PRZEGLĄD STATYSTYCZNY
R. LXIII – ZESZYT 4 – 2016

SERGIUSZ HERMAN[1]

ANALIZA PORÓWNAWCZA WYBRANYCH METOD SZACOWANIA BŁĘDU PREDYKCJI KLASYFIKATORA

1. WSTĘP

Proces konstrukcji klasyfikatora obejmuje trzy zasadnicze etapy: wybór cech opisujących klasyfikowane obiekty (zmiennych predykcyjnych), wybór metody/modelu oraz finalną ocenę jego jakości. Jakość klasyfikatora utożsamiana jest z jego umiejętnością przewidywania (prognozowania) przynależności do rozważanych populacji obiektów, dla których przynależność ta nie jest znana. Miarą tak zdefiniowanej jakości może być błąd predykcji klasyfikatora. Badania poświęcone klasyfikacji obiektów koncentrują swoją uwagę na wykorzystywaniu coraz bardziej zaawansowanych metod klasyfikacji, przyjmując powszechnie stosowane metody szacowania błędu predykcji. Jednocześnie podkreśla się, iż niezależnie od stopnia zaawansowania metody klasyfikacji, jakość decyzji podjętych na podstawie skonstruowanego za jej pomocą modelu uzależniona jest od tego, jak wiarygodnie zostanie oszacowana jego zdolność predykcyjna (Isaksson i inni, 2008).

W literaturze wskazuje się, iż ocena zdolności predykcyjnej klasyfikatora powinna być dokonana na podstawie dużej, niezależnej próby obiektów, które nie zostały uwzględnione przy konstrukcji modelu. Często jednak, szczególnie w przypadku badań dotyczących polskiego rynku kapitałowego, trudno taką próbę uzyskać. Rozwiązaniem tej sytuacji jest wówczas wykorzystanie jednej z wielu, zaproponowanych w literaturze, metod szacowania błędu predykcji klasyfikatora. W zagranicznych publikacjach można spotkać opracowania, których celem jest ich empiryczna analiza porównawcza. I tak na przykład Wehberg, Schumacher (2004) wykorzystali w tym celu metodę resubstytucji, metodę walidacji krzyżowej oraz metody wielokrotnego repróbkowania. Molinaro i inni (2005) rozszerzyli wspomnianą listę porównywanych metod o prostą metodę podziału, Braga-Neto, Dougherty (2004) oraz Kim (2009) uwzględnili z kolei w swoich badaniach powtarzaną walidację krzyżową oraz powtarzaną prostą metodę podziału. Analizy te zostały przeprowadzone na podstawie zmiennych opisujących wyniki eksperymentów genetycznych oraz danych uzyskanych w wyniku symulacji. W polskiej literaturze brakuje podobnego opracowania o charakterze empirycznym,

[1] Uniwersytet Ekonomiczny w Poznaniu, Wydział Informatyki i Gospodarki Elektronicznej, Katedra Ekonometrii, al. Niepodległości 10, 61-875 Poznań, Polska, e-mail: sergiusz.herman@ue.poznan.pl.

poświęconego metodom szacowania błędu predykcji. Przegląd metod szacowania błędu predykcji klasyfikatora można znaleźć np. w pracach Gatnara (2001, 2008).

Celem artykułu jest dokonanie przeglądu oraz empirycznej analizy porównawczej wybranych metod szacowania błędu predykcji klasyfikatora, skonstruowanego z wykorzystaniem liniowej analizy dyskryminacyjnej. W analizie porównawczej estymatorów wykorzystano trzy miary tj. odchylenie standardowe, obciążenie oraz błąd średniokwadratowy. Zbadano, czy wyniki analizy uzależnione są od wielkości próby oraz metody wyboru zmiennych do modelu. W tym celu badanie przeprowadzono dla prób o różnej liczebności oraz wykorzystano trzy statystyczne metody wyboru zmiennych do modelu.

Wyniki przeprowadzonej analizy pozwolą wskazać metodę szacowania błędu predykcji klasyfikatora, która posiada najbardziej pożądane własności, na przykładzie problemu prognozowania upadłości spółek akcyjnych w Polsce.

2. BŁĄD PREDYKCJI KLASYFIKATORA I WYBRANE METODY JEGO SZACOWANIA

Na podstawie notacji zaproponowanej m.in. przez Efrona, Tibishiraniego (1997) zagadnienie konstrukcji klasyfikatora oraz szacowania błędu predykcji można przedstawić w sposób formalny. Pierwszym krokiem badania jest zgromadzenie danych dla obiektów tworzących próbę uczącą x. Próba ta składa się z n obiektów, z których każdy opisany jest wektorem $x_i = (t_i, y_i)$, gdzie t_i to wektor zmiennych (cech) opisujących i-ty obiekt, natomiast y_i jest zmienną określającą przynależność tego obiektu do populacji. Wykorzystując tak zdefiniowany zbiór uczący x konstruowany jest klasyfikator (model) r_x, który w oparciu o wartości zawarte w wektorze cech t_i dla i-tego obiektu, pozwala określić przynależność tego obiektu do badanych populacji, oznaczanej przez $r_x(t_i)$. W przypadku, gdy problem dotyczy klasyfikacji dychotomicznej (zmienne y_i oraz $r_x(t_i)$ przyjmują wartość 0 lub 1) można w następujący sposób zdefiniować miarę rozbieżności między rzeczywistą a prognozowaną przynależnością i-tego obiektu:

$$Q[y_i, r_x(t_i)] = \begin{cases} 0 \; jeżeli \; r_x(t_i) = y_i, \\ 1 \; jeżeli \; r_x(t_i) \neq y_i. \end{cases} \tag{1}$$

Skonstruowany klasyfikator wykorzystywany jest do prognozowania przynależności obiektów, tworzących próbę testową. Oznaczając taki obiekt za pomocą $x_0 = (t_0, y_0)$ przedstawioną właśnie miarę rozbieżności (wzór 1) można dla niego zapisać w skrócie $Q[y_0, r_x(t_0)]$. Porównuje ona rzeczywistą przynależność badanego obiektu y_0, z przynależnością ustaloną za pomocą klasyfikatora r_x skonstruowanego w oparciu o zbiór uczący x, przy wykorzystaniu wektora cech t_0 opisującego klasyfikowany obiekt.

Przyjmując, iż obserwacje $x_i = (t_i, y_i)$ z próby uczącej są próbą losową z pewnego rozkładu F oraz że obiekt z próby testowej $x_0 = (t_0, y_0)$ także charakteryzuje się tym rozkładem, można zdefiniować tzw. prawdziwy błąd predykcji (ang. *true error rate*) klasyfikatora za pomocą następującej wartości oczekiwanej:

$$Err = E_{0F}Q[y_0, r_x(t_0)]. \tag{2}$$

Klasyfikator został skonstruowany w oparciu o daną próbę uczącą *x*, stąd też błąd *Err* jest także określany mianem warunkowego błędu predykcji (ang. *conditional error rate*). Oszacowanie prawdziwego błędu predykcji wymaga posiadania dużej, niezależnej próby testowej. W przypadku, gdy badacz nie dysponuje nią, jego zadaniem jest oszacowanie prawdziwego błędu predykcji w oparciu o posiadaną próbę uczącą *x* (Efron, Tibshirani, 1997).

Najprostszym sposobem oszacowania prawdziwego błędu predykcji jest wykorzystanie wszystkich dostępnych obserwacji, zarówno w celu zbudowania, jak i oceny konstruowanego klasyfikatora. Oszacowaniem warunkowego błędu predykcji jest wówczas tzw. estymator resubstytucji (ang. *resubstitution estimator*):

$$\overline{err} = \frac{1}{n}\sum_{i=1}^{n} Q[y_i, r_x(t_i)]. \tag{3}$$

Zgodnie z tym zapisem, klasyfikator zostaje skonstruowany w oparciu o cały zbiór uczący *x*. Następnie dla każdego obiektu *i* zostaje porównana jego rzeczywista przynależność y_i z tą, prognozowaną za pomocą klasyfikatora – $r_x(t_i)$. Na tej podstawie zostaje obliczona średnia przedstawiona we wzorze (3). Uzyskany za pomocą przedstawionej metody estymator błędu predykcji jest nadmiernie „optymistyczny", to znaczy nie doszacowuje wartości ryzyka warunkowego (McLachlan, 1992, s. 339–340).

Drugą, bardzo często stosowaną w badaniach metodą szacowania warunkowego błędu predykcji klasyfikatora, jest tzw. prosta metoda podziału (ang. *split sample/ holdout metod*). Polega ona na jednokrotnym podziale dostępnych danych, zgodnie z wcześniej ustaloną proporcją *p*, na próbę uczącą i testową. Klasyfikator jest konstruowany w oparciu o pierwszą z nich. Następnie jest on wykorzystywany do prognozowania przynależności obiektów z próby testowej. Błąd predykcji wyrażany jest jako udział błędnie zaklasyfikowanych obiektów z próby testowej $x_0 = (t_0, y_0)$ w ogólnej liczebności tej próby. Zaletą tej metody jest jej prostota i fakt, że nie pociąga za sobą obszernych obliczeń, ponieważ klasyfikator jest konstruowany tylko raz. Jej wadą natomiast jest fakt, iż każdy obiekt przyporządkowany zostaje tylko do jednej z analizowanych prób. Ma to istotny wpływ na uzyskane wyniki, szczególnie w przypadku występowania w zgromadzonych danych obserwacji odstających. Z prostej metody podziału należy korzystać tylko wtedy, gdy dysponuje się dostatecznie szerokim zbiorem danych, które pozwolą na wyodrębnienie odpowiednio licznych, niezależnych zbiorów: uczącego i testowego (Ripley, 1996, s. 67).

W wielu dziedzinach nauki badania empiryczne muszą być przeprowadzane na podstawie zbiorów danych o ograniczonej, niewielkiej liczbie obserwacji. W tej sytuacji przedstawiona prosta metoda podziału, ze względu na wspomniane wady, nie gwarantuje wiarygodnych wyników. Z tego też powodu opracowano szereg metod

szacowania błędu predykcji, z których każda bazuje na odpowiednim, wielokrotnym podziale dostępnego zbioru danych na próbę uczącą oraz testową.

Pierwszą z takich metod jest walidacja krzyżowa (Lachenbruch, Mickey, 1968; Geisser, 1975). W metodzie tej klasyfikator konstruowany jest *n*-krotnie. Za każdym razem z dostępnej próby usuwany jest jeden obiekt, który wykorzystywany jest jako jednoelementowy zbiór testowy, reszta obiektów służy natomiast do uczenia modelu. Tym samym formułę wykorzystywaną przy szacowaniu warunkowego błędu predykcji można zapisać następująco:

$$\widehat{Err}^{cv1} = \frac{1}{n}\sum_{i=1}^{n} Q\left[y_i, r_{x_{(i)}}(t_i)\right], \tag{4}$$

gdzie $x_{(i)}$ oznacza próbę uczącą po usunięciu z niej *i*-tego obiektu. Opisana metoda szacowania błędu predykcji określana jest mianem walidacji krzyżowej typu „pozostaw jedną poza" (ang. *leave-one-out cross-validation*).

Innym wariantem tej metody jest *k*-krotna walidacja krzyżowa. W tym przypadku dostępna próba zostaje podzielona na *k* części. Następnie *k*-krotnie klasyfikator jest konstruowany na podstawie $k - 1$ części, oraz testowany na tej, nieuwzględnionej w uczeniu. Oszacowaniem błędu predykcji jest średnia z uzyskanych w ten sposób *k* wyników pośrednich. Zaletą walidacji krzyżowej jest fakt, iż każda z obserwacji zostaje uwzględniona zarówno przy szacowaniu modelu, jak i przy jego testowaniu. Wadą metody jest większy (w porównaniu z metodami wcześniej opisanymi) koszt obliczeniowy.

Losowy podział próby na *k* części w przypadku walidacji krzyżowej powoduje, iż otrzymany za pomocą tej metody estymator charakteryzuje wysoka zmienność – określana w literaturze mianem wariancji wewnętrznej (ang. *internal variance*) (Braga-Neto, Dougherty, 2004). Jednym ze sposobów jej redukcji jest wielokrotne powtórzenie całego procesu (tj. dzielenia dostępnej próby, konstruowania klasyfikatora oraz szacowania błędu predykcji). Metoda ta określana jest mianem powtarzanej *k*-krotnej walidacji krzyżowej (ang. *repeated cross-validation*)

Kolejną metodą szacowania błędu predykcji jest powtarzana prosta metoda podziału (ang. *repeated hold-out*) (Kim, 2009). Polega ona na losowym, *k*-krotnym podziale badanej próby (składającej się z *n* obiektów) na dwa podzbiory: uczący i testowy zgodnie z ustaloną proporcją *p*. Dla każdego powtórzenia $n \cdot p$ obserwacji jest przyporządkowywanych do zbioru testowego, pozostałe $n \cdot (1 - p)$ obiektów traktowane jest jako zbiór testowy. Błąd predykcji szacowany jest jako średnia z wszystkich *k* iteracji – losowań. Przewagą tej metody nad przedstawioną wcześniej walidacją krzyżową jest fakt, iż w tym przypadku możliwe jest uzyskanie większej liczby, różnych w stosunku do siebie, podziałów pierwotnego zbioru obserwacji.

Ostatnią grupą metod szacowania błędu predykcji, które zostały uwzględnione w badaniu są metody wielokrotnego repróbkowania (ang. *bootsrapping*). Ta grupa metod bazuje na generowaniu *B* prób typu bootstrap x^{*1}, x^{*2}, x^{*3}, …, x^{*B} w taki

sposób, że każda z nich powstaje poprzez *n*-krotne losowanie proste ze zwracaniem obiektów z dostępnej próby *n* obiektów $\{x_1, x_2, \ldots, x_n\}$. Próby te wykorzystywane są następnie jako próby uczące. Obiekty niewylosowane w kolejnych próbach stanowią próbę testową. Miarę rozbieżności między rzeczywistą a prognozowaną przynależnością *i*-tego obiektu, dla uczącej próby *b*, można wyrazić w następujący sposób:

$$Q_i^b = Q\left[y_i, r_{x^{*b}}(t_i)\right]. \tag{5}$$

Niech N_i^b odpowiada liczbie przypadków, w których obiekt x_i został wylosowany w *b*-tej próbie bootstrap. Dla każdego obiektu *i* oraz każdej próby bootstrap *b* określona zostanie następująca zmienna dychotomiczna:

$$I_i^b = \begin{cases} 1 \; jeżeli \; N_i^b = 0, \\ 0 \; jeżeli \; N_i^b > 0. \end{cases} \tag{6}$$

Biorąc pod uwagę powyższe oznaczenia, estymator błędu predykcji typu „pozostaw jedną poza" (ang. *leave-one-out bootstrapp*) można przedstawić następująco (Efron, 1983):

$$\widehat{Err}^{(1)} = \frac{\sum_i \sum_b I_i^b \, Q_i^b}{\sum_i \sum_b I_i^b}. \tag{7}$$

Badania pokazały, iż otrzymany w ten sposób estymator warunkowego błędu predykcji przeszacowuje jego wartość. Dzieje się tak, ponieważ każda z wylosowanych podprób uczących zawiera w przybliżeniu tylko 0,632*n* unikatowych (niepowtarzających się) obserwacji. Z tego też powodu Efron (1983) zaproponował jego modyfikację. W celu uniknięcia problemu przeszacowania, estymator opisany wzorem (7) został skorygowany o przedstawiony wcześniej, niedoszacowany estymator resubstytucji (wzór 3) w następujący sposób:

$$\widehat{Err}^{(0,632)} = 0{,}368\overline{err} + 0{,}632\widehat{Err}^{(1)}. \tag{8}$$

Jednak badania empiryczne przeprowadzone przez Efrona, Tibshiraniego (1997) wskazały na znaczącą wadę tego estymatora. Okazało się, że posiada on tendencję do nadmiernego, ujemnego obciążenia, zwłaszcza w przypadku klasyfikatorów mających tendencję do nadmiernego przeuczenia – dopasowania do obiektów, na podstawie których jest on konstruowany. W takich sytuacjach estymator resubstytucji przyjmuje wartości bliskie zeru, a tym samym wartość błędu predykcji (wzór 8) jest zbyt optymistyczna. Efron oraz Tibshirani zaproponowali, by zwiększyć wagę przypisaną estymatorowi $\widehat{Err}^{(1)}$ wtedy, kiedy skala nadmiernego dopasowania (przeuczenia) mierzona za pomocą różnicy między wartościami $\widehat{Err}^{(1)}$ a $\overline{err}$ wzrasta. Przedstawiono estymator następującej postaci (Efron, Tibshirani, 1997):

$$\widehat{Err}^{(0,632+)} = (1 - \hat{w}) \cdot \overline{err} + \hat{w} \cdot \widehat{Err}^{(1)}, \tag{9}$$

gdzie:
$\overline{err}$ – estymator opisany wzorem (3)

$$\hat{w} = \frac{0{,}632}{1 - 0{,}368\hat{R}}, \tag{10}$$

$$\hat{R} = \frac{\widehat{Err}^{(1)} - \overline{err}}{\hat{\gamma} - \overline{err}}, \tag{11}$$

$$\hat{\gamma} = \sum_{i=1}^{n}\sum_{j=1}^{n} Q[y_i, r_x(t_j)]/n^2. \tag{12}$$

Występująca w liczniku wzoru (11) miara przeuczenia $\widehat{Err}^{(1)} - \overline{err}$ skalowana jest z wykorzystaniem poziomu tzw. błędu braku informacji $\hat{\gamma}$ (ang. *no-information error rate*), który pojawiłby się wtedy, gdyby t_i oraz y_i były niezależne (wzór 12). W przypadku dychotomicznego problemu klasyfikacji błąd ten można przedstawić za pomocą zależności:

$$\hat{\gamma} = \hat{p}_1(1 - \hat{q}_1) + \hat{q}_1(1 - \hat{p}_1), \tag{13}$$

gdzie:
$\hat{p}_1$ – oznacza zaobserwowaną częstość y_i przyjmujących wartość 1,
$\hat{q}_1$ – oznacza zaobserwowaną częstość $r_x(t_j)$ przyjmujących wartość 1.

Uzyskany w ten sposób względny poziom przeuczenia $\hat{R}$, przyjmuje wartości z przedziału od 0 (dla braku przeuczenia, gdy $\widehat{Err}^{(1)} = \overline{err}$), do wartości 1 (kiedy poziom przeuczenia odpowiada wartościowo poziomowi $\hat{\gamma} - \overline{err}$). Waga $\hat{w}$ we wzorze (9) może przyjmować wówczas wartości z przedziału 0,632 do 1, a tym samym wartość estymatora $\widehat{Err}^{(0,632+)}$ jest nie mniejsza od $\widehat{Err}^{(0,632)}$ oraz nie większa od $\widehat{Err}^{(1)}$.

Efron, Tibshirani (1997) podkreślają w swojej pracy, iż mogą się zdarzyć sytuacje, gdy $\hat{\gamma} \leq \overline{err}$ lub $\overline{err} < \hat{\gamma} \leq \widehat{Err}^{(1)}$. Zaproponowali, by wówczas zmodyfikować wcześniej przedstawione zależności (wzory 7 i 11) w następujący sposób:

$$\widehat{Err}^{(1)\prime} = \min(\widehat{Err}^{(1)}, \hat{\gamma}), \tag{14}$$

$$\hat{R}' = \begin{cases} \dfrac{\widehat{Err}^{(1)} - \overline{err}}{\hat{\gamma} - \overline{err}} & jeżeli\ \widehat{Err}^{(1)} > \overline{err}\ i\ \hat{\gamma} > \overline{err}, \\ 0 & w\ pozostałych\ przypadkach. \end{cases} \tag{15}$$

W badaniach empirycznych wykorzystano oba opisane estymatory prawdziwego błędu predykcji (wzory 8 i 9). Należy jednak zaznaczyć, iż istnieje szereg innych, opisanych w literaturze metod repróbkowania służących do szacowania prawdziwego błędu predykcji (np. Jiang, Simon, 2007). Estymatory $\widehat{Err}^{(0,632)}$ oraz $\widehat{Err}^{(0,632+)}$ zostały wybrane przez autora ze względu na fakt, iż są one najbardziej popularne w literaturze.

3. WYKORZYSTANA PRÓBA BADAWCZA

Analiza porównawcza metod szacowania błędu predykcji klasyfikatora wymagała zgromadzenia odpowiedniej próby badawczej, reprezentującej dwie rozłączne grupy obiektów. W tym celu wykorzystano dane finansowe przedsiębiorstw o złej oraz dobrej kondycji finansowej. Kryterium decydującym o zaklasyfikowaniu przedsiębiorstw do pierwszej grupy był fakt ogłoszenia przez odpowiedni sąd ich upadłości. W celu wyselekcjonowania próby wykorzystano informacje zawarte w Internetowym Monitorze Sądowym i Gospodarczym. Postanowiono ograniczyć selekcję przedsiębiorstw do spółek akcyjnych, reprezentujących trzy różne branże gospodarki. W ten sposób, biorąc także pod uwagę dostępność danych finansowych, wyselekcjonowano:

– 30 spółek akcyjnych z branży budownictwo (PKD 41.10-43.99z),
– 30 spółek akcyjnych z branży przetwórstwo przemysłowe (PKD 10.11-33.20z),
– 30 spółek akcyjnych z branży handel hurtowy i detaliczny (PKD 46.11-47.99z).

Do każdej z nich została dobrana spółka akcyjna o dobrej kondycji finansowej. Za kryteria dopasowania poszczególnych par przyjęto: sektor, działalność główną oraz wielkość aktywów. Dane finansowe spółek upadłych pochodziły z ich sprawozdań finansowych z roku, poprzedzającego ten w którym złożono pierwszy wniosek o ogłoszenie upadłości. Dotyczyły one lat 2000–2013. Sprawozdania finansowe dla spółek zdrowych pochodziły z tych samym okresów. Źródłem danych były bazy firm Notoria Serwis i Bisnode Dun & Bradstreet oraz Monitor Polski B.

Tabela 1.

Lista wskaźników finansowych wykorzystanych w badaniu

Symbol	Formuła wskaźnika
Wskaźniki rentowności	
ROA	zysk netto / średnia wartość aktywów
ROE	zysk netto / średnia wartość kapitałów własnych
ZB	zysk brutto / średnia wartość aktywów
ZS	zysk ze sprzedaży / średnia wartość aktywów
MZ	zysk brutto / przychody ze sprzedaży
MZ2	zysk netto / przychody ze sprzedaży
MZO	zysk operacyjny / przychody ze sprzedaży

Tabela 1. (cd.)

Symbol	Formuła wskaźnika
Wskaźniki płynności	
KP	kapitał pracujący / suma bilansowa
WBP	majątek obrotowy / zobowiązania krótkoterminowe
WSP	(majątek obrotowy-zapasy)/zobowiązania krótkoterminowe
WPP	(majątek obrotowy-zapasy-należności)/zobowiązania krótkoterminowe
Wskaźniki struktury kapitałowo-majątkowej	
ZO	zobowiązania ogółem / aktywa ogółem
ZD	zobowiązania długoterminowe / aktywa ogółem
KW	kapitał własny / aktywa ogółem
KWZ	kapitał własny / zobowiązania ogółem
Wskaźniki sprawności działania	
RN	średnia wartość należności / przychody ze sprzedaży netto*365
RZ	średnia wartość zapasów / przychody ze sprzedaży netto*365
RZob	średnia wartość zobowiązań / przychody ze sprzedaży*365
Rakt	średnia wartość aktywów/przychody ze sprzedaży*365

Źródło: opracowanie własne.

W badaniach empirycznych wykorzystano 19 wskaźników finansowych charakteryzujących rentowność, płynność, strukturę kapitałowo-majątkową oraz sprawność działania przedsiębiorstw (tabela 1).

Ich wyboru dokonano na podstawie przeglądu literatury – są to wskaźniki pojawiające się najczęściej w modelach prognozowania upadłości. W wyborze kierowano się także dostępnością danych w sprawozdaniach finansowych spółek. Wartości wskaźników finansowych zostały obliczone dla roku poprzedzającego ten, w którym złożono pierwszy wniosek o ogłoszenie upadłości przedsiębiorstwa – dla spółek, wobec których ogłoszono upadłość oraz dla tego samego roku dla odpowiadających im spółek zdrowych.

Wśród założeń przyjmowanych przy konstruowaniu funkcji dyskryminacyjnej są te dotyczące rozkładu normalnego oraz równości wariancji zmiennych opisujących obiekty w badanych grupach. W literaturze podkreśla się jednak, iż niespełnienie tych wymagań nie pogarsza istotnie wyników uzyskanych za pomocą omawianej metody (Hand, 1981, s. 27; Hadasik, 1998, s. 94). Z tego też powodu – pomimo niespełnienia wspomnianych założeń[2] – w badaniu wykorzystano liniową analizę dyskryminacyjną.

[2] Jedynie w przypadku trzech wskaźników (*KP*, *ZO* oraz *KW*) dla spółek zdrowych oraz jednego wskaźnika (*WBP*) dla spółek, wobec których ogłoszono upadłość, spełnione jest założenie o ich rozkładzie normalnym. Tylko w przypadku trzech wskaźników sprawności działania (*RN*, *RZ* oraz *Rakt*) nie

4. EMPIRYCZNA ANALIZA PORÓWNAWCZA WYBRANYCH METOD SZACOWANIA BŁĘDU PREDYKCJI KLASYFIKATORA

Celem badania empirycznego jest porównanie różnych metod szacowania prawdziwego błędu predykcji. Przeprowadzona analiza składa się z M iteracji. W każdym powtórzeniu m ($m = 1, 2, \ldots, M$) następuje, spośród przedstawionej grupy 180 spółek akcyjnych, losowanie stratyfikowanej[3] podpróby, składającej się z n obiektów. Każda z nich pełni w badaniu dwojaką rolę. Po pierwsze, na jej podstawie zostają oszacowane, zgodnie z przedstawionymi wcześniej metodami, estymatory prawdziwego błędu predykcji $\widehat{Err}$. Po drugie, traktując każdą z wylosowanych podprób jako próbę uczącą, a pozostałe $180 - n$ obiektów jako dużą, niezależną próbę testową, oszacowano wartość prawdziwego błędu predykcji Err (zgodnie z wzorem (2)). W celu porównania różnych metod szacowania wykorzystano trzy następujące miary: odchylenie standardowe (*SD*) zmiennej $\widehat{Err}$ oraz obciążenie (*Bias*) i błąd średniokwadratowy (*MSE*) zmiennej $\widehat{Err} - Err$ (wzory 16–18).

$$SD = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\widehat{Err}_{n,m} - \overline{\widehat{Err}_n}\right)^2}, \tag{16}$$

$$Bias = \frac{1}{M}\sum_{m=1}^{M}\left(\widehat{Err}_{n,m} - Err_{n,m}\right), \tag{17}$$

$$MSE = \frac{1}{M}\sum_{m=1}^{M}\left(\widehat{Err}_{n,m} - Err_{n,m}\right)^2. \tag{18}$$

W przypadku każdej metody szacowania prawdziwego błędu predykcji, proces uczenia poprzedzony był każdorazowo selekcją zmiennych do modelu. Inne podejście, polegające na tylko jednorazowym wyborze zmiennych do modelu, skutkowałoby obciążonym estymatorem prawdziwego błędu predykcji (Simon i inni, 2003; Jiang, Simon, 2007).

W celu przeprowadzenia badania empirycznego przyjęto następujące założenia:

– liczba iteracji M równa jest 1000,
– liczba obiektów losowanych do podprób w kolejnych iteracjach $n = 60$, $n = 90$, $n = 120$,

---

ma podstaw do odrzucenia hipotezy zerowej mówiącej o tym, iż populacje, z których pochodzą obiekty mają jednakową wariancję.

[3] Każda wylosowana podpróba charakteryzuje się taką samą proporcją spółek, które ogłosiły upadłość i spółek zdrowych, jak miało to miejsce w oryginalnej próbie 180 spółek, czyli 1:1.

– dla prostej metody podziału przyjęto proporcje p podziału próby na uczącą oraz testową $p = \frac{1}{5}$, $p = \frac{1}{3}$, $p = \frac{1}{2}$.
– dla k-krotnej walidacji krzyżowej przyjęto $k = n$, $k = 3$, $k = 5$.
– dla powtarzanej k-krotnej walidacji krzyżowej oraz powtarzanej prostej metody podziału przyjęto liczbę powtórzeń równą 10,
– dla metod wielokrotnego repróbkowania przyjęto liczebność losowanych prób typu bootstrap $B = 50$.

Listę badanych estymatorów błędu predykcji przedstawiono w tabeli 2.

Tabela 2.

Estymatory błędu predykcji porównywane w badaniu

Oznaczenie estymatora	Metoda szacowania
RS	metoda resubstytucji – próba testowa odpowiada próbie uczącej
Split 1/5	metoda prosta podziału (próba testowa – 1/5 całej próby)
Split 1/3	metoda prosta podziału (próba testowa – 1/3 całej próby)
Split 1/2	metoda prosta podziału (próba testowa – 1/2 całej próby)
CV3	walidacja krzyżowa – podział próby na 3 części
CV5	walidacja krzyżowa – podział próby na 5 części
CV3 r10	powtarzana walidacja krzyżowa – podział próby na 3 części, 10 powtórzeń
CV5 r10	powtarzana walidacja krzyżowa – podział próby na 5 części, 10 powtórzeń
LOOCV	walidacja krzyżowa typu „pozostaw jedną poza”
0,632	wielokrotne repróbkowanie, estymator 0,632, liczba losowanych prób typu bootstrap $B = 50$
0,632+	wielokrotne repróbkowanie, estymator 0,632+, liczba losowanych prób typu bootstrap $B = 50$
rSplit	powtarzana 50-krotna prosta metoda podziału (próba testowa – 1/5 całej próby)

Źródło: opracowanie własne.

Do konstrukcji modelu klasyfikacyjnego (klasyfikatora) wykorzystano, jak wspomniano wcześniej, liniową analizę dyskryminacyjną. W celu uniknięcia skorelowania zmiennych opisujących obiekty przyjęto, iż każdorazowo przed procesem uczenia usuwane są te zmienne, które są silnie skorelowane z pozostałymi (współczynnik korelacji Pearsona wyższy od 0,90)[4]. Dodatkowo w celu wyselekcjonowania tych wskaźników finansowych, które charakteryzują się najwyższą zdolnością dyskryminacyjną w badaniu wykorzystano 3 statystyczne metody wyboru zmiennych:

[4] W przypadku gdy dwie zmienne są silnie skorelowane z dalszej analizy usuwana jest ta z nich, dla której średnia z wartości bezwzględnych współczynników korelacji między tą zmienną z pozostałymi jest wyższa.

– wybór 4 zmiennych, których wartość bezwzględna statystyki *t*-studenta dla testu porównującego średnią wartość zmiennej dla badanych grup jest najwyższa (metoda oznaczona dalej jako *t4zmienne*),
– metoda krokowa w przód, przyjęto poziom istotności dla wartości statystyki F równy 0,1 (*krokowa*),
– dobór zmiennych silnie skorelowanych ze zmienną grupującą – współczynnik korelacji Pearsona jest statystycznie istotny przy poziomie $\alpha = 0{,}05$ (*korelacja*).

Całość obliczeń została wykonana z wykorzystaniem środowiska statystycznego R.

W tabelach 3–5 przedstawiono charakterystyki tj. odchylenie standardowe (*SD*), obciążenie (*Bias*) oraz błąd średniokwadratowy (*MSE*) estymatorów błędu predykcji, uzyskane dla wymienionych statystycznych metod wyboru zmiennych do modelu.

Analizując dane zawarte w tabelach można zauważyć, iż w przypadku estymatora resubstytucji analizowana zmienna $\widehat{Err} - Err$, bez względu na liczebność losowanych w kolejnych iteracjach podprób, zawsze charakteryzuje się ujemnym obciążeniem. W przypadku estymatorów uzyskanych prostą metodą podziału odchylenie standardowe zmiennej $\widehat{Err}$ jest zdecydowanie wyższe, od pozostałych porównywanych w badaniu. Zestawiając te wartości z błędami predykcji oszacowanymi za pomocą metod walidacji krzyżowej wyraźnie widać, iż te drugie charakteryzują się zarówno znacznie mniejszym odchyleniem standardowym, jak i mniejszymi wartościami błędów średniokwadratowych. Biorąc pod uwagę obciążenie analizowanej zmiennej, najlepsze własności (wśród estymatorów szacowanych metodą jednokrotnej walidacji krzyżowej) posiada estymator CV5.

Tabela 3.

Analiza porównawcza estymatorów błędu predykcji. Metoda wyboru zmiennych do modelu: *t4zmienne*

Estymator	$n = 60$			$n = 90$			$n = 120$		
	SD	Bias	MSE	SD	Bias	MSE	SD	Bias	MSE
RS	0,0600	-0,0317	0,0046	0,0549	-0,0232	0,0036	0,0581	-0,0189	0,0037
Split 1/5	0,1283	-0,0041	0,0165	0,1098	-0,0047	0,0121	0,0981	-0,0010	0,0096
Split 1/3	0,1036	-0,0073	0,0108	0,0868	-0,0048	0,0076	0,0852	-0,0066	0,0073
Split 1/2	0,0918	-0,0189	0,0088	0,0825	-0,0140	0,0070	0,0803	-0,0119	0,0066
CV3	0,0702	-0,0082	0,0050	0,0632	-0,0066	0,0040	0,0639	-0,0074	0,0041
CV5	0,0671	-0,0036	0,0045	0,0607	-0,0042	0,0037	0,0630	-0,0037	0,0040
CV3r10	0,0625	-0,0087	0,0040	0,0570	-0,0072	0,0033	0,0610	-0,0063	0,0038
CV5r10	0,0621	-0,0046	0,0039	0,0569	-0,0041	0,0033	0,0606	-0,0028	0,0037
LOOCV	0,0724	-0,0163	0,0055	0,0601	-0,0111	0,0037	0,0625	-0,0072	0,0040
0,632	0,0592	-0,0021	0,0035	0,0548	-0,0004	0,0030	0,0594	-0,0019	0,0035
0,632+	0,0598	-0,0003	0,0035	0,0551	-0,0011	0,0030	0,0591	-0,0008	0,0035
rSplit	0,0645	-0,0053	0,0042	0,0575	-0,0044	0,0033	0,0617	-0,0027	0,0038

Źródło: opracowanie własne.

Tabela 4.

Analiza porównawcza estymatorów błędu predykcji. Metoda wyboru zmiennych do modelu: *krokowa*

Estymator	$n = 60$			$n = 90$			$n = 120$		
	SD	Bias	MSE	SD	Bias	MSE	SD	Bias	MSE
RS	0,0615	-0,0494	0,0062	0,0573	-0,0399	0,0049	0,0605	-0,0376	0,0051
Split 1/5	0,1314	-0,0083	0,0173	0,1097	-0,0061	0,0121	0,1008	-0,0001	0,0102
Split 1/3	0,1074	-0,0181	0,0119	0,0888	-0,0077	0,0079	0,0831	-0,0042	0,0069
Split 1/2	0,0953	-0,0246	0,0097	0,0813	-0,0182	0,0069	0,0794	-0,0090	0,0064
CV3	0,0716	-0,0129	0,0053	0,0647	-0,0081	0,0043	0,0652	-0,0020	0,0043
CV5	0,0683	-0,0082	0,0047	0,0633	-0,0052	0,0040	0,0649	-0,0006	0,0042
CV3r10	0,0590	-0,0132	0,0036	0,0557	-0,0083	0,0032	0,0610	-0,0033	0,0037
CV5r10	0,0607	-0,0080	0,0038	0,0562	-0,0047	0,0032	0,0611	-0,0012	0,0037
LOOCV	0,0743	-0,0221	0,0060	0,0680	-0,0160	0,0049	0,0679	-0,0097	0,0047
0,632	0,0581	-0,0006	0,0034	0,0551	-0,0019	0,0030	0,0600	-0,0060	0,0036
0,632+	0,0587	-0,0047	0,0035	0,0554	-0,0013	0,0031	0,0603	-0,0008	0,0036
rSplit	0,0614	-0,0083	0,0038	0,0570	-0,0052	0,0033	0,0621	-0,0017	0,0039

Źródło: opracowanie własne.

Tabela 5.

Analiza porównawcza estymatorów błędu predykcji. Metoda wyboru zmiennych do modelu: *korelacja*

Estymator	$n = 60$			$n = 90$			$n = 120$		
	SD	Bias	MSE	SD	Bias	MSE	SD	Bias	MSE
RS	0,0631	-0,0730	0,0093	0,0558	-0,0532	0,0060	0,0609	-0,0431	0,0056
Split 1/5	0,1333	-0,0110	0,0179	0,1063	-0,0065	0,0114	0,1017	-0,0057	0,0104
Split 1/3	0,1065	-0,0099	0,0114	0,0884	-0,0108	0,0079	0,0850	-0,0084	0,0073
Split 1/2	0,1011	-0,0229	0,0108	0,0815	-0,0232	0,0072	0,0801	-0,0174	0,0067
CV3	0,0705	-0,0136	0,0052	0,0649	-0,0119	0,0044	0,0657	-0,0100	0,0044
CV5	0,0730	-0,0077	0,0054	0,0601	-0,0056	0,0036	0,0642	-0,0057	0,0042
CV3r10	0,0620	-0,0144	0,0041	0,0564	-0,0118	0,0033	0,0618	-0,0095	0,0039
CV5r10	0,0630	-0,0073	0,0040	0,0563	-0,0063	0,0032	0,0618	-0,0052	0,0038
LOOCV	0,0727	-0,0185	0,0056	0,0620	-0,0126	0,0040	0,0649	-0,0117	0,0044
0,632	0,0599	-0,0070	0,0036	0,0538	-0,0045	0,0029	0,0603	-0,0046	0,0037
0,632+	0,0611	-0,0021	0,0037	0,0543	-0,0005	0,0029	0,0606	-0,0015	0,0037
rSplit	0,0647	-0,0085	0,0043	0,0566	-0,0069	0,0033	0,0620	-0,0051	0,0039

Źródło: opracowanie własne.

Prezentując teoretyczne podstawy różnych metod szacowania błędu predykcji zwrócono uwagę na to, iż błędy predykcji estymowane metodami (jednokrotnej) walidacji krzyżowej „narażone są" na dodatkową zmienność, określaną mianem wewnętrznej. Przedstawionym sposobem jej redukcji było wielokrotne powtórzenie procesu szacowania błędu predykcji – powtarzana *k*-krotna walidacja krzyżowa. Weryfikując skuteczność takiego postępowania, należy porównać charakterystyki estymatorów CV3 oraz CV5 z ich odpowiednikami, uwzględniającymi 10-krotne powtarzanie procesu szacowania błędu: CV3r10 oraz CV5r10. Analizując wyniki zawarte w tabelach można stwierdzić, iż uwzględnienie powtórzeń w procesie szacowania błędu predykcji w każdym przypadku redukuje, zgodnie z oczekiwaniami, zmienność wyników uzyskanych z wykorzystaniem badanych estymatorów. Ma to swoje odzwierciedlenie w zmniejszonych wartościach błędu średniokwadratowego. Na tej podstawie można wyciągnąć wniosek, iż większy koszt obliczeniowy, związany z 10-krotnym powtarzaniem procesu szacowania błędu predykcji, pozwala osiągnąć korzystniejsze własności uzyskanych estymatorów.

Analiza danych zawartych w tabelach 3–5 pozwala stwierdzić, iż niezależnie od przyjętej metody wyboru zmiennych do modelu oraz wielkości wylosowanej podpróby, wartości odchylenia standardowego oraz błędu średniokwadratowego są najniższe w przypadku estymatorów uzyskanych metodami wielokrotnego repróbkowania (oznaczone jako 0,632 oraz 0,632+). Obciążenie analizowanej zmiennej, w przypadku estymatora 0,632 (wzór 8) jest zawsze najniższe, jednak bardzo często ujemne. Potwierdzono tym samym, o czym wspomniano w pierwszym podrozdziale, iż estymator wielokrotnego repróbkowania 0,632 (wzór 8) niedoszacowuje wartości błędu predykcji.

5. PODSUMOWANIE I WNIOSKI

Celem niniejszego artykułu były przegląd oraz empiryczna analiza porównawcza wybranych metod szacowania błędu predykcji klasyfikatora, skonstruowanego z wykorzystaniem liniowej analizy dyskryminacyjnej. Badanie empiryczne zostało przeprowadzone na przykładzie problemu prognozowania upadłości spółek akcyjnych w Polsce.

W przeprowadzonej analizie zbadano własności statystyczne 12 estymatorów prawdziwego błędu predykcji klasyfikatora (ang. *true error rate*). Podsumowując wyniki przeprowadzonego badania sformułować można następujące wnioski, dotyczące metod szacowania błędu predykcji klasyfikatora:

– Estymator resubstytucji ma znaczącą tendencję do niedoszacowywania wartości prawdziwego błędu predykcji. Uzyskane za jego pomocą wyniki są nadmiernie optymistyczne.
– Estymatory uzyskane za pomocą prostych metod podziału charakteryzują się najwyższą zmiennością spośród wszystkich poddanych analizie.
– Uwzględnienie powtórzeń w metodach walidacji krzyżowej poprawia jakość uzyskanych za ich pomocą estymatorów. Związane jest to jednak z wyższym kosztem obliczeniowym.

– Estymatory 0,632 szacowane metodą wielokrotnego repróbkowania charakteryzują się w większości przypadków ujemnym obciążeniem, co oznacza, że niedoszacowują one wartości prawdziwego błędu predykcji.
– Metoda wyboru zmiennych do modelu oraz liczebność losowanych podprób n nie wpływa na wnioski, jakie można sformułować na podstawie analizy porównawczej badanych estymatorów.

Podsumowując, uzyskane wyniki świadczą o tym, iż estymator prawdziwego błędu predykcji 0,632+ charakteryzuje się najbardziej pożądanymi własnościami w przypadku prognozowania upadłości spółek akcyjnych w Polsce.

Szacowanie błędu predykcji jest kluczowym etapem konstrukcji każdego klasyfikatora. Z tego też powodu, należy kontynuować przedstawiony kierunek badań. W badaniu wykorzystano wybrane estymatory punktowe. W ostatnich latach w literaturze podkreśla się, iż uzyskane w ten sposób wyniki, szczególnie w przypadku małolicznych prób badawczych, mogą być niewiarygodne. Z tego też powodu wskazuje się, iż lepszym, choć też nie idealnym, rozwiązaniem może być określanie przedziałów ufności dla szacowanych błędów predykcji klasyfikatora (Hanczar, Dougherty, 2013).

LITERATURA

Braga-Neto U. M., Dougherty E. R., (2004), Is Cross-validation for Small-sample Microarray Classification?, *Bioinformatics*, 20 (3), 374–380.

Efron B., (1983), Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation, *Journal of the American Statistical Association*, 78 (382), 316–331.

Efron B., Tibshirani R. J., (1997), Improvements on Cross-Validation: The .632+ Bootstrap Method, *Journal of the American Statistical Association*, 92 (438), 548–560.

Gatnar E., (2001), *Nieparametryczna metoda dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa.

Gatnar E., (2008), *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa.

Geisser S., (1975), The Predictive Sample Reuse Method With Applications, *Journal of the American Statistical Association*, 70, 320–328.

Hadasik D., (1998), *Upadłość przedsiębiorstw w Polsce i metody jej prognozowania*, Zeszyty naukowe – seria II, Prace habilitacyjne, Zeszyt 153, Akademia Ekonomiczna w Poznaniu, Poznań.

Hanczar B., Dougherty E. R., (2013), The Reliability of Estimated Confidence Intervals for Classification Error Rates When Only a Single Sample is Available, *Pattern Recognition*, 46, 1067–1077.

Hand D. J., (1981), *Discrimination and Classification*, John Wiley & Sons, Chichester.

Isaksson A., Wallman M., Goransson H., Gustafsson M. G., (2008), Cross-Validation and Bootstrapping are Unreliable in Small Sample Classification, *Pattern Recognition*, 29, 1960–1965.

Jiang W., Simon R., (2007), A Comparison of Bootstrap Methods and an Adjusted Bootstrap Approach for Estimating Prediction Error in Microarray Classification, *Statistics in Medicine*, 26, 5320–5334.

Kim J. H., (2009), Estimating Classification Error Rate: Repeated Cross-Validation, Repeated Hold-Out and Bootstrap, *Computational Statistics and Data Analysis*, 53, 3735–3745.

Lachenbruch P. A., Mickey M. R., (1968), Estimation of Error Rates in Discriminant Analysis, *Technometrics*, 10, 1–11.

McLachlan G. J., (1992), *Discriminant Analysis and Statistical Pattern Recognition*, John Wiley & Sons, Inc.

Molinaro A. M., Simon R., Pfeiffer R. M., (2005), Prediction Error Estimation: A Comparison of Resampling Methods, *Bioinformatics*, 21, 3301–3307.

Ripley B. D., (1996), *Pattern Recognition and Neural Networks*, Cambrige University Press.

Simon R., Radmacher M. D., Dobbin K., McShane L. M., (2003), Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification, *Journal of the National Cancer Institute*, 95 (1), 14–18.

Wehberg S., Schumacher M., (2004), A Comparison of Nonparametric Error Rate Estimation Methods in Classification Problems, *Biometrical Journal*, 46, 35–47.

ANALIZA PORÓWNAWCZA WYBRANYCH METOD SZACOWANIA BŁĘDU PREDYKCJI KLASYFIKATORA

Streszczenie

Klasyfikacją nazywamy algorytm postępowania, który przydziela badane obserwacje/obiekty, bazując na ich cechach do określonych populacji. W tym celu konstruowany jest odpowiedni model – klasyfikator. Miarą jego jakości jest przede wszystkim zdolność predykcyjna, mierzona m.in. za pomocą prawdziwego błędu predykcji. Wartość tego błędu, ze względu na brak odpowiednio dużej, niezależnej próby testowej, musi być często szacowana na podstawie dostępnej próby uczącej.

Celem artykułu jest dokonanie przeglądu oraz empirycznej analizy porównawczej wybranych metod szacowania błędu predykcji klasyfikatora, skonstruowanego z wykorzystaniem liniowej analizy dyskryminacyjnej. Zbadano, czy wyniki analizy uzależnione są od wielkości próby oraz metody wyboru zmiennych do modelu. Badanie empiryczne zostało przeprowadzone na przykładzie problemu prognozowania upadłości spółek akcyjnych w Polsce.

Słowa kluczowe: błąd predykcji, walidacja krzyżowa, prosta metoda podziału, wielokrotne reprób-kowanie, upadłość przedsiębiorstw, klasyfikacja

COMPARATIVE ANALYSIS OF SELECTED METHODS FOR ESTIMATING THE PREDICTION ERROR OF CLASSIFIER

Abstract

Classification is an algorithm, which assigns studied companies, taking into consideration their attributes, to specific population. An essential part of it is classifier. Its measure of quality is especially predictability, measured by true error rate. The value of this error, due to lack of sufficiently large and independent test set, must be estimated on the basis of available learning set.

The aim of this article is to make a review and compare selected methods for estimating the prediction error of classifier, constructed with linear discriminant analysis. It was examined if the results of the analysis depends on the sample size and the method of selecting variables for a model. Empirical research was made on example of problem of bankruptcy prediction of join-stock companies in Poland.

Keywords: prediction error, cross-validation, holdout method, bootstrapping, corporate bankruptcy, classification