoviny. Na základe dát z poskytnutej databázy je potrebné vybrať čo najlepšiu klasifikačnú (diagnostickú) metódu, ktorá bude s najväčšou presnosťou zaraďovať viacrozmerné dáta do tried 'chorí' a 'zdraví'.
Príspevok sa zaoberá úpravou a klasifikáciou dát získaných analýzou vydychovaných plynov pomocou Fisherovho lineárneho klasifikátora (FLK), metódou oporných bodov (SVM) a doprednými neurónovými sieťami (ANN).

DATABÁZA

V databáze je 458 pozorovaných objektov X_i, z čoho 20 objektov patrí do triedy chorí (ω_1) a zvyšných 438 objektov do triedy zdraví (ω_2). V databáze sú objekty rozdelené do štyroch skupín, kde každá skupina bola meraná iným autorom. Objekty v jednotlivých skupinách patria do tej istej triedy. Každý objekt je vyjadrený 11 rozmerným vektorom, ktorého jednotlivé prvky predstavujú číselné vyjadrenia hustoty chemických látok vydýchnutého vzduchu s rôznymi molekulovými hmotnosťami.
Niektoré charakteristiky objektov sú v databáze nahradené výrazom NaN. Jedná sa o prípad, keď koncentrácia vo vydýchnutom a vdychovanom plyne bola rovnaká. Tento výraz je nutné pred klasifikáciou vhodne upraviť (712 NaN z 5038 prkov databázy).

ÚPRAVA DÁT

Možným spôsobom náhrady výrazu NaN číslom, je vyjadriť strednú hodnotu z nameraných dát a NaN nahradiť týmto výrazom. Stredná hodnota sa môže vyjadriť viacerými spôsobmi, a to ako stredná hodnota všetkých nameraných objektov (vNaN=1), pre každú skupinu osobitne (vNaN=2) a pre jednotlivé triedy (vNaN=3). Ďalším spôsobom úpravy dát je znásobenie dát chorých (438 chorých) na rovnaký počet ako je počet zdravých a výraz NaN nahradiť strednou hodnotou všetkých objektov (vNaN=4).

METÓDY

Pre upravené dáta sa snažíme nájsť na testovacej množine (pomer testovacej ku trénovacej množine = 0,6) také pravidlo s najlepšími parametrami, ktoré by tieto dve triedy (ω_1, ω_2) oddelili s čo najväčšou presnosťou.
Na klasifikáciu dát bol použitý Fisherov lineárny klasifikátor [2] s rozdeľovacím prahom rovným 5%-nému kvantilu výstupov klasifikácie pre skupinu chorých. V metóde oporných bodov [3] sa použilo štandardné Gaussovské jadro, ktoré nelineárne transformuje charakteristiky objektov a parametre modelu boli nájdené metódou cross-validácie. Pre model doprednýchneurónových sietí [4] bola zvolená jedna skrytá vrsta s tromi neurónmi pre 1000 vnútorných epôch.
Presnosť klasifikátora budeme vyjadrovať tromi pravdepodobnosťami pri počte opakovacích cyklov 1000, a to:

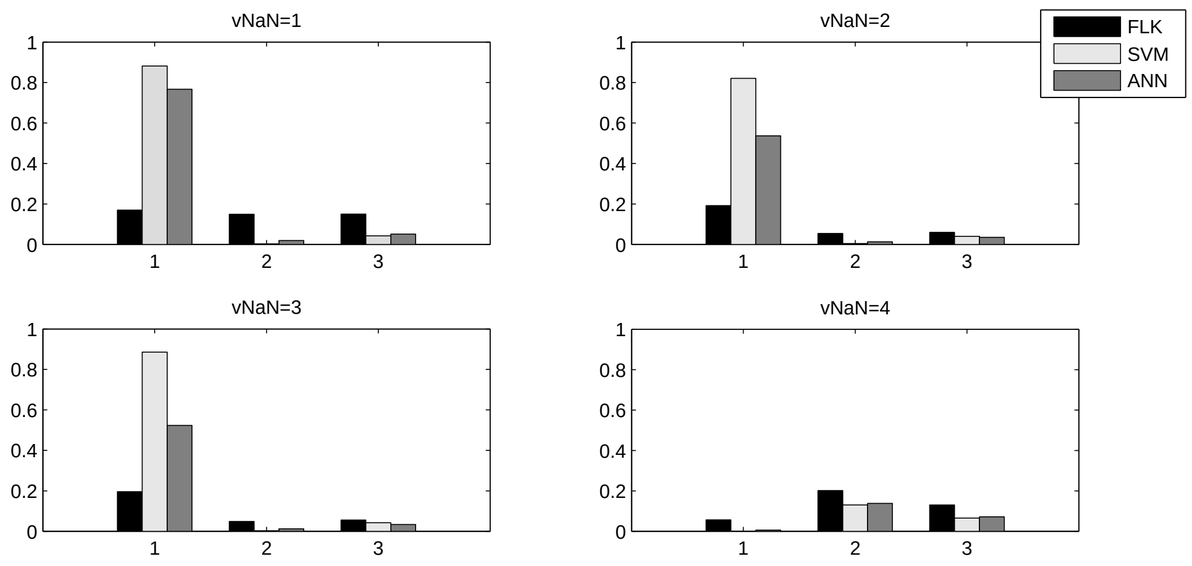
$$P1 = \frac{\text{počet zaradených objektov do triedy 'zdraví' z triedy 'chorí'}}{\text{počet testovaných objektov v triede 'chorí'}}$$
$$P2 = \frac{\text{počet zaradených objektov do triedy 'chorí' z triedy 'zdraví'}}{\text{počet testovaných objektov v triede 'zdraví'}}$$
$$P3 = \frac{\text{počet zle zaradených objektov}}{\text{počet testovaných objektov}}$$

VÝSLEDKY

V tabuľke sú výsledky pravdepodobností zlého zatriedenia do tried (P1, P2) a celková chyba (P3) podľa jednotlivých klasifikačných metód (FLK, SVM, ANN) pre rôzne upravené dáta (vNaN).

	vNaN=1			**vNaN=2**			**vNaN=3**			**vNaN=4**		
	P1	P2	P3	P1	P2	P3	P1	P2	P3	P1	P2	P3
FLK	.170	.149	.150	.192	.054	.060	.196	.049	.056	.057	.202	.130
SVM	.882	.003	.042	.821	.004	.040	.886	.003	.042	.000	.130	.065
ANN	.767	.019	.051	.537	.013	.035	.523	.012	.034	.006	.138	.072

Na obrázku je grafické znázornenie výsledkov.



V prípade klasifikácie dát, kde počet zdravých ($n = 438$) bol mnohonásobne vyšší než počet chorých ($n = 20$) je vidieť, že pravdepodobnosť zlého zatriedenia chorých je vo všetkých troch prípadoch väčšia ako pravdepodobnosť zlého zatriedenia zdravých. Extrémne hodnoty pravdepodobnosti P1 v metódach SVM a ANN sa dajú oddôvodniť tým, že metódy sa snažia o najmenšiu celkovú chybu a zlé zatriedenie objektov z triedy 'chorí' k nej prispieva veľmi málo. Najlepšie výsledky celkovej chyby boli dosiahnuté vo výsledkoch metódy ANN.
V prípade klasifikácie dát, kde počet zdravých ($n = 438$) je rovný počtu chorých, sú výsledky viditeľne odlišné od predchádzajúcich troch prípadov. Vo výsledkoch FLK je pravdepodobnosť zlého zatriedenia chorých približne rovná nastavenej hodnote kvantilu výstupov klasifikátora pre chorých. Pravdepodobnosť zlého zatriedenia zdravých je však 20 percent. V prípade SVM ide o takmer bezchybné zatriedenie chorých, ale zdravé objekty boli zle zatriedené na 13 percent. Táto metóda však dosiahla najmenšiu celkovú pravdepodobnosť zlého zatriedenia. V prípade ANN sú výsledky pobobné (o niečo horšie) ako pri metóde SVM.

ZÁVER

V štúdii sme porovnali tri rôzne klasifikačné metódy. Z výsledkov sa dá konštatovať, že v prípade nepomeru objektov v jednotlivých triedach sú výsledky klasifikácie nevyhovujúce hlavne pre zatriedenie menej početnej triedy. V prípade rovnosti počtu objektov je klasifikácia dát omnoho lepšia.

PLÁNY

Prezentovaná práca je prvým krokom v spoznávaní dát získaných meraním vydychovaných plynov. Cieľom je podľa rôznych štatistikých metód nájsť čo najlepšie vlastnosti dát tak, aby bolo ich zatriedenie pri klasifikácii do triedy 'chorí' a 'zdraví' čo najlepšie.

POĎAKOVANIE. Za cenné rady a podporu patrí moja vďaka K. Hornišovej a B. Arendackej. Práca bola podporená grantami VEGA 1/3016/06 a VEGA 2/4026/04.

REFERENCIE

[1] Amann, A. (2005). *Breath analysis for clinical diagnosis and therapeutic monitoring.*World Scientific, Singapore.
[2] Arendacká, B. (2005). *Classification of objects in segmented EPO images.* In Proc. of: MEASUREMENT 2005, 5^{th} International Conference on Measurement, 194-198.
[3] Hornišová, K. (2005). *Klasifikácia erytropoetínových obrazcov metódou oporných bodov (support vector machines).* Forum Statisticum Slovacum 3, 174-183.
[4] Witkovský, V. a kol. (2005). *Alternative Approaches to Band Classiffication in EPO images for the GASepo software.* ARC Seidersdorf research, Seidersdorf.