KRZYWE ROC, CZYLI OCENA JAKOŚCI KLASYFIKATORA I POSZUKIWANIE OPTYMALNEGO PUNKTU ODCIĘCIA

Grzegorz Harańczyk, StatSoft Polska Sp. z o.o.

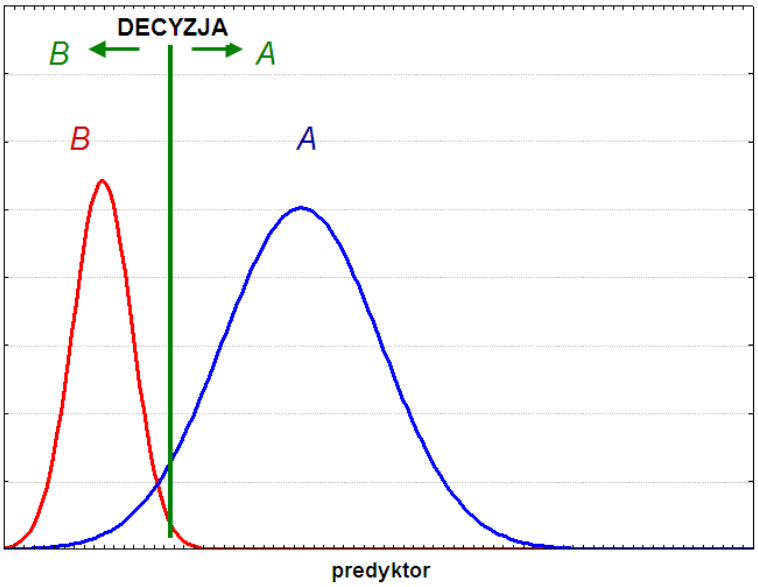
Krzywa ROC (*Receiver Operating Characteristic*) jest narzędziem do oceny poprawności klasyfikatora, zapewnia ona łączny opis jego czułości i specyficzności. Ten sposób wspomagania systemu decyzyjnego jest szeroko wykorzystywany w różnych zastosowaniach, również w diagnostyce medycznej. Zaprezentowane zostanie rozszerzenie programu *STATISTICA* do wykreślania krzywych ROC oraz wyznaczania optymalnego punktu odcięcia.

Zdefiniowanie problemu – model decyzyjny

Podejmowanie decyzji na ogół jest sprawą trudną, jeszcze bardziej komplikuje się właśnie w sytuacji podejmowania decyzji w warunkach niepewności. Jednym z zadań analizy danych są problemy klasyfikacyjne, polegające na znalezieniu zbioru reguł, tzw. modelu, na podstawie którego można obiekt przyporządkować do jednej z kilku klas. Budowanie modeli klasyfikujących jest bardzo powszechne w zastosowaniach medycznych. Stosowane są one między innymi do: diagnostyki na podstawie danych medycznych (np. zapis EKG, zdjęcia USG, dane laboratoryjne itp.), przypisywania badanych do grup ryzyka zachorowania na daną chorobę, wystąpienia powikłań podczas leczenia czy jako narzędzie wspomagające w przypisywaniu najefektywniejszej terapii.

My będziemy zajmować się problemem przynależności do dwóch klas (np. przyporządkowanie pacjenta do jednej z dwóch grup: zdrowy – chory). W szczególnym przypadku decyzja może być podejmowana na podstawie jednej zmiennej – wskaźnika diagnostycznego. Badamy wówczas zależność pomiędzy tą zmienną, najczęściej mierzoną na skali ilościowej, a wybraną zmienną dwustanową (np. wiek pacjenta a występowanie powikłań po zabiegu). Często zakłada się, że badana zależność jest monotoniczna (wraz ze wzrostem wartości zmiennej diagnostycznej rosną szanse na wystąpienie badanego zjawiska lub wraz ze wzrostem wartości zmiennej diagnostycznej maleją szanse na wystąpienie badanego zjawiska, nie ma zależności o „mieszanym" wpływie). Wówczas znalezienie reguły decyzyjnej sprowadza się do wybrania pewnej wartości czynnika diagnostycznego, która najlepiej dzieli badaną zbiorowość na dwie grupy: jedną, w której często występowało badane zdarzenie, i drugą, w której częstość występowania zdarzenia była mała (np. dla młodszych

pacjentów ryzyko powikłań po badanym zabiegu jest mniejsze, a od pewnego wieku wyższe - tu wiek jest czynnikiem diagnostycznym). W dalszej części zdefiniujemy miary mówiące o jakości reguły decyzyjnej, a wyznaczony na podstawie tych kryteriów punkt będziemy nazywali punktem odcięcia.



Rys. 1. Okno definiowania zakresu danych do analizy, wiersza z nazwami zmiennych i kolumny z nazwami przypadków.

W praktyce czasem pojawia się problem bardziej złożony – wnioskowanie na podstawie kilku zmiennych diagnostycznych. Sprowadza się on jednak również do omówionego powyżej. Przedstawimy teraz tę ogólną sytuację. Załóżmy, że mamy przewidzieć wartość zmiennej zależnej (określającej przynależność do jednej z dwóch klas), na podstawie wartości pewniej liczby zmiennych niezależnych – cech charakteryzujących badany obiekt (np. mężczyzna, wiek=65, HR=80). W tym kontekście te cechy nazywany również predyktorami. Reguła może być na przykład postaci: „jeśli mężczyzna, wiek>65, HR>90, …, to wystąpią powikłania po zabiegu”. Próbę, na podstawie której tworzy się regułę decyzyjną, nazywa się próbą uczącą. Oczywiście w praktyce zależy nam na tym, aby zbudowane reguły działały na całej populacji, z której została wybrana próba ucząca, a w szczególności dla nowych danych podczas stosowania modelu w przyszłości.

Jeśli mamy wiele zmiennych, które podejrzewamy o wpływ na wystąpienie wyróżnionego zdarzenia, to chcemy wykorzystać całą informację, jaką one niosą – zbudować reguły wykorzystujące wiele zmiennych jednocześnie. Jest wiele metod służących do budowania reguł tego typu, np. drzewa klasyfikacyjne, regresja logistyczna, metoda *k*-najbliższych sąsiadów, sieci neuronowe. W wyniku zastosowania takich modeli otrzymujemy prawdopodobieństwo przynależności do wybranej klasy (tzw. *scoring*). Zatem wartość zmiennej zależnej należy do przedziału *(0,1)*, toteż oprócz zbudowania modelu ważne jest również wybranie odpowiedniego punktu odcięcia, czyli takiej wartości *k* z przedziału *(0,1)*, że jeśli $y<k$, to obiekt przyporządkowywujemy do klasy kodowanej przez *0*, jeśli $y \geq k$, to do klasy kodowanej przez *1*.

Wynik takiego modelu możemy zatem również traktować jako wskaźnik diagnostyczny. Widzimy zatem, że na jednym z etapów problem sprowadza się do szukania optymalnego punktu odcięcia dla jednej zmiennej, tu dla wyniku modelu. Dzięki rozważaniu większej liczby predyktorów możemy zbudować dokładniejsze reguły decyzyjne, uwzględniające więcej cech badanych obiektów, czy też badać wpływ interesującej nas cechy skorygowany o inne cechy.

Błędne decyzje

Nieodłącznym elementem podejmowania decyzji jest popełnianie błędów (w tym przypadku złych zaklasyfikowań). Błędne decyzje modelu klasyfikującego są nieuniknione, ponieważ często klasy nie są całkowicie separowalne. Łatwo wyobrazić sobie taką sytuację, kiedy dwa obiekty charakteryzowane są za pomocą takich samych wartości zmiennych niezależnych, ale należą do dwóch różnych klas. Często wynika to z niepełnej wiedzy o badanym zjawisku lub po prostu niekompletnych danych.

Naszym celem jest przewidywanie klasy wyróżnionej i w związku z tym poprawne decyzje to: prawidłowe wskazanie wyróżnionej klasy (TP – *true positive*) oraz prawidłowe niewskazanie drugiej z klas (TN – *true negative*). Błędy popełniamy w sytuacji, gdy niepoprawnie wskazujemy wyróżniona klasę (FP – *false positve*) lub niewskazujemy klasy wyróżnionej w sytuacji, gdy powinniśmy ją wskazać (FN – *false negative*).

Tab. 1. Macierz klasyfikacji – stan faktyczny i wskazanie modelu.

	Zaobserwowano stan wyróżniony	Nie zaobserwowano stanu wyróżnionego
Przewidywano stan wyróżniony	TP	FP
Nie przewidywano stanu wyróżnionego	FN	TN

W powyższej tabeli TP, FP, FN oraz TN oznaczają liczbę obserwacji, które trafiły do danej komórki tabeli. Tabela taka podsumowywuje wyniki klasyfikacji dla danej reguły decyzyjnej, porównując stan faktyczny ze wskazaniem modelu (np. wynikiem testu diagnostycznego). Dobry model to taki, który minimalizuje liczbę błędów, czyli FN oraz FP. Nie zawsze oba te błędy traktowane są tak samo. W niektórych zastosowaniach te błędne klasyfikacje do dwóch klas mogą mieć bardzo różny koszt. Na przykład w klasyfikowaniu pacjentów do grup ryzyka gorszym błędem jest traktowanie chorego pacjenta jako zdrowego niż odwrotnie.

Miary mierzące wartość predykcyjną reguły decyzyjnej

Tak więc najlepsza reguła decyzyjna ma nam zapewnić najlepsze wyniki – jak najmniejszą liczbę błędów. Aby móc precyzyjnie zdefiniować odpowiednie kryterium wprowadza się miary jakości reguł decyzyjnych. Definiujemy zatem dwie główne miary: specyficzność (ang. *specificity*) oraz czułość (ang. *sensitivity*). Czułość definiujemy jako

$$\text{Czułość}=\frac{TP}{TP+FN},$$

natomiast specyficzność jako

$$\text{Specyficzność}=\frac{TN}{TN+FP}.$$

Wprowadza się także inne miary jakości reguł predykcyjnych. Najpopularniejsze z nich to: wartość predykcyjna dodatniego wyniku:

$$\text{PPV}=\frac{TP}{TP+FP}$$

wartość predykcyjna ujemnego wyniku:

$$\text{NPV}=\frac{TN}{TN+FN}$$

Jeśli dodatkowo zdefiniujemy częstość występowania wyróżnionego zdarzenia (np. choroby) (PV, ang. *prevalence*) jako

$$\text{PV}=\frac{TP+FN}{TP+TN+FN+FP},$$

to możemy zdefiniować tzw. skuteczność reguły decyzyjnej (ACC, ang. *accuracy*) jako

$$\text{ACC}=\frac{TP+TN}{TP+TN+FN+FP}=$$

$$=\frac{TP}{TP+FN}\cdot PV+\frac{TN}{TN+FP}\cdot(1-PV)=$$

$$=\text{Czułość PV+Specyficzność (1-PV)}$$

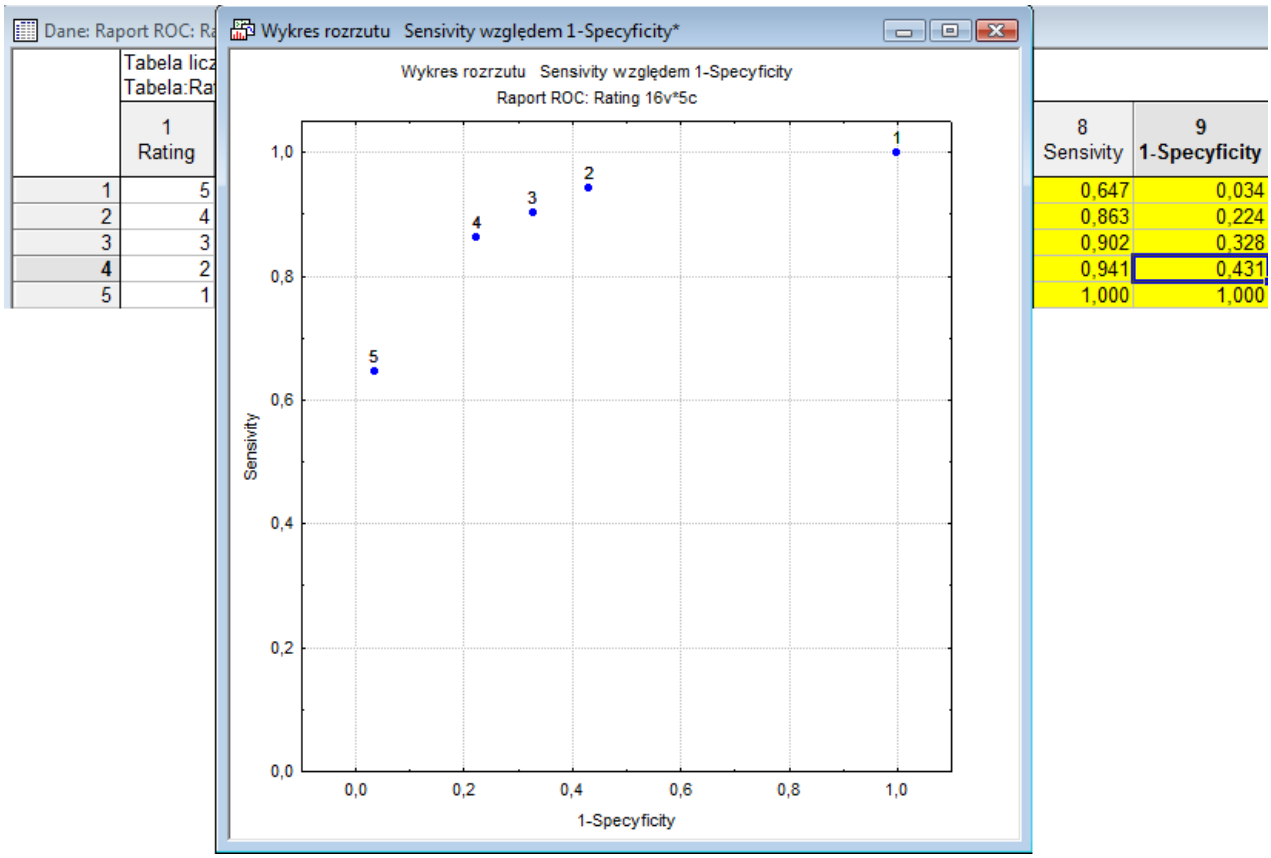
Definiujemy także iloraz wiarygodności (LR) jako

$$\text{LR}=\frac{TP}{TP+FN}\Big/\frac{FP}{FP+TN}.$$

Czułość i specyficzność są jednak najczęściej wykorzystywanymi miarami i to one są podstawą w konstrukcji krzywych ROC.

Konstrukcja krzywej ROC

Do tej pory rozważaliśmy związek pomiędzy dwoma zmiennymi dwuwartościowymi. Natomiast naszym celem jest zbadanie związku pomiędzy ciągłym predyktorem a zmienną dwustanową. Postępujemy następująco: dla każdej wartości wskazanego predyktora (pojedynczej zmiennej lub wyniku modelu) możemy stworzyć regułę decyzyjną, wybierając jako punkt odcięcia właśnie daną wartość. Każdą taką regułę oceniamy na podstawie kryterium specyficzności i czułości - dobra decyzja to taka, która maksymalizuje obie te wielkości. Jeśli mamy do czynienia z milionem obserwacji, to mamy milion potencjalnych punktów odcięcia, czyli milion potencjalnych tabel dwa na dwa do przeanalizowania, a mamy wybrać tę optymalną z najlepszy podziałem. Aby dokonać takiego wyboru, warto wykorzystać krzywe ROC, nie tylko po to, aby znaleźć optymalny punkt, ale również całościowo ocenić jakość skonstruowanej reguły decyzyjnej. Krzywa ROC ilustruje związek miedzy czułością a specyficznością dla danego modelu.



Rys. 2. Przykład konstrukcji krzywej ROC.

Krzywe ROC, jako technika analizy danych, zostały wprowadzone podczas II wojny światowej do analizy danych pochodzących z radarów. Ich zadaniem było pomagać operatorom

radarów zdecydować, czy zaobserwowany sygnał to wrogi czy sojuszniczy statek, czy też tylko szum. Po pięćdziesięciu latach krzywe ROC wykorzystywane są w wielu obszarach analizy danych [5,8], szczególnie popularne są w analizie danych medycznych.

Reasumując, dla każdego z możliwych punktów odcięcia obliczamy czułość i specyficzność, a następnie zaznaczamy otrzymane wyniki na wykresie. Tradycyjnie zaznaczamy je w układzie współrzędnych, gdzie na osi odciętych jest (1-specyficzność), a na osi rzędnych czułość (por. rys. 2). Uzyskane punkty ze sobą łączymy. Im więcej różnych wartości badanego wskaźnika, tym gładsza uzyskana krzywa.

Jeśli przyjmujemy równe koszty błędnych klasyfikacji, to optymalnym punktem odcięcia jest punkt krzywej ROC znajdujący się najbliżej punktu o współrzędnych (0,1). Punkt o współrzędnych (0,1) to punkt o czułości równej 1 (wszystkie obiekty wybranej klasy wykryto) i swoistości równej 1 (nie uznano błędnie żadnego obiektu za obiekt wyróżnionej klasy). Jeśli dla pewnego punktu odcięcia klasy są całkowicie separowane i wskazania modelu dobre, to krzywa ROC przechodzi przez ten punkt.

Może się zdarzyć, że model wskazuje klasy „na odwrót", wówczas krzywa ROC przebiega poniżej przekątnej $y=x$. Gdy rozkłady w obu grupach się pokrywają, wówczas krzywa ROC pokrywa się z przekątną $y=x$ (decyzja podejmowana na podstawie modelu jest tak samo dobra jak losowe wybieranie klasy dla danego obiektu).

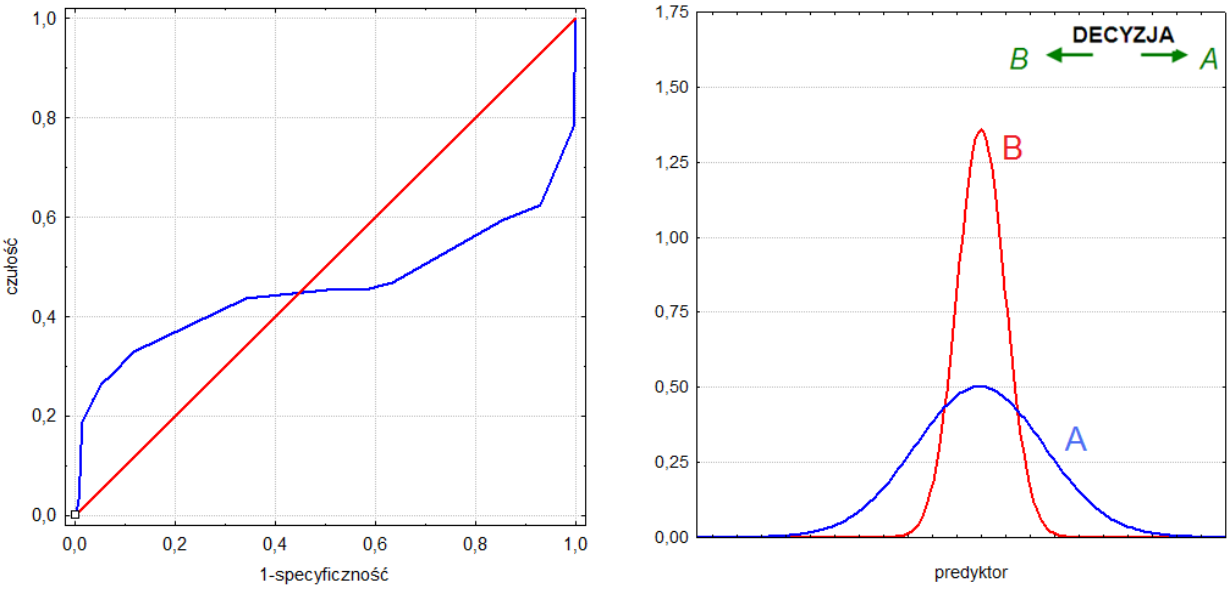
Pole pod wykresem krzywej ROC (AUC)

Krzywa ROC bywa często wykorzystywana jako narzędzie od oceny i porównywania między sobą modeli klasyfikacyjnych. Bardzo popularnym podejściem jest wyliczanie pola pod wykresem krzywej ROC, oznaczanego jako AUC (ang. *area under curve*), i traktowanie go jako miarę dobroci i trafności danego modelu [4, 2]. Wartość wskaźnika AUC przyjmuje wartości z przedziału [0,1]; im większa, tym lepszy model.

W ten sposób porównywanie dwóch krzywych można by ograniczyć do porównywania jedynie wskaźników AUC, bez wykonywania samych wykresów. Rodzi to od razu pytanie, jaka wartość AUC jest wystarczająco dobra dla modelu klasyfikacyjnego. Nie ma jednak uniwersalnej odpowiedzi na to pytanie, nie można powiedzieć, że wartość AUC na pewnym poziomie jest dobra albo zła, bo zależy to od dziedziny i specyfiki rozważanego problemu i jego trudności.

Dodatkowo, porównywanie samych wskaźników AUC nie jest najlepszym podejściem. Podobnie jak porównywanie samych współczynników korelacji liniowej, bez tworzenia odpowiednich wykresów rozrzutu (por. kwartet Anscombe'a [1]). Sama krzywa ROC, jej kształt i przebieg również daje pewne informacje o charakterze wpływu badanej zmiennej na wyróżniony stan [7]. Dla przykładu: możliwe jest zaobserwowanie krzywej jak po lewej stronie na rys. 3. Ten dość nietypowy przebieg mówi nam o sytuacji, w której dla pewnych wartości zmiennej diagnostycznej należy zmienić decyzję, tzn. stan wyróżniony A występuje dla niskich oraz dla wysokich wartości predyktora (w medycynie może to oznaczać po

prostu występowanie pewnych zdarzeń przy odchyleniu od normy, nieważne czy w górę, czy w dół). Taki uproszczony schemat przedstawiony jest po prawej stronie rys. 3.



Rys. 3. Krzywa ROC (po lewej) oraz sytuacja decyzyjna, na podstawie której została wygenerowana (po prawej).

Przykład analizy

W poniższym przykładzie zilustrowane zostanie działanie rozszerzenia programu *STATISTICA* do kreślenia krzywych ROC. Otwieramy plik danych ***ROCdata.sta***.

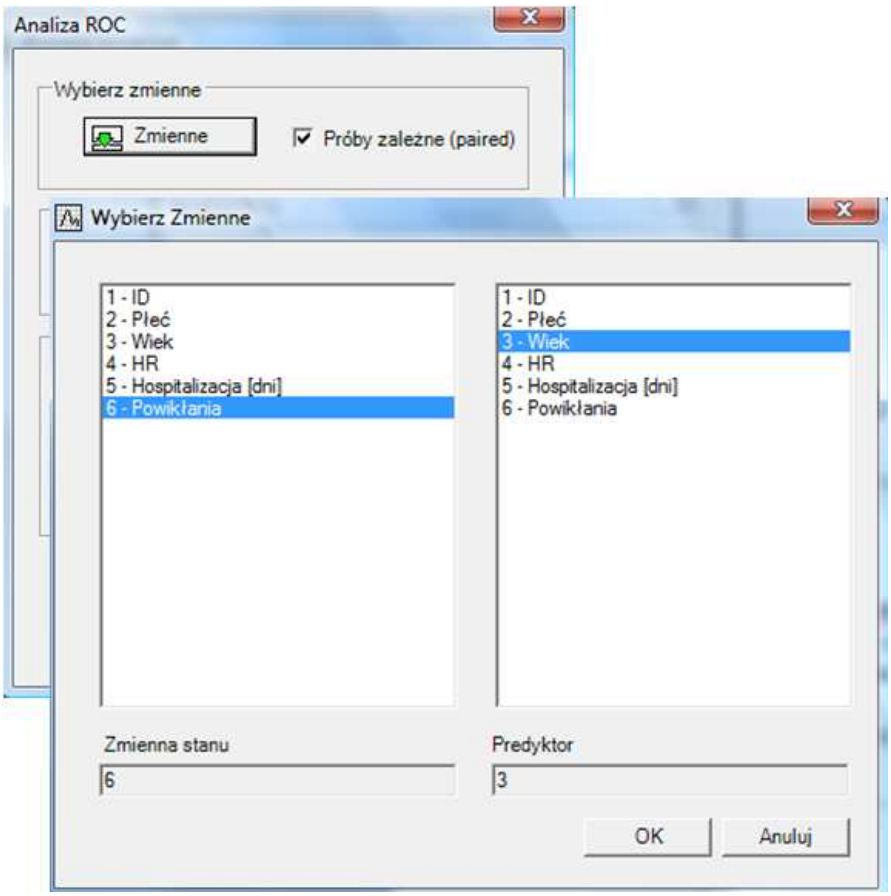
Dane: ROCdata.sta (6 zmn. * 1000 prz.)

	1 ID	2 Płeć	3 Wiek	4 HR	5 Hospitalizacja [dni]	6 Powikłania
4	4	mężczyzna	65	74	6	0
5	5	mężczyzna	76	97	10	0
6	6	mężczyzna	60	71	8	0
7	7	kobieta	52	81	4	0
8	8	kobieta	56	92	7	0
9	9	mężczyzna	73	91	4	0
10	10	kobieta	68	105	5	0
11	11	kobieta	67	91	4	0
12	12	mężczyzna	77	69	12	0
13	13	mężczyzna	89	76	8	0
14	14	mężczyzna	71	121	2	0
15	15	mężczyzna	64	76	7	0
16	16	kobieta	76	75	9	0
17	17	mężczyzna	51	71	8	0

Rys. 4. Arkusz z danymi wykorzystanymi podczas przykładowej analizy.

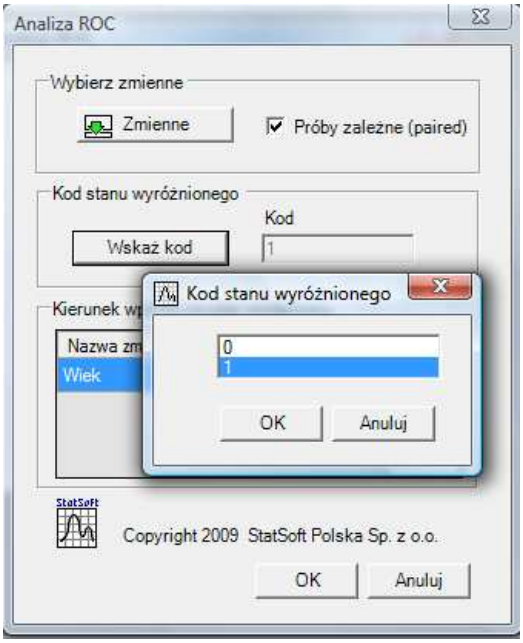
Plik ten zawiera informacje o 1000 pacjentach, którzy poddani byli pewnemu zabiegowi. Każdy pacjent scharakteryzowany jest za pomocą 4 zmiennych: płci, wieku, HR w momencie wykonywania zabiegu oraz liczby dni hospitalizacji po zabiegu. Dodatkowo w zbiorze zawarta jest informacja, czy podczas leczenia wystąpiły powikłania.

Klikamy w przycisk ***Krzywe ROC***. Klikając przycisk ***Zmienne***, określamy zmienne do analizy.



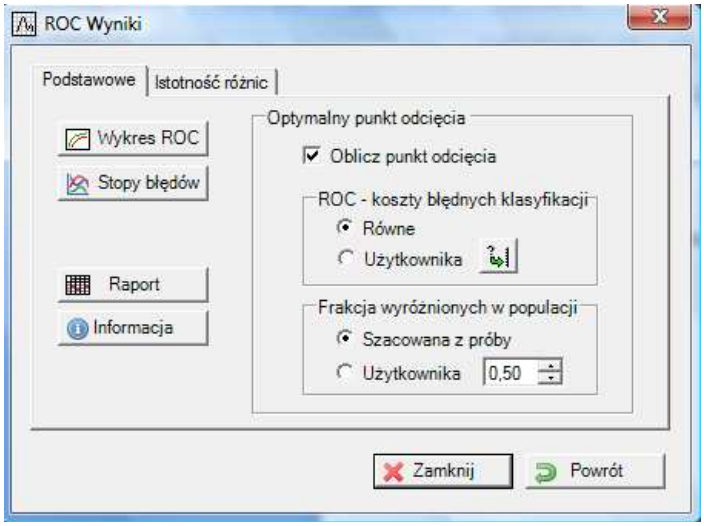
Rys. 5. Okno wyboru zmiennych.

Wybieramy jako zmienną stanu zmienną ***Powikłania***, natomiast jako predyktory zmienną ***Wiek.*** Następnie klikamy w przycisk ***Wskaż kod***, aby wskazać kod stanu wyróżnionego. Wybieramy kod ***1***. Na tym etapie musimy już zdecydować, czy ***Wiek*** jest stymulantą czy destymulantą, czyli czy podejrzewany, że większe wartości zmiennej wiek wpływają na częstsze występowanie stanu wyróżnionego, czy też nie.



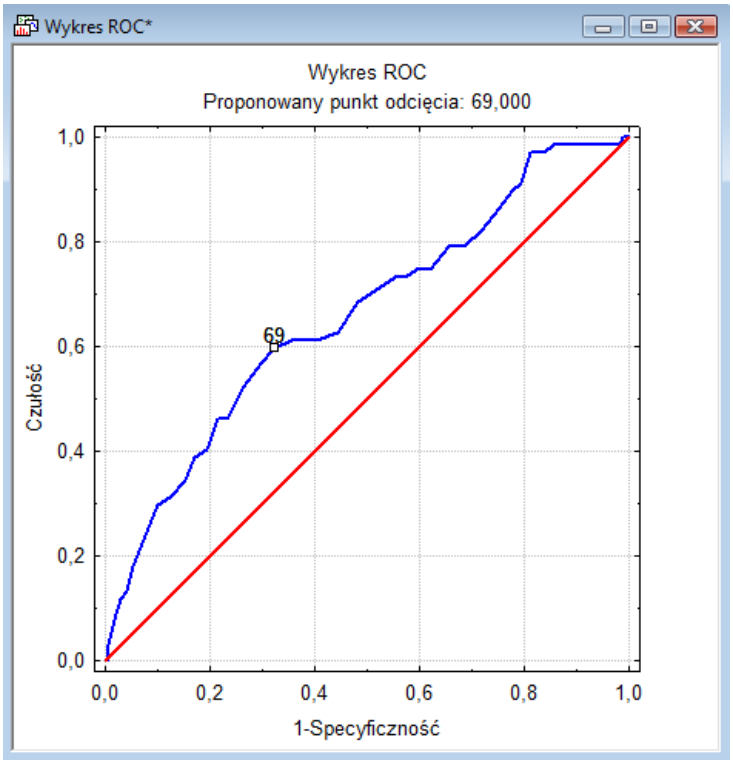
Rys. 6. Okno definiowania stanu wyróżnionego.

W polu ***Kierunek wpływu na stan wyróżniony*** określamy więc, czy wybrany przez nas czynnik jest stymulantą czy destymulantą. Klikamy ***OK*** i przechodzimy do karty z wynikami analizy.



Rys. 7. Okno z wynikami analizy.

Klikając na przycisk ***Wykres ROC***, otrzymujemy wykres ROC z wybranym optymalnym punktem odcięcia. Punkt ten wybrany jest dla równych kosztów błędnych klasyfikacji, możemy te koszty błędnych klasyfikacji zmodyfikować w polu ***ROC – koszty błędnych klasyfikacji – Użytkownika***.



Rys. 8. Arkusz z danymi wykorzystanymi podczas przykładowej analizy.

Klikając w przycisk ***Raport*** otrzymujemy między innymi wartość pola pod wykresem krzywej ROC (obliczane według [4]). Wybierając kilka zmiennych możemy porównać te zmienne jako klasyfikatory oraz przetestować istotność różnic pomiędzy polami pod wykresami dla tych klasyfikatorów (obliczane według [6]).

Podsumowanie

Oprócz wspomagania wyboru optymalnego punktu odcięcia krzywa ROC używana jest do porównywania różnych modeli, czy to zbudowanych na podstawie różnych zmiennych niezależnych czy też różnymi metodami. Zaletą tej metody jest to, że pokazuje siłę wpływu predyktora na występowanie wybranej klasy dla wszystkich możliwych punktów odcięcia. Zatem krzywe ROC są narzędziem do wyboru progu decyzyjnego, ale też narzędziem do wizualizacji całej sytuacji decyzyjnej.

Literatura

1. Anscombe F. J., *Graphs in Statistical Analysis*, The American Statistician, 27 (1973), 17-21.
2. Bradley A. P., *The use of the area under the ROC curve in the evaluation of machine learning algorithms*, Pattern Recognition, 1997, 30 (7), 1145-59.

3. Fisher R. A. *The Use of Multiple Measurements in Taxonomic Problems*, Annals of Eugenics, 1936, 7, 179–88.
4. Hanley J.A., McNeil B. J., *The meaning and use of the area under receiving operating characteristic (ROC) curve*, Radiology 1982, 43, 29-36.
5. Hanley J. A., *Receiver operating characteristic (ROC) methodology: the state of the art*, Crit Rev Diagn Imaging. 1989; 29(3), 307-35.
6. Hanley J.A., Hajian-Tilaki K.O., *Sampling Variability of Nonparametric Estimates of the Areas under Receiver Operating Characteristic Curves: An Update*, Academic Radiology, 1997, 4, 49-58.
7. Harańczyk G., Stępień M., *Ilustrowana sztuka podejmowania decyzji*, Matematyka Społeczeństwo Nauczanie, 41 (2008), 12-15.
8. Swets J. A., Dawes R. M., Monahan J. *Better decision through science*, Scientific American, 2000, October, 82-7.

