



WYBRANE METODY EKSPLORACJI DANYCH I WSPOMAGANIA PROCESÓW DECYZYJNYCH

Jerzy Moczko

Akademia Medyczna im. K. Marcinkowskiego w Poznaniu, Katedra i Zakład Informatyki i Statystyki

Wprowadzenie

Przedstawione poniżej analizy matematyczne zostały przeprowadzone w Katedrze Informatyki i Statystyki AM w Poznaniu dla potrzeb pracy habilitacyjnej doktora Bogusława Dębniaka pt.: „Ultrasonografia dopplerowska w ocenie ukrwienia złośliwych i niezłośliwych zmian gruczołów sutkowych”[1].

Najczęściej występującym nowotworem złośliwym u kobiet w Polsce jest rak gruczołu sutkowego. Stanowi on około 20 procent wszystkich nowotworów złośliwych u kobiet i jest najczęstszą przyczyną zgonów (około 13% w roku 1999). Ponieważ zachorowalność na tę jednostkę chorobową ma ciągle tendencję zwyżkową (przy bardzo słabym spadku umieralności), wczesna diagnostyka odgrywa zasadniczą rolę w procesie leczenia. Obok metod tak znanych i powszechnie stosowanych jak badanie palpacyjne, mammografia czy kseromammografia coraz częściej zwraca się uwagę na techniki ultrasonograficzne. Aczkolwiek pierwsze próby ze stosowaniem diagnostyki ultradźwiękowej prowadzono już w latach pięćdziesiątych ubiegłego wieku, to dopiero wprowadzenie głowic o wysokiej częstotliwości umożliwiło jej wykorzystanie w różnicowaniu złośliwych i niezłośliwych zmian w obrębie gruczołu sutkowego.

Podstawą stosowania angiografii dopplerowskiej w diagnostyce zmian gruczołów sutkowych (różnicowanie zmian złośliwych i niezłośliwych) jest spostrzeżenie, że przepływy krwi w naczyniach normalnych i naczyniach powstających w rosnącym guzie w procesie neoangiogenezy różnią się istotnie.

Metodyka badania

W ocenie pomiaru przepływu krwi wykorzystano pięć wskaźników:

- ◆ indeks pulsacji (PI) wg Goslinga,
- ◆ indeks oporu (RI) wg Pourceleta,
- ◆ średnią maksymalną prędkość przepływu krwi (TAMAX),



- ♦ prędkość minimalną (późnorozkurczową) fali przepływu krwi (VMIN),
- ♦ prędkość maksymalną fali przepływu krwi (VMAX).

Kryterium diagnostyczne określające charakter zmiany (złośliwa lub niezłośliwa) oparto na wynikach biopsji otwartej lub ocenie preparatów pooperacyjnych.

Wyniki statystyki opisowej

Wejściowy zbiór danych stanowi plik zawierający 6 kolumn z wynikami badań 497 pacjentek, którego fragment ilustruje rysunek 1. Zmienna GRUPA zawiera wyniki badań histopatologicznych pomierzonych w skali nominalnej (kod 1 – zmiany łagodne, kod 2 – zmiany złośliwe), pozostałe zmienne pomierzone w skali interwałowej obejmują wyniki wymienionych w poprzednim rozdziale parametrów ultrasonograficznych. Na podstawie tych danych spróbujemy odpowiedzieć na następujące pytania:

Dane: doppler_kolor (6 zm. * 497 prz.)						
	1 GRUPA	2 PI	3 RI	4 TAMAX	5 VMIN	6 VMAX
197	1.000	1.100	0.620	7.000	5.000	13.000
198	1.000	1.190	0.690	9.000	5.000	16.000
199	1.000	0.580	0.440	7.000	5.000	9.000
200	1.000	1.060	0.620	8.000	5.000	13.000
201	1.000	0.880	0.570	9.000	6.000	14.000
202	1.000	0.780	0.540	9.000	6.000	13.000
203	2.000	1.130	0.670	7.000	4.000	14.000
204	2.000	2.330	1.000	11.000	0.000	25.000
205	2.000	1.100	0.670	15.000	8.000	24.000
206	2.000	0.980	0.640	17.000	10.000	11.000
207	2.000	1.080	0.640	13.000	8.000	22.000
208	2.000	1.860	1.000	18.000	0.000	33.000
209	2.000	1.070	1.000	4.000	0.000	4.000
210	2.000	1.280	0.680	10.000	6.000	19.000
211	2.000	1.620	0.760	18.000	9.000	38.000
212	2.000	1.180	0.670	8.000	5.000	15.000

Rys. 1. Struktura wejściowego zbioru danych

1. Czy pięć pomierzonych w badaniu parametrów ultrasonograficznych może być przydatnych we wczesnej nieinwazyjnej diagnostyce zmian nowotworowych gruczołów sutkowych,
2. Czy wszystkie pięć zmiennych musi być wykorzystane w modelu klasyfikacyjnym, czy też zawierają one informację redundantną i można część z nich odrzucić.

Przeprowadzona analiza opisowa wskazuje na bardzo silną prawostronną skośność rozkładów wszystkich parametrów, zarówno w grupie zmian łagodnych, jak i złośliwych (wynik testu normalności Shapiro-Wilka $p < 0.000001$). Nie możemy zatem użyć klasycznego testu parametrycznego t-Studenta do oceny, czy wartości średnich arytmetycznych parametrów USG różnią się w obu typach schorzenia czy też nie. Aby przekonać się o istotności statystycznej różnic rozkładu parametrów posłużymy się nieparametrycznym testem Manna-Whitney'a, którego moc jest bardzo zbliżona do mocy testu t-Studenta (95.5%) [1]. Wyniki testu (rys. 2.) wskazują na fakt, iż rozkłady czterech

spośród badanych parametrów (PI, RI, TAMAX oraz VMAX) bardzo silnie różnicują badane podgrupy i warto jest przyjrzeć im się bliżej.

Test U Manna-Whitneya (doppler_kolor)									
Wzg.zmienn. GRUPA									
Zaznaczone wyniki są istotne z p < .05000									
zmienna	Sum.rang Grupa 1	Sum.rang Grupa 2	U	Z	poziom p	Z popraw.	poziom p	N ważn. Grupa 1	N ważn. Grupa 2
PI	37545.00	86208.00	17042.00	-8.10964	0.000000	-8.10996	0.000000	202	295
RI	38634.00	85119.00	18131.00	-7.41715	0.000000	-7.43659	0.000000	202	295
TAMAX	46184.50	77568.50	25681.50	-2.61578	0.008903	-2.62386	0.008694	202	295
VMIN	52789.50	70963.50	27303.50	1.58435	0.113116	1.59537	0.110630	202	295
VMAX	42613.50	81139.50	22110.50	-4.88658	0.000001	-4.89104	0.000001	202	295

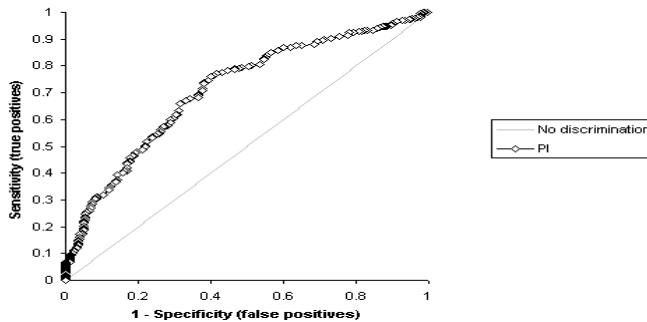
Rys. 2. Wyniki porównania rozkładów badanych parametrów USG w grupie zmian łagodnych (grupa 1) i złośliwych (grupa 2)

Analiza krzywych ROC

Spróbujemy teraz oszacować zdolność dyskryminacyjną poszczególnych parametrów w układzie jednowymiarowym. Posłużymy się w tym celu krzywymi ROC (Receiver Operating Characteristic) [2]. Na rysunku 3 prezentujemy fragment wyników otrzymanych dla parametru PI.

GRUPA	n
1	202
2	295

Curve	Area	SE	p	95% CI of Area	GRUPA = r	2
PI	0.714	0.0233	<0.0001	0.668 to 0.760	have higher values	



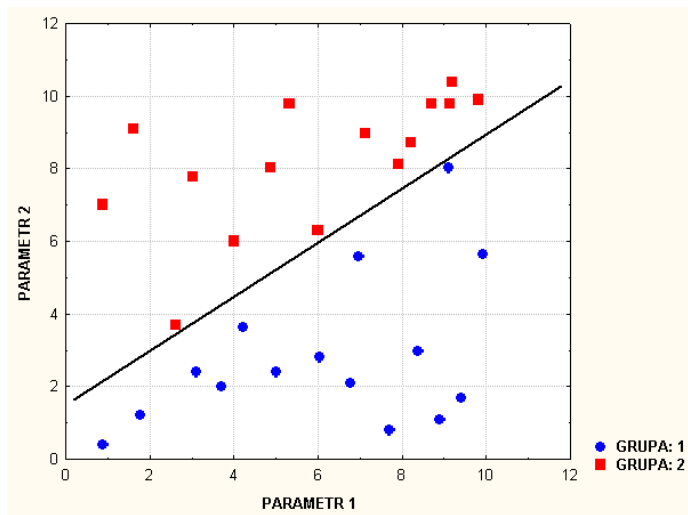
Rys. 3. Analiza krzywej ROC dla parametru PI

Analiza krzywych ROC pozwala odpowiedzieć na pytanie, czy badany parametr umożliwia prawidłowe zaklasyfikowanie analizowanego obiektu do jednej z dwóch klas w sposób maksymalnie jednoznaczny. W tym celu dla określonej wartości parametru (tzw. punktu odcięcia) zlicza się liczbę przypadków prawdziwie i fałszywie pozytywnych, liczbę przypadków prawdziwie i fałszywie negatywnych i na podstawie tych wielkości wyznacza się czułość i swoistość metody klasyfikacyjnej opartej na badanym parametrze. Następnie

zmienia się wartość punktu odcięcia i ponownie wyznacza się czułość i swoistość. Przechodząc w ten sposób przez cały zakres zmienności parametru, konstruujemy dwuwymiarowy wykres, na którym nanosimy wyznaczone uprzednio wartości czułości i swoistości. Ponieważ zarówno czułość, jak i swoistość mogą przyjmować wartości z przedziału domkniętego $<0,1>$, zatem maksymalna wartość pola pod krzywą wynosi 1 i odpowiada pełnej dyskryminacji (liczba przypadków fałszywie dodatnich i ujemnych jest równa zero). Przebieg krzywej ROC wzdłuż przekątnej odpowiada prawdopodobieństwu 0.5, co oznacza, że parametr nie ma właściwości dyskryminacyjnych. Wyniki przedstawione na rysunku 3. wskazują, że użycie samego parametru PI do rozróżniania typu guza sutka pozwala na istotną statystycznie klasyfikację ($p < 0.0001$). Z tablic konstruowanych do wyznaczenia krzywej ROC można dodatkowo odczytać optymalną wartość parametru odcięcia, tzn. minimalizującą błąd klasyfikacyjny. Dla badanego parametru PI wynosi ona 1.06 (dla tej wartości liczba przypadków prawdziwie pozytywnych $TP=228$, prawdziwie negatywnych $TN=118$, fałszywie pozytywnych $FP=84$, a fałszywie negatywnych $FN=67$). Założyliśmy tutaj po cichu, że wagi popełnienia błędu I i II rodzaju są jednakowe. W praktyce można również przyjąć inny stosunek wag, co wpłynie oczywiście na wartość odcięcia. Przeprowadzenie analogicznej analizy dla pozostałych parametrów potwierdza wstępne wyniki uzyskane z testu Manna-Whitney'a, a dodatkowo umożliwia określenie wartości odcięcia dla każdego z nich i odpowiadającą im macierz klasyfikacji.

Podejście wielowymiarowe

Do tej pory próbowaliśmy podjąć decyzję na podstawie każdego z pięciu parametrów z osobna. Narzuca się zatem pytanie, czy jednoczesne użycie dwóch lub większej liczby parametrów poprawi jakość klasyfikacji. Idea polepszenia jakości klasyfikacyjnej poprzez zwiększenie wymiarowości przestrzeni badania przedstawiona jest na rysunku 4.



Rys. 4. Przykład rozkładu danych nieseparowalnych przy użyciu pojedynczego parametru



Łatwo zauważyć, że nie możemy podjąć jednoznacznej decyzji co do przynależności obiektu do określonej grupy na podstawie wyłącznie jednego parametru, obojętnie – pierwszego czy drugiego. W obu sytuacjach istnieje obszar niejednoznaczności powodujący generowanie niezerowych wartości fałszywie pozytywnych i fałszywie negatywnych. Jednakże jeżeli jednocześnie weźmiemy pod uwagę oba parametry, to łatwo dostrzec, iż w zbudowanej dwuwymiarowej przestrzeni można skonstruować prostą jednoznacznie separującą obie klasy (liczba rozpoznań fałszywie pozytywnych i fałszywie negatywnych jest równa zero). Nasze rozumowanie możemy rozszerzyć na większą liczbę wymiarów, np. gdy weźmiemy pod uwagę jednocześnie trzy pomiary, to w przestrzeni trójwymiarowej będziemy próbowali konstruować hiperpłaszczyznę rozdzielającą jednoznacznie interesujące nas klasy przypadków. Oczywiście przeciętnemu człowiekowi trudno sobie wyobrazić większą liczbę wymiarów i generowanych w nich hiperprzestrzeni rozdzielających, jednakże istnieją techniki matematyczne, które dają sobie radę z dowolną liczbą wymiarów (nawet nieskończoną). W dalszej części pracy przedstawimy kilka takich technik i porównamy ich zalety i wady.

Analiza dyskryminacyjna

U podstaw analizy dyskryminacyjnej leży pytanie, czy klasyfikowane grupy różnią się ze względu na wartości średnie pewnych zmiennych i w związku z tym, czy można te zmienne wykorzystać do badania przynależności badanych obiektów do wspomnianych grup [3]. Jak widać, sformułowanie zagadnienia jest niezwykle zbliżone do jednoczynnikowej analizy wariancji. Analogiczne są również podstawowe założenia, które muszą być spełnione, aby analiza dyskryminacyjna mogła być bezpiecznie stosowana. Jak pamiętamy, jednym z nich było pochodzenie danych z populacji o wielowymiarowym rozkładzie normalnym. W naszym przypadku założenie to jest naruszone, lecz mimo to spróbujemy z pewną ostrożnością zastosować analizę dyskryminacyjną do klasyfikacji typu nowotworu sutka. Przeprowadzone w ostatnich latach badania symulacyjne przy użyciu metod Monte Carlo wykazały bowiem empirycznie, że wpływ na uzyskane wyniki niespełnienia założenia normalności jest znikomy. Wyniki przeprowadzonej analizy dyskryminacyjnej uwidocznione są na rysunku 5. Wynika z nich, że spośród pięciu użytych predyktorów istotną rolę w dyskryminacji wykazuje parametr PI. Wartości cząstkowej Lambdy Wilksa wskazują wkład poszczególnych zmiennych w modelu predykcyjnym. Na jej podstawie możemy stwierdzić, że po zmiennej PI największy udział w modelu odgrywa zmienna VMAX, następnie TAMAX, RI, a na samym końcu VMIN. Wielkość tolerancji określa stopień redundancji zmiennej. Wynika stąd, że zmienne TAMAX i VMAX niosą informację redundantną w stosunku do informacji wnoszonej przez pozostałe zmienne użyte w modelu. Test chi-kwadrat przeprowadzony w analizie kanonicznej wskazuje na wysoką istotność wygenerowanej funkcji dyskryminacyjnej. Przytoczona macierz klasyfikacji post-hoc podkreśla fakt, iż guzy złośliwe są wykrywane przez użyty model lepiej (77.63 %) niż guzy łagodne (tylko 54.46%).



Podsumowanie analizy funkcji dyskryminacyjnej. (doppler_kolor)						
Zmiennych w modelu: 5;Grupująca: GRUPA (2 grup)						
Lambda Wilksa: .85926 przybl. F (5,491)=16.084 p<.0000						
N=497	Lambda Wilksa	Cząstk. Wilksa	F usun. (1,491)	poziom p	Toler.	1-Toler. (R-kwad)
PI	0.867939	0.990003	4.957933	0.026424	0.145935	0.854065
RI	0.860171	0.998943	0.519411	0.471436	0.163710	0.836291
TAMAX	0.859357	0.999890	0.054025	0.816299	0.043520	0.956480
VMIN	0.859708	0.999481	0.255023	0.613787	0.182235	0.817765
VMAX	0.860485	0.998578	0.699045	0.403512	0.049430	0.950570

Testy chi-kwadrat kolejnych pierwiastków (doppler_kolor)						
Pierw. Usunięte	Wartość własna	Kanonicz R	Lambda Wilksa	chi-kwad	df	poziom p
0	0.163789	0.375150	0.859262	74.70301	5	0.000000

Macierz klasyfikacji (doppler_kolor)			
Wiersze: obserwowana klasyfik.			
Kolumny: Przewidywana klasyfikacja			
Grupa	Procent Poprawne	G_1:1 p=.40644	G_2:2 p=.59356
G_1:1	54.45544	110	92
G_2:2	77.62712	66	229
Razem	68.20926	176	321

Rys. 5. Wyniki analizy dyskryminacyjnej

Spróbujemy teraz przeprowadzić analizę dyskryminacyjną krokową. Zacniemy od analizy postępującej (rys. 6.). Do modelu zostają wprowadzone jedynie dwie zmienne: PI oraz VMAX. Pozostałe zmienne znajdują się poza modelem. Jak widać, model zbudowany na tych dwóch zmiennych ma niemal takie same zdolności klasyfikacyjne, jak model klasyczny uwzględniający wszystkie zmienne wejściowe. Procedura krokowa pozwoliła nam zatem na redukcję wymiarowości naszego systemu informacyjnego.

Podsumowanie analizy funkcji dyskryminacyjnej. (doppler_kolor)						
Krok 2, N zmn. w modelu: 2;Grupująca: GRUPA (2 grup)						
Lambda Wilksa: .86028 przybl. F (2,494)=40.117 p<.0000						
N=497	Lambda Wilksa	Cząstk. Wilksa	F usun. (1,494)	poziom p	Toler.	1-Toler. (R-kwad)
PI	0.947244	0.908188	49.94014	0.000000	0.953224	0.046776
VMAX	0.879588	0.978044	11.08981	0.000933	0.953224	0.046776

Macierz klasyfikacji (doppler_kolor)			
Wiersze: obserwowana klasyfik.			
Kolumny: Przewidywana klasyfikacja			
Grupa	Procent Poprawne	G_1:1 p=.40644	G_2:2 p=.59356
G_1:1	54.45544	110	92
G_2:2	77.96610	65	230
Razem	68.41046	175	322

Rys. 6. Wyniki analizy dyskryminacyjnej krokowej postępującej

Identyczne wyniki otrzymamy, stosując procedurę wsteczną. Jest to o tyle istotne, że z reguły obie techniki prowadzą do nieco innego zestawu zmiennych włączonych do



modelu. Zgodność wyników z obu procedur potwierdza fakt, że te właśnie zmienne opisujące przepływ krwi odgrywają istotną rolę w predykcji typu zmian guza sutka.

Analiza logistyczna

Przyjrzyjmy się teraz, jakie rezultaty otrzymamy, próbując rozwiązać ten sam problem techniką analizy logistycznej [4]. Można z powodzeniem próbować zastosować to podejście, gdyż zmienna zależna GRUPA jest zmienną dychotomiczną (z tej przyczyny użycie klasycznej regresji wielokrotnej jest niemożliwe). Warunek dostatecznej liczebności danych niezbędnych do prowadzenia analizy logistycznej ($N > 10 \cdot (k+1)$, gdzie k jest liczbą estymowanych parametrów) jest spełniony. Wyniki analizy przedstawione są na rysunku 7.

Model: Regr. logistyczna (logit) (doppler_kolor)						
Zmn. zal: GRUPA Strata: najw.wiaryg. bł.średnkw.skala do 1						
Końc.strata 294.27604704 Chi2(5)=82.931 p=.00000						
N=497	Stała B0	PI	RI	TAMAX	VMIN	VMAX
Ocena	-1.37807	1.42	-0.89506	-0.024	-0.07388	0.07046
Błąd standard.	0.85658	0.68	1.68657	0.086	0.09222	0.04767
t(491)	-1.60880	2.09	-0.53070	-0.278	-0.80115	1.47791
poziom p	0.10830	0.04	0.59587	0.781	0.42343	0.14007
-95%CL	-3.06110	0.09	-4.20884	-0.194	-0.25508	-0.02321
+95%CL	0.30495	2.75	2.41872	0.146	0.10732	0.16412
Chi-kwadrat Walda	2.58824	4.37	0.28164	0.077	0.64184	2.18422
poziom p	0.10767	0.04	0.59563	0.781	0.42305	0.13944
iloraz szans z.jedn.	0.25206	4.12	0.40858	0.976	0.92878	1.07300
-95%CL	0.04684	1.09	0.01486	0.824	0.77485	0.97706
+95%CL	1.35656	15.60	11.23141	1.157	1.11329	1.17836
iloraz szans zakr.		191.47	0.49751	0.261	0.19682	2321.85200
-95%CL		1.37	0.03752	0.000	0.00365	0.07783
+95%CL		26695.87	6.59688	3489.039	10.60092	69268670.00000

Klasyfikacja przypadków (doppler_kolor)			
Il. szans: 4.0811			
Obszew.	Przew.	Przew.	Procent Popraw.
1.000000	113	89	55.94059
2.000000	70	225	76.27119

Rys. 7. Wyniki analizy logistycznej przy włączeniu do modelu wszystkich parametrów

Jak łatwo zauważyć, użyty model z włączonymi wszystkimi deskryptorami różni się istotnie statystycznie od modelu zawierającego wyłącznie wyraz wolny. Z drugiej jednak strony w modelu tym jedynie zmienna PI jest istotnie związana z klasyfikacją guza. Pozostałe zmienne nie mają istotnego wpływu na jakość klasyfikacji. Jak widać, w przeciwieństwie do analizy dyskryminacyjnej, model analizy logistycznej nie wykazał istotności parametru VMAX. Spróbujmy zatem zbudować model logistyczny oparty wyłącznie na zmiennej PI (rys. 8), a następnie włączmy parametr VMAX i sprawdzimy, czy jego dołączenie zmienia jakość klasyfikacyjną.



		Model: Regr. logistyczna (logit) (doppler_kolor)	
		Zmn. zal: GRUPA Strata: najw.wiaryg. bł.średnkw.skala	
		Końc.strata 301.56722800 Chi2(1)=68.348 p=.00000	
		Stała B0	PI
N=497			
Ocena		-1.76899	1.709
Błąd standard.		0.29764	0.233
t(495)		-5.94346	7.333
poziom p		0.00000	0.000
-95%CL		-2.35377	1.251
+95%CL		-1.18420	2.167
Chi-kwadrat Walda		35.32473	53.779
poziom p		0.00000	0.000
iloraz szans z.jedn.		0.17051	5.523
-95%CL		0.09501	3.494
+95%CL		0.30599	8.731
iloraz szans zakr.			566.943
-95%CL			103.709
+95%CL			3099.305

Klasyfikacja przypadków (doppler_kolor)			
Il. szans: 4.2310			
	Przew.	Przew.	Procent
Obserw.	1.000000	2.000000	Popraw.
1.000000	108	94	53.46535
2.000000	63	232	78.64407

Rys. 8. Wyniki analizy logistycznej przy włączeniu do modelu wyłącznie zmiennej PI

		Model: Regr. logistyczna (logit) (doppler_kolor)		
		Zmn. zal: GRUPA Strata: najw.wiaryg. bł.średnkw.skala do 1		
		Końc.strata 294.92553974 Chi2(2)=81.632 p=.00000		
		Stała B0	PI	VMAX
N=497				
Ocena		-2.18516	1.536	0.0362
Błąd standard.		0.32999	0.238	0.0106
t(494)		-6.62193	6.461	3.4024
poziom p		0.00000	0.000	0.0007
-95%CL		-2.83351	1.069	0.0153
+95%CL		-1.53680	2.003	0.0571
Chi-kwadrat Walda		43.84992	41.738	11.5766
poziom p		0.00000	0.000	0.0007
iloraz szans z.jedn.		0.11246	4.647	1.0369
-95%CL		0.05881	2.912	1.0154
+95%CL		0.21507	7.414	1.0588
iloraz szans zakr.			298.610	53.7505
-95%CL			52.768	5.3846
+95%CL			1689.797	536.5522

Klasyfikacja przypadków (doppler_kolor)			
Il. szans: 4.0719			
	Przew.	Przew.	Procent
Obserw.	1.000000	2.000000	Popraw.
1.000000	111	91	54.95050
2.000000	68	227	76.94915

Rys. 9. Wyniki analizy logistycznej po włączeniu do modelu z rysunku 8 zmiennej VMAX



Zauważamy, że model logistyczny oparty wyłącznie na zmiennej PI daje gorsze rezultaty (zwiększa iloraz szans z początkowej wartości 4.08 do wartości 4.23) w stosunku do modelu pierwotnego. Pogorszenie to polega na zwiększeniu zarówno liczby przypadków fałszywie dodatnich, jak i liczby przypadków fałszywie ujemnych. Dodanie nieistotnego statystycznie parametru VMAX (patrz rys.7.) znakomicie polepsza uzyskane rezultaty i czyni je niemal identycznymi z wynikami uzyskanymi w modelu logistycznym włączającym wszystkie predyktory (rys.9).

Podsumowując możemy stwierdzić, że dwie różne techniki statystyczne – analiza dyskryminacyjna i analiza logistyczna, bazujące na odmiennych założeniach, doprowadziły do analogicznych wyników. Pamiętajmy jednak, że sytuacja taka nie zawsze ma miejsce, co zostanie wyjaśnione w dalszej części opracowania.

Analiza przy użyciu drzew regresyjno-klasyfikacyjnych CRT

Przytoczone w poprzednich rozdziałach techniki eksploracyjne reprezentowały podejście czysto statystyczne. Ostatnio coraz większym zainteresowaniem cieszą się również metody oparte na sztucznej inteligencji (sztuczne sieci neuronowe, zbiory rozmyte, zbiory przybliżone) i teorii przetwarzania sygnałów. Techniki te wraz z drzewami regresyjno-klasyfikacyjnymi (zwanymi również drzewami decyzyjnymi) składają się na metodykę badawczą określaną w literaturze anglosaskiej jako „data mining”. Odpowiednik w języku polskim nie został jeszcze ostatecznie zdefiniowany i różni autorzy prac tłumaczą go jako „zgłębianie danych”, „drażnienie danych”, „kopanie w danych”, a nawet „torturowanie danych”. Wydaje się, że pierwszy z przytoczonych terminów przyjmie się jako standard. Zgłębianie danych definiuje się jako proces automatycznego lub półautomatycznego badania dużych ilości danych w celu znalezienia istotnych zależności, wzorców i reguł. Należy podkreślić fakt, że wnioski uzyskiwane metodami zgłębiania danych mają charakter indukcyjny. Oznacza to, że u źródła tworzonych modeli nie tkwią apriorycznie przyjęte abstrakcyjne teorie, lecz sama struktura analizowanych zbiorów danych. Przejdziemy teraz do jednej z najprostszych interpretacyjnie technik eksploracyjnych – drzew regresyjno-klasyfikacyjnych [5-8].

Zasadniczą zaletą drzew decyzyjnych jest brak jakichkolwiek założeń wstępnych dotyczących rozkładów danych. Szczególnie przydatne są one w sytuacjach, w których występują skorelowane ze sobą dane (w analizach wykorzystujących ogólny model liniowy generują one źle uwarunkowane macierze, których rozwiązywanie prowadzi najczęściej do uzyskiwania niestabilnych numerycznie rozwiązań). Ponadto wygenerowane reguły logiczne są dla lekarzy praktyków łatwiejsze w interpretacji niż różnego rodzaju funkcje klasyfikacyjne. Na podstawie drzew łatwo jest też opracować rozmaite standardy postępowania diagnostycznego czy terapeutycznego. Spróbujemy zatem zastosować je do przeanalizowania naszego przykładu.

Zmienna GRUPA stanowi, podobnie jak poprzednio, zmienną zależną, natomiast pozostałe zmienne są predyktorami porządkowymi. W analizie drzew decyzyjnych stosowane są trzy podstawowe metody wyboru podziału: dyskryminacyjne podziały jednowymiarowe dla zmiennych porządkowych i nominalnych, dyskryminacyjne podziały na podstawie



kombinacji liniowej wyłącznie dla predyktorów porządkowych oraz metoda CRT wyczerpującego poszukiwania podziałów jednowymiarowych. W naszym przypadku stosowana jest każda z wymienionych metod. Drugim ważnym czynnikiem jest wybór opcji zatrzymania tworzenia drzewa. Łatwo bowiem wykazać, że dla dowolnego zestawu danych można zbudować mniej lub bardziej skomplikowane drzewo, które dokona idealnej 100% zgodnej klasyfikacji. Problem przypomina klasyczne zagadnienie regresji dwuwymiarowej. Mając N par danych (jedną zmienną zależną i jedną zmienną niezależną) można wygenerować wielomian N -tego rzędu idealnie dopasowany do rozkładu danych. Otrzymamy zatem 100-procentowe wyjaśnienie zmienności zmiennej zależnej na podstawie zmian zmiennej niezależnej, lecz otrzymany model nie będzie przydatny w praktyce – będzie zbyt skomplikowany.

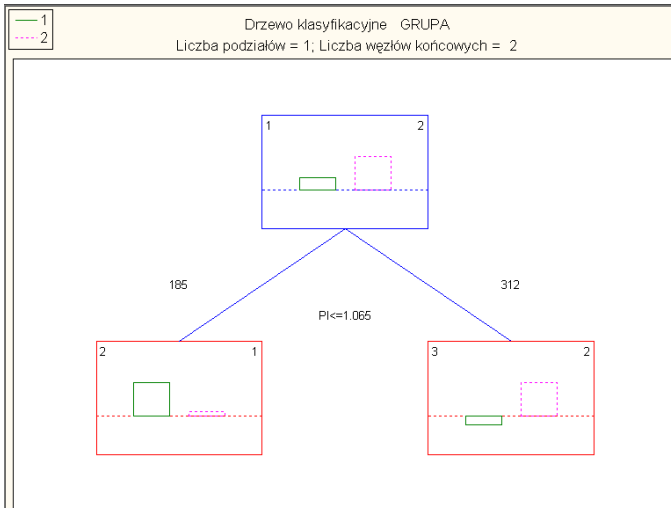
Zbyt wiele szczegółów utrudni badaczowi interpretację i uogólnienie wyników. Będziemy nazywali go modelem „przeuczonym” lub modelem nauczonego „na pamięć”. Odtworzy on idealnie wszystkie przypadki, których użyliśmy do tworzenia modelu, natomiast najczęściej „zglupieje”, widząc przypadek uprzednio nie przedstawiony. Przypomina on zatem ucznia, który wykuł na pamięć wszystkie wzory i reguły i potrafi je idealnie wyrecytować. Jednakże nie daje sobie rady, gdy zapytamy go o cokolwiek zbliżonego – nie potrafi bowiem uogólniać informacji i kojarzyć faktów.

Podobnie jest z konstrukcją drzew decyzyjnych. Źródło błędu tkwi w fakcie, że dane, które poddajemy analizie, zawierają oprócz informacji istotnej pewien poziom szumów. Model przeuczony oprócz istotnych związków zaczyna również odtwarzać szumy, a więc to, czego najchętniej chcielibyśmy się pozbyć. Czasami lepiej jest skonstruować drzewo prostsze o gorszej jakości klasyfikacji typu „post-hoc” (a więc klasyfikacji tych samych przypadków, których użyliśmy do tworzenia tego drzewa), ale łatwiejsze w interpretacji i mające zbliżone lub lepsze zdolności uogólniające. O metodach sprawdzania jakości budowanego modelu wspomnimy jeszcze w dalszej części rozdziału, a na razie powróćmy do wyboru opcji zatrzymania tworzenia drzewa. Stosowane są trzy podstawowe procedury: przytnij przy błędzie złej klasyfikacji, przytnij przy odchyleniu oraz dokonaj bezpośredniego zatrzymania typu FACT. W pracy przedstawimy jedynie dwie spośród dziewięciu możliwych kombinacji użytych metod podziału i opcji zatrzymania. Jako pierwsze zanalizujemy wyniki otrzymane z metody CRT wyczerpującego poszukiwania podziałów jednowymiarowych powiązanego z opcją bezpośredniego zatrzymania typu FACT. Na rysunku 10 przedstawiono strukturę drzewa decyzyjnego, fragment tablicy parametrów opisującej drzewo (m.in. zmienne podziału oraz ich wartości progowe), wyniki klasyfikacji typu „post-hoc” oraz wyniki klasyfikacji w procedurze pięciokrotnej walidacji krzyżowej (zwanej też walidacją skrośną lub krosswalidacją). Technika walidacji skrośnej polega na losowym podziale posiadanego zbioru danych na dwie części: uczącą i testującą. Pierwszego podzbioru używamy do wytworzenia modelu, drugiego zaś do testowania wytworzonego modelu. Procedura ta może być powtarzana wielokrotnie. Jeżeli wytworzone drzewa bazowały głównie na szumie, a nie na istotnej informacji, wynik walidacji skrośnej będzie bardzo słaby.



((57+66)/497). Dla porównania technika ROC dała stopę błędu około 30%, a analiza dyskryminacyjna i logistyczna – około 32%). Wydaje się zatem, że otrzymany model drzewa decyzyjnego jest zdecydowanie najlepszy. Jest to prawdą, ale jedynie w zakresie odtwarzania widzianych już uprzednio przez model przypadków. Rezultat pięciokrotnej walidacji skróśnej wynosi niemal 39% ((89+103)/497), czyli wytworzony model jest dla nas w zasadzie mało przydatny. Po prostu nauczył się on zbioru danych „na pamięć”, nieźle go odtwarza, ale gdy próbujemy mu pokazać coś nowego – zwyczajnie się gubi.

Strukt. drzewa (doppler_kolor)							
Węzły-potomkowie, n obserw. klas przewidywane klasy i warunki podziału dla węzłów							
Węzeł	Lewostr. gałąź	Prawostr. gałąź	n klas 1	n klas 2	Przewid. klasa	Podział stała	Podział zmienna
1	2	3	202	295	2	-1.06500	PI
2			118	67	1		
3			84	228	2		



Przewidywane a obserwowane klasy (doppler_kolor)					
Przewidywane(wiersz), obserwowane(kolumna)					
N próby uczącej = 497					
Klasa	Klasa 1	Klasa 2			
1	118	67			
2	84	228			

Błędne klasyfikacje w globalnym s. krzyż. (doppler_kolor)					
Przewidywane(wiersz), obserwowane(kolumna)					
Globalne koszty= .32193; odch. std. = .02096					
Klasa	Klasa 1	Klasa 2			
1		80			
2	80				

Rys. 11. Wyniki analizy wyczerpującego poszukiwania podziałów + zatrzymanie typu przytnij przy błędzie złej klasyfikacji



Zastosujemy teraz drugą kombinację: metodę CRT wyczerpującego poszukiwania podziałów jednowymiarowych powiązaną z opcją zatrzymania przytnij przy błędzie złej klasyfikacji (minimalna liczebność przypadków do badania dalszego podziału wynosi 5) - rys.11. Otrzymane drzewo jest niezwykle proste – zawiera jedynie jeden podział oparty na zmiennej PI. Uzyskana reguła brzmi: jeżeli pomierzona wartość PI jest mniejsza od 1.065, to mamy najprawdopodobniej do czynienia ze zmianami łagodnymi (ten warunek spełnia 185 przypadków). W przeciwnym razie podejrzewamy zmiany złośliwe (312 przypadków). Proszę zwrócić uwagę na określenia „najprawdopodobniej”, „podejrzewamy”. Skonstruowany model myli się w około 30 procentach. Dla przykładu, wśród 185 przypadków zaklasyfikowanych jako łagodne postawił rozpoznanie zgodne z wzorcem histopatologicznym w 118 przypadkach (pomylił się w 67 przypadkach). Pośród 312 przypadków rozpoznanych przez drzewo jako złośliwe uzyskaliśmy potwierdzenie badania histopatologicznego w 228 przypadkach, pomyłka nastąpiła w 84 przypadkach.

Przeprowadzenie pięciokrotnej walidacji krzyżowej doprowadza do stopy błędu około 32%. Jak zatem widzimy, jest to rezultat w pełni zbieżny z wynikami analizy dyskryminacyjnej i logistycznej. Zaprezentowany przykład w żadnym razie nie wyczerpuje wszystkich możliwości generowania drzew decyzyjnych. Z oczywistych względów ograniczyliśmy się tu jedynie do najprostszych opcji, mając na celu zilustrowanie idei konstrukcji drzew i porównania ich właściwości z innymi technikami klasyfikacyjnymi.

W pakiecie *STATISTICA* dostępne są między innymi dwie bardziej zaawansowane opcje: drzewa interakcyjne oraz drzewa CRT ze wzmacnianiem (tzw. boosted trees). Jak już wiemy, podstawowa idea konstrukcji drzewa opiera się na rekurencyjnym podziale zbioru danych ze względu na każdą z badanych cech. W każdym kolejnym kroku wybierany jest taki podział, który daje w wyniku najsilniej rozseparowane podzbiory. Problem stanowi najczęściej podjęcie decyzji o zakończeniu prowadzenia dalszych podziałów. Kryteria automatycznego zatrzymywania rozwoju drzewa oparte np. na minimalnej liczbie obiektów w liściu, na liczbie dopuszczonych poziomów drzewa lub na dopuszczalnej frakcji mylnie rozpoznanych obiektów są czasami mało elastyczne. Użycie opcji drzew interaktywnych usuwa tę niedogodność, gdyż badacz może zgodnie ze swoją wolą rozwijać dalsze podziały w niektórych lub wszystkich gałęziach, bądź też pójść w drugim kierunku, „zwijając” pewne rozgałęzienia. Można w ten sposób uzyskać drzewo optymalnie dopasowane do rzeczywistych potrzeb eksperymentatora.

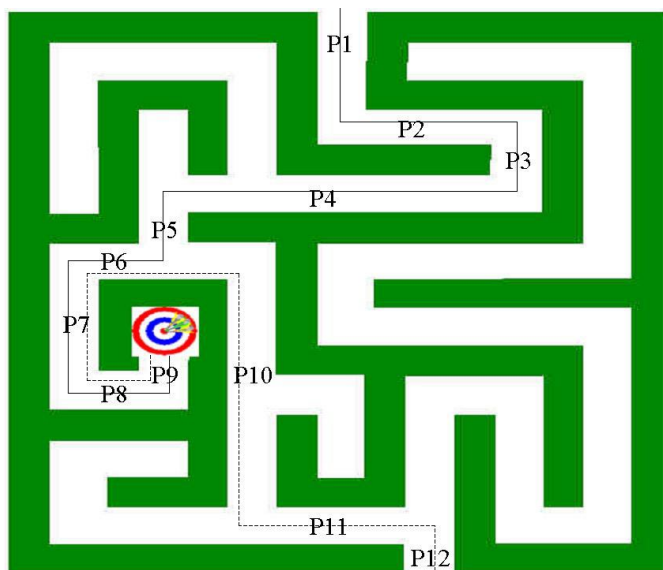
Nieco inna idea przyświeca procedurze drzew wzmacnianych. W podejściu klasycznym w procesie uczenia każda z obserwacji używana jest z tą samą wagą. Otrzymujemy w rezultacie drzewo klasyfikujące o określonej zdolności predykcyjnej. Gdy pracujemy na rzeczywistym, a nie symulowanym materiale badawczym, będzie ono obok prawidłowych wyników generowało mylne prognozy (odpowiadające przypadkom fałszywie dodatnim lub fałszywie ujemnym). W tej sytuacji można przyporządkować tym obiektom, które opierają się prawidłowej klasyfikacji, wyższe wagi, w przeciwieństwie do obiektów rozpoznawanych bezproblemowo. Procedura ta może być prowadzona rekurencyjnie, generując sekwencję drzew, w której kolejne drzewo lepiej klasyfikuje przypadki, z którymi drzewo poprzednie nie dawało sobie rady.



Dyskusja uzyskanych wyników

W poprzednich rozdziałach przedstawiliśmy kilka możliwych podejść do analizy tego samego problemu medycznego. Każda z metod wskazała na parametr PI jako optymalny we wspomaganie klasyfikowania zmian nowotworowych w obrębie guza sutka na łagodne lub złośliwe. Dodanie drugiego parametru (VMAX) nie zwiększało zdolności klasyfikacyjnej w radykalny sposób. Powstaje zatem pytanie, czy zawsze otrzymamy taką zgodność. Otóż na ogół uzyskane rezultaty będą znacząco różne. Może się okazać, że część metod wybierze jeden zestaw zmiennych, inne metody wykażą przydatność klasyfikacyjną innej grupy zmiennych. Dla wielu osób sytuacja ta wydaje się paradoksalna i traci wiary w stosowane techniki eksploracyjne. Otóż każda z użytych technik zwraca uwagę na nieco inne aspekty informacyjne niesione przez analizowane dane. Aby w prosty sposób wyjaśnić ten problem, posłużmy się prostym przykładem. W statystyce opisowej do opisu miary tendencji centralnej zbioru danych używamy różnych mierników, m.in. średniej arytmetycznej, mediany i wartości modalnej. W skali interwałowej możemy użyć teoretycznie każdego z nich, lecz z reguły kierujemy się następującą wskazówką. Gdy dane są jednorodne (jednomodalne), a ich rozkład jest w przybliżeniu symetryczny, jako miarę tendencji centralnej w skali interwałowej wykorzystujemy najczęściej średnią arytmetyczną. Dla danych homogenicznych, lecz z wyraźną skośnością rozkładu, optymalną miarą tendencji centralnej będzie mediana. Dla rozkładów niehomogenicznych stosujemy wartości modalne. Mimo, że każda z wymienionych wielkości jest miarą tendencji centralnej, to jednak uwypukla inne cechy badanej próby. Średnia arytmetyczna jest niezwykle czuła na wartości pomiarów skrajnych, mediana jest mało stabilna przy przejściu od próby do próby, wartość modalna może być niejednoznaczna. Podobnie jest w przypadku stosowania rozmaitych technik wielowymiarowych. Każda z nich może (ale nie musi) prowadzić do wyboru innego zestawu predyktorów, co więcej nawet takie same wybrane predyktory mogą mieć inną wagę. Często lekarze są zakłopotani tą różnorodnością uzyskiwanych wyników. A tymczasem sytuację można sobie intuicyjnie wyobrazić w sposób następujący. Drogi labiryntu przedstawionego na rysunku 12. reprezentują pomierzone przez nas parametry, wejścia do labiryntu – różne techniki badawcze. Do celu (tarcza) można dojść różnymi drogami. Jak widać, stosując jedną metodę zaangażujemy do osiągnięcia celu parametry od P1 do P9, używając innej metody do celu doprowadzą nas zmienne P12, P11, P10 oraz P6-P9. Pewne zmienne (P6-P9) są wspólne dla obu modeli, pewne zaś odmienne.

Kolejnym problemem jest określenie kosztów błędów klasyfikacyjnych. W rozważanym przykładzie założyliśmy po cichu, że popełnienie obu typów błędów (I i II rodzaju) jest równoważne. W praktyce często spotykamy się z sytuacją, w której popełnienie jednego z wymienionych błędów jest bardziej brzemienne w skutkach niż drugiego. Dla naszego przykładu popełnienie błędu I rodzaju polega na zaklasyfikowaniu pacjentki do grupy ze zmianami złośliwymi, podczas gdy w rzeczywistości występowały u niej zmiany łagodne. Błąd II rodzaju występuje w sytuacji, gdy metodą USG stwierdzamy zmiany łagodne, tymczasem tak naprawdę pacjentka miała zmiany złośliwe.



Rys. 12. Ilustracja osiągnięcia tego samego celu różnymi drogami

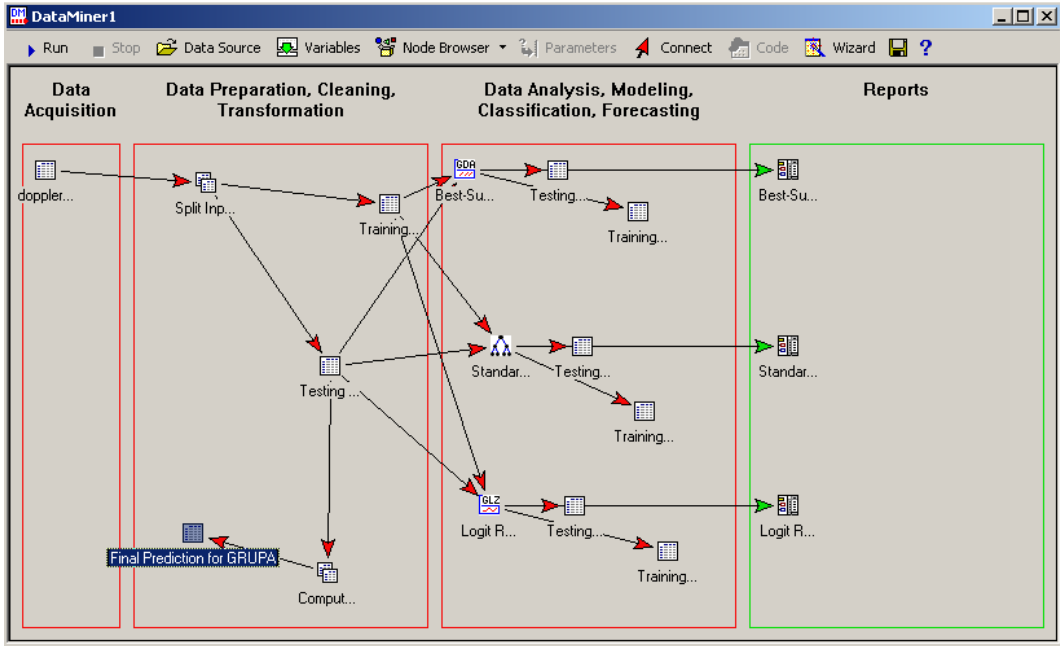
Niestety zmniejszenie prawdopodobieństwa popełnienia błędu I rodzaju pociąga za sobą wzrost prawdopodobieństwa popełnienia błędu II rodzaju i vice versa. Sytuacja ta jest częściowym analogiem znanej fizykom zasady nieoznaczoności Heisenberga. Jedynym sposobem jednoczesnego zmniejszenia obu rodzajów błędu jest zwiększenie liczebności badanej próby, co niestety w medycynie często jest niemożliwe. Dlatego też to lekarz (a nie statystyk) musi zawsze ocenić, który z błędów należy minimalizować. Gdy statystyk otrzyma taką informację, może modyfikować wagi, z jakimi oceniane są oba błędy. Dla przykładu, w panelu definiującym drzewa klasyfikacyjne w zakładce metody w opcji koszty błędnych klasyfikacji można wprowadzić w miejsce wartości domyślnych (1,1) wagi zdefiniowane przez użytkownika. Z błędem I i II rodzaju wiąże się jednoznacznie liczba przypadków prawdziwie i fałszywie pozytywnych, liczba przypadków prawdziwie i fałszywie negatywnych, a jak już wspomniano w rozdziale poświęconym analizom krzywych ROC, na ich podstawie oszacowuje się między innymi czułość i swoistość metody badawczej. Czułość odpowiada zdolności do detekcji osób rzeczywiście chorych na daną jednostkę chorobową, swoistość jest zdolnością do wykluczenia osób, u których nie występuje badana choroba.

Trzeci problem związany jest z prawidłowym oszacowaniem błędu klasyfikacyjnego. Omawiając krzywe ROC, analizę dyskryminacyjną i logistyczną wykorzystywaliśmy do oceny jakości klasyfikacji błędów typu post-hoc. Jak już wspomnieliśmy, omawiając drzewa klasyfikacyjne, ocena ta jest mało obiektywna i reprezentuje bardziej zdolność modelu do odtwarzania danych, na których model był konstruowany, niż do oceny jego prawdziwej zdolności klasyfikacyjnej dla danych, których model do tej pory nie widział. Podobnie jak w przypadku drzew można zastosować techniki walidacji krzyżowej lub



osobnego zbioru testującego do oceny rzeczywistej zdolności klasyfikacyjnej modeli zbudowanych przy użyciu wspomnianych analiz.

Najłatwiej zrealizować to, używając efektywnego narzędzia eksploracyjnego *Data Miner* opracowanego przez firmę StatSoft. Dla przykładu na rysunku 13 prezentujemy projekt zgłębiania danych opracowany w środowisku *STATISTICA Data Miner v. 6.1*.



Rys. 13. Projekt data mining opracowany w środowisku *STATISTICA Data Miner v.6.1*

Przestrzeń robocza podzielona jest na cztery strefy. W strefie Data Acquisition umieszczamy ikonę reprezentującą plik danych wejściowych. Łączymy ją z ikoną Split Input Data into Training and Testing Samples, która to reprezentuje losowy podział danych wejściowych na zbiór uczący (Training Data), oraz zbiór testujący (Testing File). Ikony wspomnianego węzła oraz wygenerowanych plików umieszczone są w drugiej strefie – Data Preparation, Cleaning and Transformation. Pierwszy ze zbiorów wykorzystywany jest do tworzenia modeli na bazie analizy dyskryminacyjnej (Best Subset and Stepwise GDA ANCOVA with Deployment), drzew klasyfikacyjnych (Standard Classification Trees with Deployment) oraz modelu logistycznego (Logit Regression with Deployment), które to węzły zostały umieszczone w strefie trzeciej – Data Analysis, Modelling, Classification and Forecasting. Wytworzone modele weryfikowane są przy użyciu zbioru testującego. Wyniki dotyczące każdej z analiz (m.in. parametry modelu, macierze klasyfikacji) umieszczone są w raportach znajdujących się w czwartej strefie Reports.

Każdy z modeli może być wykorzystany praktycznie jako klasyfikator do nowych danych, których reprezentujące ikony umieszczamy w strefie pierwszej. Należy wówczas koniecznie zaznaczyć opcję Data for deployed project; do not re-estimate models.



W strefie drugiej umieszczono ikonę Compute Best-Predicted Classification from All Models. Węzeł ten umożliwia automatyczne porównanie wyników klasyfikacji wykonanych przez każdy z użytych modeli oraz wybór odpowiedzi optymalnej.

Literatura

1. Dębniak B., *Ultrasonografia dopplerowska w ocenie ukrwienia złośliwych i niezłośliwych zmian gruczołów sutkowych*, Seminaria z Medycyny Perinatalnej, t.VII, Ośrodek Wydawnictw Naukowych, Poznań 2003.
2. Hanley J.A., Mc Neil B.J., *The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve*, Radiology 143:29-36, April 1982.
3. Morrison D.F., *Wielowymiarowa Analiza Statystyczna*, PWN, Warszawa 1990.
4. Hosmer DW, Lemeshow S, *Applied Logistic Regression*, Wiley, New York 1989.
5. Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J., *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software 1984.
6. Loh, W.-Y., & Shih, Y.-S., *Split selection methods for classification trees*. Statistica Sinica, 7, 815-840, 1997.
7. Loh, W.-Y., & Vanichestakul, N., *Tree-structured classification via generalized discriminant analysis*, Journal of the American Statistical Association, 83, 715-728, 1988.
8. Lim, T.-S., Loh, W.-Y., & Shih, Y.-S., *An empirical comparison of decision trees and other classification methods*, Technical Report 979, Department of Statistics, University of Wisconsin, Madison 1997.