PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 328

Taksonomia 23
Klasyfikacja i analiza danych – teoria i zastosowania

Redaktorzy naukowi
Krzysztof Jajuga, Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2014

Redaktor Wydawnictwa: Barbara Majewska
Redaktor techniczny: Barbara Łopusiewicz
Korektor: Barbara Cibis
Łamanie: Beata Mazur
Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
w Dolnośląskiej Bibliotece Cyfrowej www.dbc.wroc.pl,
The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon
http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 328 • 2014
Taksonomia 23 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania

Małgorzata Misztal
Uniwersytet Łódzki

WYBRANE METODY OCENY JAKOŚCI KLASYFIKATORÓW – PRZEGLĄD I PRZYKŁADY ZASTOSOWAŃ

Streszczenie: W przypadku zagadnienia klasyfikacji obiektów do dwóch klas popularnym narzędziem oceny i porównywania różnych modeli klasyfikacyjnych jest krzywa ROC oraz wielkość pola pod krzywą (AUC). W ostatnich latach w publikacjach o tematyce medycznej pojawiło się kilka nowych metod pozwalających ocenić zdolność predykcyjną klasyfikatorów. Wymienić tu należy zaproponowaną przez Cook [2008] metodę reklasyfikacji (*Reclassification*) oraz zaproponowane przez Pencinę i in. [2008] wskaźniki: NRI (*Net Reclassification Improvement*) i IDI (*Integrated Discrimination Improvement*). W artykule zwięźle scharakteryzowano wymienione metody oraz zaprezentowano ich możliwości aplikacyjne.

Słowa kluczowe: jakość klasyfikatora, krzywa ROC, reklasyfikacja.

1. Uwagi wstępne

Zagadnienie klasyfikacji obiektów do jednej z dwóch wyróżnionych klas jest najczęściej spotykane w praktycznych zastosowaniach metod klasyfikacji. Wymienić tu można np. klasyfikację pacjentów do grupy wysokiego ryzyka (np. zagrożonej zgonem) i do grupy niskiego ryzyka, klasyfikację kredytobiorców do grupy zagrożonej windykacją i grupy kredytów spłacanych czy klasyfikację przedsiębiorstw do grupy zagrożonej lub niezagrożonej bankructwem. Zauważmy też, że każde zagadnienie klasyfikacji do większej liczby klas można sprowadzić do klasyfikacji binarnej lub zestawu zagadnień klasyfikacji binarnych.

W praktyce dysponujemy zbiorem uczącym, złożonym z obiektów opisanych zestawem zmiennych objaśniających, których przynależność do klas znamy. Na podstawie wartości tych zmiennych budowana jest reguła decyzyjna (klasyfikacyjna) – klasyfikator. Reguła decyzyjna sprowadza pomiar wielowymiarowy do pojedynczej wartości – może to być etykieta klasy, prawdopodobieństwo należenia obiektu do klasy lub wartość skoringowa.

Jeżeli w wyniku zastosowania klasyfikatora dostajemy wartość liczbową (prawdopodobieństwo lub skoring), mówimy o klasyfikatorze ciągłym. W tym

przypadku przypisanie obiektu do klasy polega na porównaniu tej wartości liczbowej z pewną wartością progową (punktem odcięcia) – obiekty z wynikiem powyżej punktu odcięcia przypisywane są do jednej klasy, a obiekty z wynikiem poniżej punktu odcięcia do drugiej klasy.

Jeżeli w wyniku zastosowania klasyfikatora dostajemy numer (etykietę) klasy, mówimy o klasyfikatorze dyskretnym. W przypadku klasyfikatora dyskretnego istnieje zwykle możliwość uzyskania wartości skoringowych czy prawdopodobieństw należenia obiektów do klas – np. szczegółowe procedury dla drzew klasyfikacyjnych przedstawiają Provost i Domingos [2001].

Istotnym problemem w zagadnieniach klasyfikacji jest ocena jakości klasyfikatora, przy czym jakość najczęściej rozumiana jest jako zdolność prognostyczna klasyfikatora (zdolność do przewidywania w poprawny sposób przynależności obiektów do badanych klas). Często pojawia się także problem porównywania modeli klasyfikacyjnych po uwzględnieniu w modelu nowych (dodatkowych) zmiennych objaśniających.

Do oceny zdolności predykcyjnej klasyfikatorów najczęściej wykorzystywany jest ogólny błąd klasyfikacji. Popularnym narzędziem jest także krzywa ROC oraz wielkość pola pod krzywą ROC (AUC). W ostatnich latach w publikacjach o tematyce medycznej pojawiło się kilka nowych metod oceniających jakość klasyfikatorów dla różnych zestawów zmiennych i pozwalających porównywać je między sobą. Wymienić tu należy metodę reklasyfikacji (*Reclassification*) zaproponowaną przez Cook [2008] oraz mierniki: NRI (*NetReclassificationImprovement*) i IDI (*IntegratedDiscriminationImprovement*) Penciny i in. [2008].

Celem pracy jest prezentacja wybranych metod oceny jakości klasyfikatorów wraz z krótkim omówieniem ich wad i zalet. Rozważania zilustrowano przykładami zastosowań omawianych metod. Do obliczeń wykorzystano środowisko R.

2. Macierz klasyfikacji i miary oceniające wartość predykcyjną klasyfikatora

W przypadku dwóch klas wyniki zastosowania reguły klasyfikacyjnej przedstawia się zwykle w postaci tzw. macierzy klasyfikacji (zwanej też macierzą pomyłek lub kontyngencji – *classification*/*confusion*/*contingency matrix*), w której stan obserwowany porównywany jest ze wskazaniem reguły decyzyjnej – por. tab. 1.

Dwie najczęściej wykorzystywane miary jakości reguły decyzyjnej to dokładność (*accuracy – ACC*) oraz błąd klasyfikacji (*missclassification/error rate – ERR*):

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} = 1 - ERR, \tag{1}$$

$$ERR = \frac{FP+FN}{TP+TN+FP+FN} = 1 - ACC. \tag{2}$$

Tabela 1. Schemat macierzy klasyfikacji

Stan przewidywany	Stan obserwowany		Σ
	klasa wyróżniona (P)	klasa niewyróżniona (N)	
Klasa wyróżniona (P)	TP (*true positives*) obiekty z klasy wyróżnionej poprawnie zaklasyfikowane	FP (*false positives*) obiekty z klasy niewyróżnionej błędnie zaklasyfikowane	TP+FP
Klasa niewyróżniona (N)	FN (*false negatives*) obiekty z klasy wyróżnionej błędnie zaklasyfikowane	TN (*true negatives*) obiekty z klasy niewyróżnionej poprawnie zaklasyfikowane	FN+TN
Σ	TP+FN	FP+TN	TP+FP+FN+TN

Źródło: opracowanie własne.

Krzanowski i Hand [2009] przytaczają wyniki badań, w których pokazano, że w większości publikacji dotyczących jakości reguł decyzyjnych wnioski podejmowane są właśnie na podstawie porównania błędów klasyfikacji poszczególnych klasyfikatorów.

Autorzy ci jednocześnie wymieniają kilka powodów, dla których błąd klasyfikacji nie może być dobrym miernikiem wartości predykcyjnej klasyfikatora. Po pierwsze, błędne klasyfikacje obiektów z klasy wyróżnionej P jako obiektów z klasy niewyróżnionej N i odwrotnie są tak samo traktowane, a przecież czym innym jest np. zaklasyfikowanie osoby chorej jako zdrowej i zaklasyfikowanie osoby zdrowej jako chorej. Po drugie, miara ta nie uwzględnia różnych prawdopodobieństw *a priori* należenia obiektów do klas oraz problemu klas niezrównoważonych. Problemy te można częściowo rozwiązać poprzez uwzględnienie różnych wag lub kosztów błędnych klasyfikacji (o ile oczywiście potrafimy ocenić te kosz-

Tabela 2. Wybrane miary oceniające zdolność predykcyjną reguły klasyfikacyjnej

Miara (nazwa /nazwy)	Wzór
True Positives Rate/Sensitivity (czułość)/*Recall* (pamięć)	$TPR = \frac{TP}{TP + FN}$
False Positives Rate	$FPR = \frac{FP}{FP + TN}$
Specificity (swoistość)	$Specificity = \frac{TN}{FP + TN} = 1 - FPR$
Positive Predictive Value (dodatnia zdolność predykcyjna)/ *Precision* (precyzja)	$PPV = \frac{TP}{TP + FP}$
Negative Predictive Value (ujemna zdolność predykcyjna)	$NPV = \frac{TN}{TN + FN}$
F1-measure (miara F1)	$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$

Źródło: opracowanie własne na podstawie [Fawcett 2006; Fielding 2007].

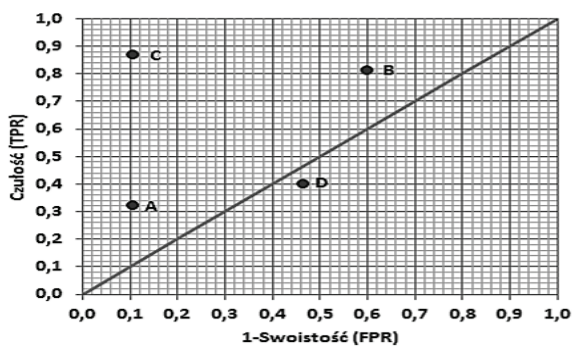
ty). Można również rozszerzyć analizę jakości klasyfikatora poprzez podanie innych niż (1) i (2) miar jakości. Przykładowe miary obliczane w oparciu o wielkości przedstawione w macierzy klasyfikacji (tab. 1) przedstawia tab. 2.

Z miar wymienionych w tabeli 2 najczęściej stosowane są dwie – czułość (zdolność klasyfikatora do przewidywania klasy wyróżnionej P) i swoistość (zdolność klasyfikatora do wykluczania obiektów z klasy P spośród obiektów z klasy niewyróżnionej N). Miary te są podstawą konstrukcji krzywych ROC (*Receiver Operating Characteristic*).

3. Przestrzeń i krzywa ROC

Krzywa ROC[1] przedstawia na dwuwymiarowym wykresie zmiany w czułości i swoistości, przy czym przyjęło się na osi odciętych przedstawiać wartości FPR (1-swoistość), a na osi rzędnych wartości TPR (czułość).

W przypadku klasyfikatorów dyskretnych na wykresie przedstawiany jest pojedynczy punkt w przestrzeni ROC. Przykładowe klasyfikatory dyskretne w przestrzeni ROC prezentuje rysunek 1. Perfekcyjną klasyfikację obrazuje punkt o współrzędnych (0, 1). Punkty leżące na przekątnej odpowiadają klasyfikacji losowej. Jeden punkt w przestrzeni ROC jest „lepszy" od drugiego, jeżeli jest położony bardziej na „północny zachód" (C). Punkty położone bliżej lewej strony i bliżej osi X opisują klasyfikatory bardziej konserwatywne (A), a punkty położone bliżej górnej prawej strony – klasyfikatory bardziej liberalne (B). Punkt leżący poniżej przekątnej świadczy o klasyfikacji gorszej niż losowe przydzielanie obiektów do klas (D).

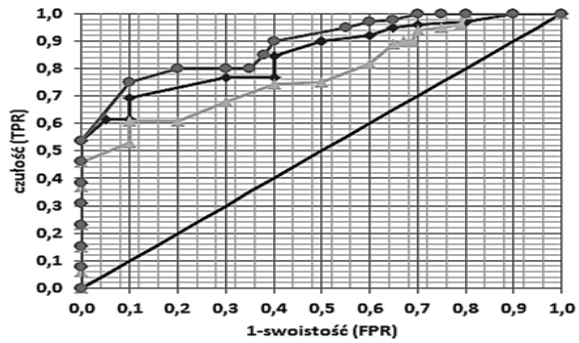


Rys. 1. Porównanie czterech przykładowych klasyfikatorów dyskretnych w przestrzeni ROC

Źródło: opracowanie własne.

[1] Ze względu na ograniczoną objętość pracy podano tylko podstawowe informacje dotyczące krzywych ROC. Szczegółowe omówienie problematyki znaleźć można m.in. w pracy Fawcetta [2006] oraz w monografii Krzanowskiego i Handa [2009].

W przypadku klasyfikatorów ciągłych dla każdego punktu odcięcia można obliczyć wartości TPR i FPR, umieścić je na wykresie, a następnie połączyć, uzyskując krzywą ROC. Przykładowe krzywe ROC prezentuje rysunek 2. Im wyżej położona krzywa ROC, tym lepsza zdolność predykcyjna reguły decyzyjnej. Optymalny punkt odcięcia można wyznaczyć np. na podstawie wskaźnika Youdena: Y=TPR-FPR [Youden 1950].



Rys. 2. Porównanie trzech przykładowych klasyfikatorów ciągłych z wykorzystaniem krzywych ROC

Źródło: opracowanie własne.

Aby porównać kilka klasyfikatorów, wygodniej jest posługiwać się jedną wielkością liczbową. W tym celu oblicza się wielkość pola pod krzywą ROC – AUC (*area under curve*). Wartość AUC jest to prawdopodobieństwo, że klasyfikator wyżej oceni losowo wybrany obiekt z klasy wyróżnionej niż losowo wybrany obiekt z klasy niewyróżnionej [Fawcett 2006]. Można wykazać, że wartość AUC jest równoważna wartości statystyki U testu Manna – Whitneya. AUC przyjmuje wartości z przedziału [0, 1], przy czym: AUC=1 oznacza doskonałą zdolność predykcyjną klasyfikatora, natomiast AUC<0,5 – brak zdolności predykcyjnej. Szczegółowe procedury estymacji wielkości AUC oraz testy statystyczne weryfikujące hipotezy o istotności AUC przedstawiają np. Krzanowski i Hand [2009].

Pole AUC może być wykorzystywane do porównania zdolności predykcyjnej różnych klasyfikatorów lub do oceny jakości tego samego klasyfikatora przed i po uwzględnieniu w modelu dodatkowych zmiennych. Niestety, jeżeli różnice w wielkości AUC są niewielkie (lub nieistotne statystycznie), pojawia się problem z wyborem optymalnego modelu klasyfikacyjnego. Stąd też w ostatnich latach w publikacjach medycznych zaproponowano nowe metody umożliwiające porównanie zdolności predykcyjnych różnych modeli.

4. Reklasyfikacja

Aby porównać dwa modele klasyfikacyjne Cook [2007, 2008] zaproponowała utworzenie tablicy reklasyfikacji (*reclassification table*), przedstawiającej klasyfi-

kację dla modelu początkowego i modelu z dodatkowym nowym predyktorem. Celem reklasyfikacji jest sprawdzenie, jak wiele obiektów zmieniło przynależność do poszczególnych grup. Jednak wysoki odsetek obiektów zmieniających przynależność do grupy nie musi oznaczać poprawy zdolności predykcyjnej modelu. Pojawia się zatem konieczność nadzorowania, czy reklasyfikacja przebiegała w dobrym kierunku. Istotność jakości reklasyfikacji można zweryfikować z wykorzystaniem modyfikacji testu Hosmera-Lemeshowa [Cook 2008].

Pencina i in. [2008] rozszerzyli ideę reklasyfikacji, proponując stosowanie wskaźnika NRI (*Net Reclassification Improvment*) i zwracając uwagę na fakt, że reklasyfikacja obiektów z grupy wyróżnionej (P) i niewyróżnionej (N) powinna być rozważana osobno. W grupie wyróżnionej zmiana „w górę – ↑" oznacza poprawę klasyfikacji (przejście do wyższej kategorii), a zmiana „w dół – ↓" – pogorszenie (przejście do niższej kategorii). W grupie niewyróżnionej – odwrotnie. Wskaźnik NRI, będący sumą wskaźników obliczanych osobno dla każdej z analizowanych grup, wyznacza się na podstawie próby uczącej w następujący sposób[2]:

$$\widehat{\mathrm{NRI}} = \widehat{NRI_P} + \widehat{NRI_N} = \left(\widehat{\mathrm{P}}(\uparrow/\mathrm{P}) - \widehat{\mathrm{P}}(\downarrow/\mathrm{P})\right) + \left(\widehat{\mathrm{P}}(\downarrow/\mathrm{N}) - \widehat{\mathrm{P}}(\uparrow/\mathrm{N})\right), \qquad (3)$$

gdzie:

$$\widehat{\mathrm{P}}(\uparrow/\mathrm{P}) = \frac{\text{liczba obiektów grupy P reklasyfikowanych} \uparrow}{\text{liczebność grupy P}}, \qquad (4)$$

$$\widehat{\mathrm{P}}(\downarrow/\mathrm{P}) = \frac{\text{liczba obiektów grupy P reklasyfikowanych} \downarrow}{\text{liczebność grupy P}}, \qquad (5)$$

$$\widehat{\mathrm{P}}(\uparrow/\mathrm{N}) = \frac{\text{liczba obiektów grupy N reklasyfikowanych} \uparrow}{\text{liczebność grupy N}}, \qquad (6)$$

$$\widehat{\mathrm{P}}(\downarrow/\mathrm{N}) = \frac{\text{liczba obiektów grupy N reklasyfikowanych} \downarrow}{\text{liczebność grupy N}}, \qquad (7)$$

Wskaźnik NRI przyjmuje wartości od -2 (jeżeli w każdej grupie zaobserwowano po 100% reklasyfikacji w nieprawidłowym kierunku) do 2 (gdy w każdej grupie zaobserwowano po 100% reklasyfikacji w prawidłowym kierunku). Interpretując wartości NRI, należy uwzględnić informacje uzyskane osobno dla każdej grupy; każda z uzyskanych wartości ocenia zysk netto z reklasyfikacji w badanej grupie.

Istnieje także wersja ciągła wskaźnika NRI, w której nie uwzględnia się podziału na grupy, natomiast porównywane są zmiany w wielkościach szacowanych prawdopodobieństw.

Kolejnym wskaźnikiem zaproponowanym przez Pencinę i in. [2008] jest IDI (*Integrated Discrimination Improvement*). Miernik ten, najogólniej biorąc, ocenia zmiany (wzrost) przeciętnej czułości modelu (TPR) po uwzględnieniu nowego

[2] W pracy przyjęto, że oznaczenia NRI i IDI dotyczą wskaźników w zbiorowości generalnej, natomiast $\widehat{NRI}$ i $\widehat{IDI}$ oznaczają oszacowania tych wskaźników uzyskane z próby uczącej.

predyktora, przy założeniu, że przeciętna swoistość nie zmniejszy się [por. Pencina i in. 2008, 2011]. Wskaźnik IDI wyznaczyć można na podstawie próby uczącej według wzoru:

$$\widehat{\mathrm{IDI}} = \left(mean\left(\hat{p}_{new,P}\right) - mean\left(\hat{p}_{old,P}\right)\right) - \left(mean\left(\hat{p}_{new,N}\right) - mean\left(\hat{p}_{old,N}\right)\right), \quad (8)$$

gdzie $mean\left(\hat{p}_{old,P}\right)$ i $mean\left(\hat{p}_{old,N}\right)$ oznaczają średnie z oszacowanych prawdopodobieństw dla grupy wyróżnionej i niewyróżnionej dla modelu początkowego, natomiast $mean\left(\hat{p}_{new,P}\right)$ i $mean\left(\hat{p}_{new,N}\right)$ – średnie z oszacowanych prawdopodobieństw dla grupy wyróżnionej i niewyróżnionej dla modelu z dodatkowym predyktorem.

Wskaźnik IDI przyjmuje wartości z przedziału [-1; 1]. Miara ta uwzględnia zmiany prawdopodobieństw przynależności obiektów do poszczególnych klas i może być interpretowana jako różnica sił dyskryminacyjnych (*di-scrimination slopes*) porównywanych modeli[3].

Wygodniej jest interpretować względną postać wskaźnika IDI (*Relative* IDI) postaci:

$$\widehat{Relative\ IDI} = \frac{mean\left(\hat{p}_{new,P}\right) - mean\left(\hat{p}_{new,N}\right)}{mean\left(\hat{p}_{old,P}\right) - mean\left(\hat{p}_{old,N}\right)}. \quad (9)$$

Dla obu wskaźników autorzy przedstawili także testy istotności dla weryfikacji hipotez zerowych postaci NRI = 0 oraz IDI = 0 [por. Pencina i in. 2008].

Wskaźniki NRI i IDI zostały zaproponowane jako miary oceniające poprawę zdolności predykcyjnej modelu klasyfikacyjnego po dodaniu do niego nowej zmiennej objaśniającej. Możliwe jest jednak wykorzystanie tych wskaźników do porównania dwóch reguł klasyfikacyjnych zbudowanych w oparciu o ten sam zestaw zmiennych objaśniających. Takie podejście zaproponowano w poniższym przykładzie.

5. Przykład

Analizie poddano dwie grupy kredytobiorców[4] (kredyty spłacane (klasa niewyróżniona)/ kredyty windykowane (klasa wyróżniona)) opisanych zestawem 6 zmiennych objaśniających: (x_1) – wiek, (x_2) – kwota kredytu, (x_3) – staż pracy, (x_4) – średni dochód z 3 miesięcy, (x_5) – rata kredytu oraz (x_6) – okres udzielenia kredytu.

[3] Siła dyskryminacyjna modelu (*di-scrimination slope*) to różnica między średnimi z oszacowanych prawdopodobieństw przynależności obiektów do klasy wyróżnionej i niewyróżnionej.

[4] Dane pochodzą z badań własnych i dotyczą łącznie 467 kredytobiorców. Na potrzeby pracy wylosowano 200 kredytobiorców (po 100 z każdej grupy), a następnie dokonano losowego podziału na próbę uczącą i testową (w każdej próbie po 50 kredytobiorców spłacających kredyt i podlegających procedurze windykacji).

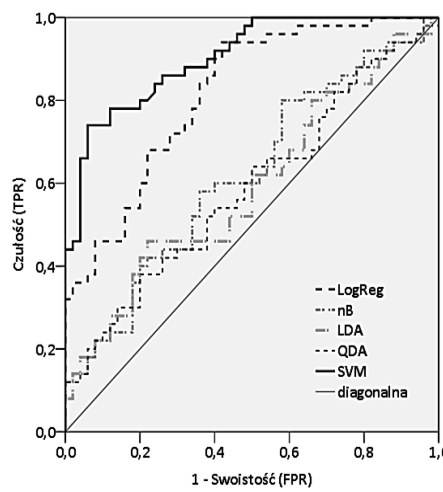
Na podstawie próby uczącej zbudowano 5 modeli klasyfikacyjnych wykorzystując regresję logistyczną (LogReg), naiwny klasyfikator Bayesa (nB), liniową i kwadratową funkcję dyskryminacyjną (LDA, QDA) oraz metodę wektorów nośnych (SVM)[5]. Zdolność predykcyjną modeli oceniano z wykorzystaniem próby testowej.

W tabeli 3 przedstawiono wartości wybranych miar jakości klasyfikatorów[6]. Na rysunku 3 wykreślono krzywe ROC dla uzyskanych modeli klasyfikacyjnych. Wartości pola pod krzywą AUC i ich charakterystykę przedstawiono w tabeli 4.

Tabela 3. Wartości wybranych miar jakości klasyfikatorów

Klasyfikator	Miara				
	ACC	TPR	1-FPR	PPV	NPV
LogReg	70%	72%	68%	69,2%	70,8%
nB	58%	60%	56%	57,7%	58,3%
LDA	55%	62%	48%	54,4%	55,8%
QDA	56%	52%	60%	56,5%	55,6%
SVM	79%	80%	78%	78,4%	79,6%

Źródło: opracowanie własne.



Rys. 3. Krzywe ROC dla analizowanych modeli klasyfikacyjnych

Źródło: opracowanie własne.

[5] Obliczenia wykonano w środowisku R z wykorzystaniem pakietów: rms, e1071, MASS.

[6] Jako wartość progową przyjęto 0,5. Ze względu na ograniczoną objętość pracy pominięto rozważania dotyczące wyznaczania innej wartości progowej.

Tabela 4. Charakterystyka pola pod krzywą ROC

Klasyfikator	AUC	SE	Poziom p	95%CI dla AUC	
LogReg	0,816	0,041	*0,0000*	0,735	0,897
nB	0,618	0,056	*0,0420*	0,508	0,728
LDA	0,592	0,057	0,1140	0,480	0,703
QDA	0,589	0,057	0,1260	0,477	0,700
SVM	0,904	0,029	*0,0000*	0,848	0,960

Źródło: opracowanie własne.

Opierając się wyłącznie na dokładności (ACC) badanych klasyfikatorów (tab. 3), za najlepszą uznać należy metodę wektorów nośnych (SVM), a następnie model regresji logistycznej. Przewagę tych metod nad pozostałymi widać wyraźnie na rysunku 3 – najwyżej położone krzywe ROC. Wartości AUC (tab. 4) potwierdzają zasadność wyboru obu tych metod do klasyfikacji kredytobiorców, przy czym metoda wektorów nośnych daje zdecydowanie najlepsze wyniki (wielkość AUC dla tej metody jest istotnie wyższa ($p < 0{,}001$) w porównaniu z wielkością AUC dla modelu regresji logistycznej). Najniższą dokładność w klasyfikacji kredytobiorców uzyskano po zastosowaniu liniowej funkcji dyskryminacyjnej (LDA).

W tabeli 5 przedstawiono wyniki porównania najlepszego klasyfikatora (SVM) z pozostałymi klasyfikatorami z wykorzystaniem wskaźników NRI i IDI opartych na reklasyfikacji.

Najwyższą wartość wskaźnika NRI uzyskano, porównując najlepszy i najgorszy klasyfikator (SVM vs LDA). Zmiany kategorii w obu grupach miały podobny wpływ na wartość wskaźnika ($\widehat{\text{NRI}}$ dla kredytobiorców spłacających wynosi 28%, a dla windykowanych 20%). Dodatnie wartości $\widehat{\text{NRI}}$ w obu grupach świadczą o przewadze zmian w prawidłowym kierunku.

Tabela 5. Miary NRI i IDI dla porównywanych modeli klasyfikacyjnych

Wskaźnik	Porównywane klasyfikatory			
	SVM vs LogReg	SVM vs nB	SVM vs QDA	SVM vs LDA
$\widehat{\textbf{NRI}}$ dla grupy kredytów windykowanych	0,10	0,22	0,30	0,20
$\widehat{\textbf{NRI}}$dla grupy kredytów spłacanych	0,08	0,20	0,16	0,28
$\widehat{\textbf{NRI}}$	0,18	0,42	0,46	0,48
95%CI dla NRI	0,0324 – 0,3276	0,1703 – 0,6697	0,2083 – 0,7117	0,2257 – 0,7343
p	*0,0169*	*0,0010*	*0,0003*	*0,0002*
$\widehat{\textbf{IDI}}$	–0,044	0,143	0,182	0,194
95%CI dla IDI	–0,0961 – 0,0081	–0,0081 – 0,294	0,0261 – 0,3369	0,0603 – 0,3267
p	0,0978	0,0636	*0,0221*	*0,0044*
$\widehat{\textbf{\textit{Relative IDI}}}$	0,8625	2,0748	3,3461	2,9204

Źródło: opracowanie własne w pakiecie `predictABEL` [Kundu i in. 2011].

Wartości miernika $\widehat{\mathrm{IDI}}$ oraz $\widehat{RelativeIDI}$ potwierdzają przewagę metody SVM nad liniową funkcją dyskryminacyjną. Siła dyskryminacyjna klasyfikatora SVM była wyższa o 19 punktów procentowych w porównaniu z siłą dyskryminacyjną klasyfikatora LDA. Z kolei $\widehat{RelativeIDI}$ = 2,92 wskazuje na prawie trzykrotny wzrost siły dyskryminacyjnej modelu SVM w stosunku do LDA.

6. Podsumowanie

Problematyka oceny jakości klasyfikatorów i sposobów ich porównywania staje się obecnie niezmiernie istotna ze względu na coraz powszechniejsze zastosowania modeli klasyfikacyjnych. Przedstawione w pracy metody pozwalają lepiej i dokładniej ocenić oraz porównać zdolność predykcyjną różnych klasyfikatorów.

W przypadku krzywych ROC wygodna jest możliwość graficznej prezentacji wyników oraz ocena jakości klasyfikatora za pomocą jednej wielkości – pola pod krzywą (AUC). Jednakże, jeżeli różnice w wielkości AUC są niewielkie, pojawia się problem z wyborem optymalnego modelu klasyfikacyjnego. Mierniki NRI oraz IDI można analizować osobno dla każdej grupy, co pozwala uzyskać dodatkowe informacje o zdolności predykcyjnej badanej reguły decyzyjnej. Może to być przydatne zwłaszcza w sytuacji, kiedy stosowane powszechnie miary jakości klasyfikacji, jak np. dokładność, nie dają jednoznacznych wskazówek pozwalających na wybór najlepszego klasyfikatora.

Omówione metody cechuje łatwość obliczania – można w prosty sposób wykorzystać dowolny arkusz kalkulacyjny bądź skorzystać z pakietów środowiska R. Wydaje się zasadne popularyzowanie tych metod w naukach innych niż medyczne, np. w ekonomii, finansach czy bankowości.

Literatura

Cook N.R. (2007), *Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction*, „Circulation", 115, s. 928-935.

Cook N.R. (2008), *Statistical Evaluation of Prognostic versus Diagnostic Models: Beyond the ROC Curve*, „Clinical Chemistry", 54, 1, s. 17-23.

Fawcett T. (2006), *An introduction to ROC analysis*, „Pattern Recognition Letters", 27, s. 861-874.

Fielding A.H. (2007), *Cluster and Classification Techniques for the Biosciences*, Cambridge University Press, Cambridge.

Krzanowski W.J., Hand D.J. (2009), *ROC Curves for Continuous Data*, CRC Press, Boca Raton – London – New York.

Kundu S., Aulchenko Y.S., Janssens A.C.J.W. (2011), *Predict ABEL: An R package for the assessment of risk prediction models*, „European Journal of Epidemiology", 26, s. 261-264.

Pencina M.J., D'Agostino R.B. Sr, D'Agostino R.B. Jr, Vasan R.S. (2008), *Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond*, „Statistics in Medicine", 27, s. 157-172.

Pencina M.J., D'Agostino R.B. Sr, Steyerberg E.W. (2011), *Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers*, „Statistics in Medicine", 30, s. 11-21.

Provost F., Domingos P. (2001), *Well-trained PETs: Improving Probability Estimation Trees*, CeDER Working Paper #IS-00-04, Stern School of Business, New York University, New York.
Youden W.J. (1950), *Index for Rating Diagnostic Tests*, „Cancer", 3, s. 32-35.

SELECTED METHODS FOR ASSESSING THE PERFORMANCE OF CLASSIFIERS – AN OVERVIEW AND EXAMPLES OF APPLICATIONS

Summary: The ROC curve and the area under the ROC curve (AUC) are popular measures for evaluating the performance of classification models for binary outcomes. Recently, several new measures have been proposed to assess the predictive ability of classifiers. These include, among others, reclassification tables [Cook 2008], net reclassification improvement (NRI) and integrated discrimination improvement (IDI) [Pencina et al. 2008]. This paper briefly describes the methods listed above and presents some examples of their application possibilities.

Keywords: classifier performance, ROC curve, reclassification.