

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 naukowci

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

Kopiowanie i powielanie w jakiegokolwiek formie
wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu
Wrocław 2014

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

ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Wstęp	11
Małgorzata Rószkiewicz , Wykorzystanie metaanalizy w budowaniu modelu pomiarowego w przypadku braku niezmienniczości zasad pomiaru na przykładzie pomiaru zadowolenia z życia.....	13
Elżbieta Sobczak , Harmonijność inteligentnego rozwoju regionów Unii Europejskiej	21
Ewa Roszkowska, Renata Karwowska , Analiza porównawcza województw Polski ze względu na poziom zrównoważonego rozwoju w roku 2010.....	30
Tadeusz Kufel, Magdalena Osińska, Marcin Błażejowski, Paweł Kufel , Analiza porównawcza wybranych filtrów w analizie synchronizacji cyklu koniunkturalnego.....	41
Marcin Salamaga , Próba konstrukcji tablic „wymierania scenicznego” spektakli operowych na przykładzie Metropolitan Opera.....	51
Iwona Foryś , Wykorzystanie analizy dyskryminacyjnej do typowania rynków podobnych w procesie wyceny nieruchomości niemieszkalnych	59
Jerzy Korzeniewski , Selekcja zmiennych w klasyfikacji – propozycja algorytmu	69
Sabina Denkowska , Testowanie wielokrotne przy weryfikacji wieloczynnikowych modeli proporcjonalnego hazardu Coxa.....	76
Ewa Chodakowska , Teoria równań strukturalnych w klasyfikacji zmiennych jawnych i ukrytych według charakteru ich wzajemnych oddziaływań	85
Iwona Konarzewska , Model PCA dla rynku akcji – studium przypadku	94
Katarzyna Wójcik, Janusz Tuchowski , Dobór optymalnego zestawu słów istotnych w opiniach konsumentów na potrzeby ich automatycznej analizy	106
Aleksandra Łuczak , Zastosowanie metody AHP-LP do oceny ważności determinant rozwoju społeczno-gospodarczego w jednostkach administracyjnych	116
Aleksandra Witkowska, Marek Witkowski , Klasyfikacja pozycyjna banków spółdzielczych według stanu ich kondycji finansowej w ujęciu dynamicznym	126
Adam Depta , Zastosowanie analizy korespondencji do oceny jakości życia ludności na podstawie kwestionariusza SF-36v2	135
Marek Lubicz, Maciej Zięba, Konrad Pawelczyk, Adam Rzechonek, Marek Marciniak, Jerzy Kołodziej , Indukcja reguł dla danych niekompletnych i niezbalansowanych: modele klasyfikatorów i próba ich zastosowania do predykcji ryzyka operacyjnego w torakochirurgii	146

Małgorzata Misztal , Wybrane metody oceny jakości klasyfikatorów – przegląd i przykłady zastosowań.....	156
Anna M. Olszewska , Wykorzystanie wybranych metod taksonomicznych do oceny potencjału innowacyjnego województw	167
Iwona Bąk , Porównanie jakości grupowań powiatów województwa zachodniopomorskiego pod względem atrakcyjności turystycznej.....	177
Agnieszka Kozera, Joanna Stanisławska, Romana Głowicka-Wołoszyn , Segmentacja gospodarstw domowych według wydatków na turystykę zorganizowaną.....	186
Agnieszka Wałęga , Podejście syntetyczne w analizie spójności ekonomicznej gospodarstw domowych.....	196
Joanna Banaś, Małgorzata Machowska-Szewczyk, Bożena Mroczek , Zastosowanie analizy korespondencji do badania wpływu elektrowni wiatrowych na jakość życia ludności	205
Joanna Banaś, Krzysztof Małecki , Klasyfikacja punktów pomiarów ankietowych kierowców na granicy Szczecina z wykorzystaniem zmiennych symbolicznych.....	214
Aneta Becker , Wykorzystanie informacji granularnej w analizie wymagań rynku pracy.....	222
Katarzyna Cheba, Joanna Holub-Iwan , Wykorzystanie analizy korespondencji w segmentacji rynku usług medycznych.....	230
Adam Depta, Iwona Staniec , Identyfikacja czynników decydujących o jakości życia studentów łódzkich uczelni.....	238
Katarzyna Dębowska, Jarosław Kilon , Reguły asocjacyjne w analizie wyników badań metodą Delphi.....	247
Anna Domagała , O wykorzystaniu analizy głównych składowych w metodzie <i>Data Envelopment Analysis</i>	254
Alicja Grześkowiak , Analiza wykluczenia cyfrowego w Polsce w ujęciu indywidualnym i regionalnym.....	264
Anna M. Olszewska, Anna Gryko-Nikitin , Pomiar postrzegania jakości kształcenia uczelni wyższej na danych porządkowych z wykorzystaniem środowiska R.....	273
Karolina Paradysz , Hierarchiczna metoda grupowania powiatów jako podejście benchmarkowe w ocenie bezrobocia według BAEL-u w wybranych typach małych obszarów	282
Radosław Pietrzyk , Porównanie metod pomiaru efektywności zarządzania portfelami funduszy inwestycyjnych.....	290
Agnieszka Przedborska, Małgorzata Misztal , Wybrane metody statystyki wielowymiarowej w ocenie skuteczności terapeutycznej głębokiej stymulacji elektromagnetycznej u pacjentów z chorobą zwyrodnieniową stawów.....	299

Wojciech Roszka, Marcin Szymkowiak , Podejście kalibracyjne w statystycznej integracji danych	308
Iwona Skrodzka , Zastosowanie wybranych metod klasyfikacji do analizy kapitału ludzkiego krajów Unii Europejskiej	316
Agnieszka Stanimir , Wielowymiarowa analiza czynników sprzyjających włączeniu społecznemu	326
Dorota Strózik, Tomasz Strózik , Przestrzenne zróżnicowanie poziomu życia w województwie wielkopolskim.....	334
Izabela Szamrej-Baran , Identyfikacja przyczyn ubóstwa energetycznego w Polsce przy wykorzystaniu modelowania miękkiego.....	343
Janusz Tuchowski, Katarzyna Wójcik , Klasyfikacja obiektów w systemie Krajowych Ram Kwalifikacji opisanych za pomocą ontologii	353
Aleksandra Matuszewska-Janica , Grupowanie krajów Unii Europejskiej ze względu na poziom feminizacji sektorów gospodarczych	361
Monika Rozkrut, Dominik Rozkrut , Identyfikacja strategii innowacyjnych przedsiębiorstw usługowych w Polsce	369

Summaries

Małgorzata Rószkiewicz , The use of meta-analysis in building the measurement model in case of the absence of measurement invariance on the example of measuring of life satisfaction.....	20
Elżbieta Sobczak , Harmonious smart growth of European Union regions.....	29
Ewa Roszkowska, Renata Karwowska , The comparative analysis of Polish voivodeships with respect to sustainable development in 2010.....	40
Tadeusz Kufel, Magdalena Osińska, Marcin Błażejowski, Paweł Kufel , Comparative analysis of chosen filters in business cycles analysis	50
Marcin Salamaga , The attempt of construction of the life tables for opera works on the example of the Metropolitan Opera	58
Iwona Foryś , Using discriminant analysis to select similar markets in non-residential property valuation process.....	68
Jerzy Korzeniewski , Variable selection in classification – algorithm proposal	75
Sabina Denkowska , Multiple testing in the verification process of multifactorial Cox proportional hazards models	84
Ewa Chodakowska , The theory of structural equations modelling in the classification of observed variables and latent constructs according to the character of their relationship.....	93
Iwona Konarzewska , Modelling stock market by PCA factor model – case study	105

Katarzyna Wójcik, Janusz Tuchowski , Selection of the optimal set of relevant words in consumers opinions in the context of the opinion mining ..	115
Aleksandra Łuczak , Application of AHP-LP to the evaluation of importance of determinants of socio-economic development in the administrative units	125
Aleksandra Witkowska, Marek Witkowski , A dynamic approach to the ranking of cooperative banks by their financial condition	134
Adam Depta , Application of correspondence analysis for the measurement of quality of life – questionnaire SF-36v2 based research	145
Marek Lubicz, Maciej Zięba, Konrad Pawelczyk, Adam Rzechonek, Marek Marciniak, Jerzy Kołodziej , Classification rules extraction for missing and imbalance data: models of classifiers and initial results in the rules-based thoracic surgery risk prediction.....	155
Małgorzata Misztal , Selected methods for assessing the performance of classifiers – an overview and examples of applications.....	166
Anna M. Olszewska , The application of selected quantitative methods to the evaluation of voivodeship innovation level potential.....	176
Iwona Bąk , The comparison of the quality of groupings of poviats of West Pomeranian Voivodeship in terms of tourism attractiveness	185
Agnieszka Kozera, Joanna Stanisławska, Romana Głowicka-Wołoszyn , Household segmentation with respect to the expenditure on organized tourism.....	195
Agnieszka Wałęga , Synthetic approach in the analysis of economic coherence of households	204
Joanna Banaś, Małgorzata Machowska-Szewczyk, Bożena Mroczek , Using the correspondence analysis to examine the impact of wind turbines on the quality of life.....	213
Joanna Banaś, Krzysztof Małecki , Classification of measurement survey points of drivers on the boundary of Szczecin using symbolic variables...	221
Aneta Becker , The use granular information in the analysis of the requirements of the labor market.....	229
Katarzyna Cheba, Joanna Hołub-Iwan , The application of the correspondence analysis of patients segmentation on the medical service market	237
Adam Depta, Iwona Staniec , Identification of the factors that determine the quality of students life at universities in Lodz.....	246
Katarzyna Dębkowska, Jarosław Kilon , Association rules in the analysis of research results the Delphi method	253
Anna Domagała , About using Principal Component Analysis in Data Envelopment Analysis	263
Alicja Grześkowiak , Analysis of the digital divide in Poland at the individual and regional level	272

Anna M. Olszewska, Anna Gryko-Nikitin , Assessment of perception of quality of teaching at an institution of higher learning based on the ordinal data with the utilization of R environment.....	281
Karolina Paradysz , The hierarchical method of grouping poviats as a benchmark approach in the assessment of unemployment by BAEL in selected types of small areas	289
Radosław Pietrzyk , Comparison of methods of measuring the performance of investment funds portfolios.....	298
Agnieszka Przedborska, Małgorzata Misztal , Selected multivariate statistical analysis methods in the evaluation of efficacy of deep electromagnetic stimulation in patients with degenerative joint disease	307
Wojciech Roszka, Marcin Szymkowiak , A calibration approach in statistical data integration	315
Iwona Skrodzka , Application of some methods of classification to the analysis of human capital in the European Union.....	325
Agnieszka Stanimir , Multivariate analysis of social inclusion factors.....	333
Dorota Strózik, Tomasz Strózik , Spatial differentiation of the standard of living in Great Poland Voivodeship	342
Izabela Szamrej-Baran , Identification of fuel poverty causes in Poland using soft modelling	352
Janusz Tuchowski, Katarzyna Wójcik , Classification of objects in the National Classification Framework described by the ontology.....	360
Aleksandra Matuszewska-Janica , Clustering of European Union states taking into consideration the levels of feminization of economic sectors..	368
Monika Rozkrut, Dominik Rozkrut , Identification of service sector innovation strategies in Poland.....	379

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 niezagrażonej bankrutstwem. 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*):

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

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

Tabela 1. Schemat macierzy klasyfikacji

Stan przewidywany	Stan obserwowany		Σ
	klasa wyróżniona (P)	klasa niewyróżniona (N)	
Klasa wyróżniona (P)	TP (<i>true positives</i>) objekty z klasy wyróżnionej poprawnie zaklasyfikowane	FP (<i>false positives</i>) objekty z klasy niewyróżnionej błędnie zaklasyfikowane	TP+FP
Klasa niewyróżniona (N)	FN (<i>false negatives</i>) objekty z klasy wyróżnionej błędnie zaklasyfikowane	TN (<i>true negatives</i>) objekty 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 nie zrównoważonych. 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
<i>True Positives Rate/Sensitivity</i> (czułość)/ <i>Recall</i> (pamięć)	$TPR = \frac{TP}{TP + FN}$
<i>False Positives Rate</i>	$FPR = \frac{FP}{FP + TN}$
<i>Specificity</i> (swoistość)	$Specificity = \frac{TN}{FP + TN} = 1 - FPR$
<i>Positive Predictive Value</i> (dodatnia zdolność predykcyjna)/ <i>Precision</i> (precyzja)	$PPV = \frac{TP}{TP + FP}$
<i>Negative Predictive Value</i> (ujemna zdolność predykcyjna)	$NPV = \frac{TN}{TN + FN}$
<i>F1-measure</i> (miara F1)	$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$

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

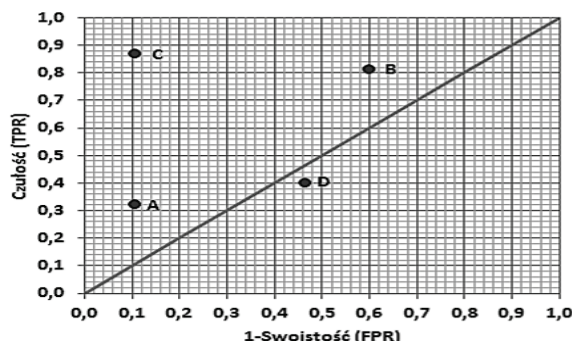
ty). Można również rozszerzyć analizę jakości klasyfikatora poprzez podanie innych niż (1) i (2) miar jakości. Przykładowe miary obliczane w oparciu o wielkości przedstawione w macierzy klasyfikacji (tab. 1) przedstawia tab. 2.

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 ROC¹ 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 dyskretnie 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).

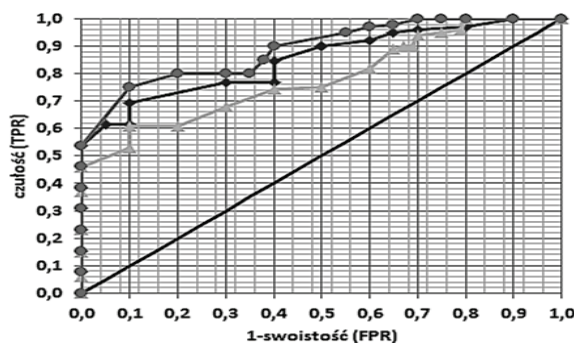


Rys. 1. Porównanie czterech przykładowych klasyfikatorów dyskretnych w przestrzeni ROC

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: $Y = \text{TPR} - \text{FPR}$ [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 Improvement*) 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 sposób²:

$$\widehat{NRI} = \widehat{NRI}_P + \widehat{NRI}_N = \left(\widehat{P}(\uparrow/P) - \widehat{P}(\downarrow/P) \right) + \left(\widehat{P}(\downarrow/N) - \widehat{P}(\uparrow/N) \right), \quad (3)$$

gdzie:

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

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

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

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

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:

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

gdzie $\text{mean}(\hat{p}_{old,P})$ i $\text{mean}(\hat{p}_{old,N})$ oznaczają średnie z oszacowanych prawdopodobieństw dla grupy wyróżnionej i niewyróżnionej dla modelu początkowego, natomiast $\text{mean}(\hat{p}_{new,P})$ i $\text{mean}(\hat{p}_{new,N})$ – 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 modeli³.

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

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

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 kredytobiorców⁴ (kredyty spłacane (klasa niewyróżniona)/ kredyty windygowane (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).

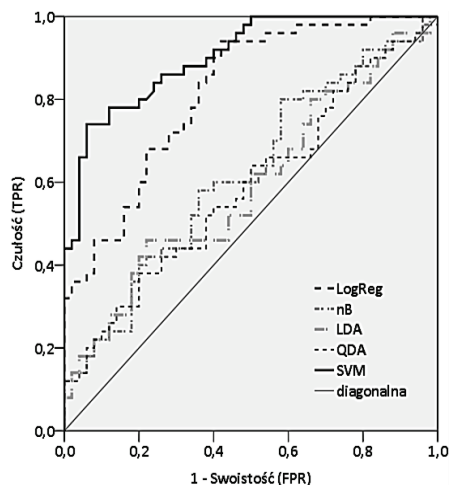
Na podstawie próby uczącej zbudowano 5 modeli klasyfikacyjnych wykorzystując regresję logistyczną (LogReg), naiwny klasyfikator Bayesa (nB), liniową i kwadratową funkcję dyskryminacyjną (LDA, QDA) oraz metodę wektorów nośnych (SVM)⁵. Zdolność predykcyjną modeli oceniano z wykorzystaniem próby testowej.

W tabeli 3 przedstawiono wartości wybranych miar jakości klasyfikatorów⁶. 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{NRI} dla kredytobiorców spłacających wynosi 28%, a dla windykowanych 20%). Dodatkowo wartości \widehat{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{NRI} dla grupy kredytów windykowanych	0,10	0,22	0,30	0,20
\widehat{NRI} dla grupy kredytów spłacanych	0,08	0,20	0,16	0,28
\widehat{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{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
$\widehat{Relative IDI}$	0,8625	2,0748	3,3461	2,9204

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

Wartości miernika \widehat{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.