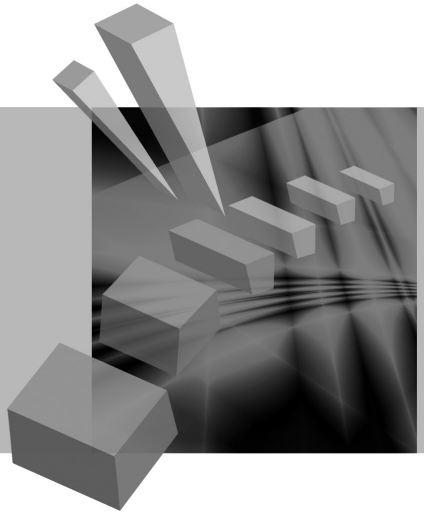
**PRACE NAUKOWE**
Uniwersytetu Ekonomicznego we Wrocławiu
**RESEARCH PAPERS**
of Wrocław University of Economics

**278**

# Taksonomia 20

## Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
**Krzysztof Jajuga**
**Marek Walesiak**

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2013

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
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 278 • 2013

Taksonomia 20. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

**Marek Lubicz, Maciej Zięba**
Politechnika Wrocławska

**Konrad Pawełczyk, Adam Rzechonek, Jerzy Kołodziej**
Akademia Medyczna we Wrocławiu

# MODELE EKSPLORACJI DANYCH NIEZBILANSOWANYCH – PROCEDURY KLASYFIKACJI DLA ZADANIA ANALIZY RYZYKA OPERACYJNEGO[1]

**Streszczenie:** W zadaniach klasyfikacji z wykorzystaniem danych rzeczywistych, na przykład w analizie danych medycznych, pojawiają się problemy konstrukcji klasyfikatorów, wynikające ze specyfiki analizowanych danych, związane m.in. z niezbilansowaniem zbiorów danych przy znacznej przewadze liczebności jednej bądź kilku klas. Celem pracy jest analiza porównawcza wybranych podejść do klasyfikacji danych niezbilansowanych. W badaniach zastosowano implementacje technik klasyfikacji w środowiskach uczenia maszynowego KEEL i WEKA. Jako dane do klasyfikacji wykorzystano zaktualizowaną bazę danych o pacjentach leczonych operacyjnie z powodu raka płuca we Wrocławskim Ośrodku Torakochirurgii w latach 2000-2011.

**Słowa kluczowe:** eksploracja danych, klasyfikacja, dane niezbilansowane, brakujące obserwacje, dane medyczne.

## 1. Wstęp

W licznych zastosowaniach analizy danych pojawiają się problemy konstrukcji klasyfikatorów, wynikające ze specyfiki analizowanych danych. Można tu wymienić przykładowo: wnioskowanie na podstawie niekompletnych danych przy braku znajomości wartości cech klasyfikowanych obiektów lub wartości etykiet klas dla niektórych obiektów ze zbioru uczącego, dane sekwencyjne lub dane ucięte, uczenie

[1] Praca naukowa finansowana ze środków budżetowych na naukę w latach 2010-2013 jako projekt badawczy N N115 090939 pt. *Modele i decyzje w systemach zdrowotnych. Koncepcje zastosowania metod badań operacyjnych i technologii informacyjnych do podejmowania decyzji zarządczych w systemach zdrowotnych* oraz współfinansowana ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

wrażliwe na koszt lub problem danych niezbilansowanych przy znacznej przewadze liczebności jednej lub kilku klas.

W problemach klasyfikacji wykorzystujących dane rzeczywiste powyższe niedoskonałości występują często łącznie, utrudniając dobór właściwego podejścia. Takim problemem jest rozważany przez autorów problem analizy ryzyka operacyjnego w torakochirurgii. W pracy [Lubicz i in. 2012] autorzy porównali efektywność kilku podejść do rozwiązania problemu klasyfikacji z brakującymi obserwacjami, uwzględniając różne klasyfikatory i metody uzupełniania braków danych. Celem tego artykułu jest analiza porównawcza wybranych podejść do klasyfikacji danych niezbilansowanych przy jednoczesnym występowaniu braków wartości obserwacji. Obliczenia, na przykładzie rzeczywistych zbiorów danych medycznych, przeprowadzono w środowisku uczenia maszynowego KEEL (www.keel.es) i porównawczo w środowisku WEKA (www.cs.waikato.ac.nz/ml/weka).

## 2. Klasyfikacja z uczeniem na podstawie próby niezbilansowanej

Klasyfikacja stanowi jedno w z podstawowych zadań wielu dyscyplin nauki, w tym statystyki, w której mówi się np. o procesie podziału zbioru obiektów na klasy lub wyznaczaniu reguł określania nieznanej wartości zmiennej zależnej na podstawie wartości atrybutów obiektu (dyskryminacja [Gatnar 2008]), jak również teorii uczenia maszynowego, w której określa się przekształcenie wektora cech opisującego dany obiekt do wartości reprezentującej jedną z możliwych klas obiektu. Funkcję przekształcającą wektor cech na zbiór klas nazywa się klasyfikatorem. Klasyfikator może być podany przez eksperta, np. w postaci zestawu reguł lub jako drzewo decyzyjne, jednak w większości przypadków jest on konstruowany w procesie uczenia nadzorowanego z wykorzystaniem zbioru uczącego. Kluczowym elementem w procesie budowy klasyfikatorów są dane zawarte w zbiorze uczącym. W odróżnieniu od danych dostępnych w repozytoriach uczenia maszynowego jakość danych rzeczywistych w większości przypadków nie pozwala na bezpośrednie wykorzystanie ich w procesie konstrukcji klasyfikatora. Jedną z istotnych przyczyn złej jakości danych mogą być dysproporcje w liczebności obiektów z poszczególnych klas w zbiorze uczącym (niezbilansowanie danych). W pracy [He, Garcia 2009] stwierdza się, że każde zadanie klasyfikacji, w którym występują różne częstości pojawiania się obiektów należących do różnych klas, należy traktować jako problem niezbilansowany. Ze względu na możliwość dekompozycji zadań wieloklasowych na zadania dwuklasowe problem dysproporcji w liczebnościach klas rozpatruje się zwykle dla dychotomicznych zagadnień decyzyjnych, w których definiuje się dwie klasy: klasę pozytywną (*positive*), będącą klasą zdominowaną (*minority*), oraz klasę negatywną (*negative*), reprezentującą klasę dominującą (*majority*). Istotą problemu jest to, iż zastosowanie klasycznych mechanizmów uczenia na niezrównoważonym zbiorze danych może prowadzić do faworyzowania przez wyuczony klasyfikator klasy dominującej kosztem klasy zdominowanej.

Problem uczenia dla danych niezbilansowanych można zdefiniować jako zadanie maksymalizacji: poprawności klasyfikacji ACC, współczynnika średniej geometrycznej jakości predykcji GM lub wskaźnika AUC, reprezentującego pole powierzchni pod krzywą ROC (*Receiver Operating Characteristic*):

$$ACC = (TP + TN) / NN, \tag{1}$$

$$GM = \sqrt{TPR \times TNR}\ , \tag{2}$$

$$AUC = 0,5 \times (1 + TPR - FPR)\ , \tag{3}$$

gdzie: *TPR, TNR, FPR* oznaczają odpowiednio: czułość klasyfikacji, swoistość klasyfikacji i odsetek błędów I rodzaju:

$$TPR = TP / (TP + FN), \tag{4}$$

$$TNR = TN / (FP + TN), \tag{5}$$

$$FPR = FP / (FP + TN), \tag{6}$$

*TP*, *TN*, *FP*, *FN* stanowią elementy macierzy kontyngencji i określają, w jaki sposób klasyfikowane były obiekty z poszczególnych klas (*T* – poprawna, *F* – niepoprawna klasyfikacja obiektu z klasy *P* – pozytywnej, *N* – negatywnej), a *NN* jest łączną liczbą obiektów.

Techniki stosowane do klasyfikacji w przypadku danych niezbilansowanych można podzielić [He, Garcia 2009; Galar i in. 2012] na trzy grupy: podejścia działające na poziomie danych (zewnętrzne), podejścia działające na poziomie algorytmu uczenia (wewnętrzne) oraz podejścia z uczeniem wrażliwym na koszt. W podejściach pierwszej grupy (jednym z najbardziej znanych jest algorytm SMOTE [Chawla i in. 2002]) obsługa danych niezrównoważonych odbywa się na poziomie przetwarzania danych (próbkowanie), niezależnie od stosowanego algorytmu uczenia klasyfikatora, a proces obsługi danych niezbilansowanych na etapie przetwarzania umożliwia stosowanie klasycznych algorytmów uczenia dedykowanych dla problemów zbilansowanych bez konieczności ich modyfikacji. Główną ideą podejść działających na poziomie algorytmu uczenia jest wzbogacenie klasycznych algorytmów uczenia o mechanizmy niwelujące negatywne skutki dysproporcji w danych. W pracy [Galar i in. 2012] wyróżnia się trzy podgrupy: podejścia wielomodelowe, łączące próbkowanie (pre-processing) z algorytmem boostingu (np. SMOTEBoost, RUSBoost) lub z algorytmem baggingu (np. OverBagging, UnderBagging) oraz podejścia hybrydowe (np. EasyEnsemble, BalanceCascade). Podejścia z uczeniem wrażliwym na koszt (przykładowo adaptacje algorytmu AdaCost [Sun i in. 2007]) stanowią kombinacje podejścia zewnętrznego i wewnętrznego, w których dane wejściowe modyfikowane są poprzez nadanie różnych wag poszczególnym obiektom, uwzględniane następnie przez algorytmy uczenia. Omówiony podział technik bilan-

sowania danych nie jest podziałem rozłącznym, gdyż niektóre algorytmy zakładają jednoczesne wykorzystanie kilku technik.

Techniki wrażliwe na koszt są powszechnie stosowane również w przypadku klasyfikatorów opartych na metodzie wektorów nośnych (SVM). Klasyczne sformułowanie problemu uczenia sprowadza się do minimalizacji funkcji, w której uwzględnia się jeden parametr $C$ kosztów związanych z błędną klasyfikacją. W uczeniu wrażliwym na koszt modyfikuje się funkcję celu, nadając różne koszty obserwacjom z klasy dominującej i zdominowanej, a minimalizowana funkcja celu jest postaci:

$$Q(a) = 0{,}5\times \underline{\mathbf{a}}^{T}\underline{\mathbf{a}} + C_{+} \times \sum_{n+\in I+} \xi_{n+} + C_{-} \times \sum_{n-\in I-} \xi_{n-}, \tag{7}$$

gdzie $\underline{\mathbf{a}}$ jest wektorem parametrów hiperpłaszczyzny liniowego klasyfikatora SVM oddzielającego dwie klasy, $C_+$ i $C_-$ są kosztami związanymi z błędną klasyfikacją dla klasy pozytywnej i negatywnej, $\xi_{n+}, \xi_{n-}$ są tzw. zmiennymi swobodnymi, które przyjmują wartości dodatnie, gdy obserwacje znajdują się wewnątrz bądź po złej stronie marginesu wyznaczonego w procesie uczenia, $I_+$ i $I_-$ oznaczają zbiory indeksów obiektów z klasy pozytywnej i negatywnej, a $N_+$ i $N_-$ – liczności tych zbiorów. Biorąc pod uwagę sugestie literaturowe [Tang i in. 2009] i wyniki prac własnych nad rozszerzeniem koncepcji opisanych w [Zięba 2011], w niniejszej pracy proponuje się zastosowanie metody C-SVM z wykorzystaniem techniki SMO uczenia klasyfikatora oraz niesymetrycznych kosztów zdefiniowanych jako:

$$C_{+} = C \times N/2N_{+},\ C_{-} = C \times N/2N_{-}. \tag{8}$$

## 3. Dane źródłowe analizy ryzyka operacyjnego i założenia analizy porównawczej

Utylitarnym celem badań było rozwinięcie wyników otrzymanych przez autorów dla problemu klasyfikacji dychotomicznej z brakującymi obserwacjami [Lubicz i in. 2012] na przypadek danych niezbilansowanych [Galar i in. 2012] występujący w rozważanej przez autorów problematyce analizy ryzyka operacyjnego. Klasami zdominowanymi, o liczebności istotnie mniejszej niż liczebność klasy dominującej, są tutaj – w zależności od przyjętej zmiennej objaśnianej – zbiór przypadków zgonu pacjenta w ciągu 30 dni lub 1 roku po operacji (ryzyko krótkookresowe) lub zbiór przypadków przeżycia przez pacjenta co najmniej $n$ ($n > 1$) lat po operacji (ryzyko długookresowe). Jednocześnie, oprócz nieproporcjonalnej liczebności klas, występują liczne braki wartości niektórych atrybutów. Jak wykazano w poprzednich badaniach, kwestię braków danych można próbować rozwiązać zastosowaniem różnych metod przetwarzania, na przykład imputacji, i właściwym doborem klasyfikatora. Pozwala to na zwiększenie dokładności klasyfikacji dla ryzyka krótkookresowego (75-95%), jednak inne wskaźniki jakości klasyfikacji pozostają nieakceptowalne

(duże różnice między TPR a TNR), czego przyczyną może być niezrównoważenie liczebności klas. Zamierzeniem autorów było zatem porównanie opisanych w literaturze przedmiotu metod klasyfikacji danych niezbilansowanych oraz próba zaproponowania własnego podejścia dla problemu analizy danych medycznych z zakresu torakochirurgii, rzadko podejmowanego w pracach o ukierunkowaniu klinicznym (nielicznymi przykładami są prace [Ferguson i in. 2008; Rivo i in. 2012]). Sformułowano pytania badawcze z perspektywy medycznej, dotyczące: modelowania ryzyka operacyjnego: przeżycia 30 dni lub $n$ ($n$ = 1, 2, ..., 5) lat po operacji i wpływu zakresu uwzględnianych danych klinicznych (dane przedoperacyjne PR, związane z zabiegiem OP, histopatologiczne HP i dane pooperacyjne PT) na jakość klasyfikacji. Zdefiniowano binarne zmienne objaśniane: R30, R1Y, ..., R5Y, odpowiadające długości przeżycia pooperacyjnego i opisanym wyżej interpretacjom klas zdominowanych (zgony pooperacyjne lub przeżycia pooperacyjne o określonej długości).

W badaniach wykorzystano dane o pacjentach leczonych operacyjnie z powodu raka płuca we Wrocławskim Ośrodku Torakochirurgii (WTO) w latach 2000-2011. Pierwszy etap badań obejmował aktualizację i weryfikację baz danych źródłowych z systemu szpitalnego WTO i spoza WTO (Regionalny Rejestr Nowotworów, Narodowy Fundusz Zdrowia). Efektem prac było utworzenie nowych badawczych baz danych o znacznie zmniejszonej, w porównaniu z poprzednimi badaniami, liczbie braków danych:

– W1: szczegółowe dane o pacjentach z resekcjami płuc z powodu pierwotnego raka płuca (2007-2011, 1203 obiekty, 139 zmiennych objaśniających, z tego 36 z okresu przedoperacyjnego, 57 związanych z zabiegiem operacyjnym, 17 dotyczących badania histopatologicznego i 29 pooperacyjnych; średnio 2,6% braków danych),
– W2: ograniczony zestaw podstawowych danych o wszystkich pacjentach, którym wykonano resekcje płuc w latach 2000-2011 (5595 obiektów, 15 zmiennych objaśniających, z tego – odpowiednio – 5, 2, 4, 4 z okresów przedoperacyjnego, operacyjnego, histopatologii i pooperacyjnego; średnio 8,3% braków danych).

W tabeli 1 porównano skalę niezbilansowania w przykładowych eksperymentalnych zbiorach danych dla cech przedoperacyjnych w zależności od zmiennej objaśnianej; wyliczono również wskaźnik niezbilansowania [Galar i in. 2012] jako iloraz liczebności klasy zdominowanej do liczebności klasy dominującej.

W drugim etapie wybrano techniki klasyfikacji danych niezbilansowanych. Biorąc pod uwagę postawione cele badawcze, zdecydowano o przeprowadzeniu eksperymentalnej analizy porównawczej w środowisku uczenia maszynowego KEEL [Alcalá-Fdez i in. 2011], w którym dostępne są 42 techniki klasyfikacji danych niezbilansowanych, reprezentujące wszystkie wymienione wcześniej kategorie technik. Spośród nich, bazując na sugestiach z pracy [Galar i in. 2012], wybrano 20 technik wymienionych w tab. 2, zaimplementowano także autorską wersję algorytmu C-SVM (wzory (3)-(4)) i włączono ją do środowiska KEEL. W trzecim etapie dla każdej bazy (2), każdej zmiennej objaśnianej (6) i zakresu klinicznego danych (4)

**Tabela 1.** Porównanie skali niezbilansowania w eksperymentalnych bazach danych

| Dane | Zmienna objaśniana | R30 | R1Y | R2Y | R3Y | R4Y | R5Y |
|---|---|---|---|---|---|---|---|
| Baza W1 | zbiór eksperymentalny | W1PRR0 | W1PRR1 | W1PRR2 | W1PRR3 | W1PRR4 | W1PRR5 |
| | liczebność | 1203 | 1203 | 1019 | 839 | 676 | 543 |
| | zgony pooperacyjne | 31 | 220 | 367 | 435 | 454 | 468 |
| | przeżycia pooperacyjne | 1172 | 983 | 652 | 404 | 222 | 75 |
| | wskaźnik niezbilansowania | 37,81 | 4,47 | 1,78 | 1,08 | 2,05 | 6,24 |
| Baza W2 | zbiór eksperymentalny | W2PRR0 | W2PRR1 | W2PRR2 | W2PRR3 | W2PRR4 | W2PRR5 |
| | liczebność | 5595 | 5595 | 5411 | 5231 | 4794 | 4403 |
| | zgony pooperacyjne | 140 | 988 | 1621 | 1986 | 2192 | 2327 |
| | przeżycia pooperacyjne | 5455 | 4607 | 3790 | 3245 | 2602 | 2076 |
| | wskaźnik niezbilansowania | 38,96 | 4,66 | 2,34 | 1,63 | 1,19 | 1,12 |

Źródło: obliczenia własne.

utworzono pliki eksperymentalne (48) i poddano je przetwarzaniu w module Imbalanced Learning w systemie KEEL wersja 2.0, wykorzystując standardowe implementacje 20 technik i własną implementację algorytmu C-SVM. Zbudowano 20-procentowe próby testowe, a jako metodę walidacji wybrano 5-częściowy sprawdzian krzyżowy. W ostatnim etapie wykonano dodatkowe badania porównawcze w środowisku WEKA przy zastosowaniu dostępnych technik (SMOTE, Resampling, a następnie zastosowanie algorytmów bazowych (SMO, NB, RF, CART), dla których w poprzednim etapie badań otrzymano najlepszą dokładność klasyfikacji). Ze względu na ograniczenia edytorskie poniżej omówiono tylko wybrane wyniki analizy porównawczej w środowisku KEEL.

## 4. Omówienie wyników badań i wnioski

Wybrane wyniki analizy porównawczej metod klasyfikacji dla danych niezbilansowanych dla przykładowych zbiorów danych eksperymentalnych przedstawiono w tab. 2-4. W tabeli 2 zamieszczono szczegółowe wyniki dla poszczególnych zastosowanych algorytmów klasyfikacji danych niezbilansowanych z pakietu KEEL dla ryzyka krótkookresowego (przeżycie roczne po operacji) i dwóch zakresów danych klinicznych: danych przedoperacyjnych oraz danych uzupełnionych o charakterystyki zabiegu operacyjnego, a w tab. 3-4 – dane uśrednione dla przeżycia rocznego i 5-letniego (ocena łączna). W tabelach zastosowano omówione wyżej oznaczenia wskaźników jakości klasyfikacji.

**Tabela 2.** Porównanie wyników klasyfikacji dla analizowanych metod przetwarzania danych niezbilansowanych i przeżycia rocznego (zbiory danych eksperymentalnych W1PRR1, W1OPR1 )

| Grupa metod | | Algorytm | Dane przedoperacyjne | | | | | Dane przed- i okołooperacyjne | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | ACC | TPR | TNR | GM | AUC | ACC | TPR | TNR | GM | AUC |
| Próbkowanie/ eliminacja + bagging | | Bagging | **0,81** | 0,08 | **0,98** | 0,28 | 0,53 | **0,80** | 0,09 | **0,96** | 0,30 | 0,52 |
| | | IIVotes | 0,76 | 0,20 | 0,89 | 0,43 | 0,55 | 0,77 | 0,22 | 0,90 | 0,44 | 0,56 |
| | | MSMOTEBagging | 0,72 | 0,29 | 0,82 | 0,49 | 0,56 | 0,76 | 0,30 | 0,87 | 0,51 | 0,58 |
| | | OverBagging2 | 0,72 | 0,34 | 0,81 | 0,52 | 0,57 | 0,74 | 0,36 | 0,83 | 0,55 | 0,60 |
| | | OverBagging | 0,72 | 0,25 | 0,83 | 0,46 | 0,54 | 0,76 | 0,21 | 0,89 | 0,43 | 0,55 |
| | | SMOTEBagging | 0,68 | 0,46 | 0,73 | 0,58 | **0,60** | 0,76 | 0,29 | 0,86 | 0,50 | 0,57 |
| | | UnderBagging2 | 0,66 | 0,50 | 0,69 | 0,59 | **0,60** | 0,67 | 0,45 | 0,71 | 0,57 | 0,58 |
| | | UnderBagging | 0,59 | 0,60 | 0,59 | **0,60** | **0,60** | 0,62 | **0,67** | 0,60 | 0,64 | **0,64** |
| | | UnderOverBagging | 0,71 | 0,42 | 0,77 | 0,57 | **0,60** | 0,72 | 0,39 | 0,79 | 0,55 | 0,59 |
| Hybrydowe | | BalanceCascade | 0,57 | 0,56 | 0,57 | 0,57 | 0,57 | 0,58 | 0,57 | 0,59 | 0,58 | 0,58 |
| | | EasyEnsemble | 0,58 | 0,58 | 0,58 | 0,58 | 0,58 | 0,59 | 0,59 | 0,59 | 0,59 | 0,59 |
| Wrażliwe na koszt | proste | C_SVM | 0,65 | 0,52 | 0,68 | 0,59 | **0,60** | 0,65 | 0,48 | 0,68 | 0,57 | 0,58 |
| | złożone | AdaBoost | 0,75 | 0,22 | 0,87 | 0,44 | 0,55 | 0,77 | 0,15 | 0,91 | 0,38 | 0,53 |
| | | AdaBoostM1 | 0,76 | 0,23 | 0,87 | 0,45 | 0,55 | 0,75 | 0,16 | 0,89 | 0,38 | 0,53 |
| | | AdaBoostM2 | 0,76 | 0,23 | 0,87 | 0,45 | 0,55 | 0,75 | 0,16 | 0,89 | 0,38 | 0,53 |
| | | AdaC2 | 0,47 | **0,70** | 0,42 | 0,54 | 0,56 | 0,52 | 0,59 | 0,51 | 0,55 | 0,55 |
| Próbkowanie + boosting | | DataBoost-IM | 0,74 | 0,30 | 0,84 | 0,50 | 0,57 | 0,77 | 0,21 | 0,90 | 0,43 | 0,55 |
| | | MSMOTEBoost | 0,66 | 0,34 | 0,73 | 0,49 | 0,53 | 0,72 | 0,31 | 0,81 | 0,51 | 0,56 |
| | | RUSBoost | 0,64 | 0,54 | 0,67 | **0,60** | **0,60** | 0,65 | 0,50 | 0,69 | 0,58 | 0,59 |
| | | SMOTEBoost | 0,69 | 0,41 | 0,76 | 0,56 | 0,58 | 0,74 | 0,22 | 0,85 | 0,43 | 0,53 |

Źródło: obliczenia własne.

Analiza wyników szczegółowych (tab. 2) nie wskazuje na zdecydowaną przewagę jednej z metod klasyfikacji danych niezbilansowanych w znaczeniu otrzymania zarówno akceptowalnej poprawności klasyfikacji (ACC, GM lub AUC), jak i zrównoważenia dokładności klasyfikacji obiektów z klasy zdominowanej (TPR względem TNR). Najbardziej zrównoważone wyniki dawały klasyfikatory, dla których jednocześnie wskaźnik AUC był największy (w większości były to metody z grupy próbkowanie + bagging), w szczególności algorytm UnderBagging, wykorzystujący klasyczny algorytm C4.5 Quinlana, niemniej wartości wskaźników GM i AUC na poziomie 0,60 nie wskazywały na szczególnie dobre zdolności predykcyjne. Rozszerzenie zbioru uczącego o dane okołooperacyjne PR (prawa część tab. 2) nie polepsza w istotny sposób jakości klasyfikacji.

Odmienna sytuacja występuje dla analizy ryzyka 5-letniego (tab. 4). W przypadku wykorzystania jako zbioru uczącego danych przedoperacyjnych PR zdecydowanie najlepsze wyniki (GM i AUC na poziomie 0,73) daje zastosowanie zaproponowano-

**Tabela 3.** Porównanie średnich wyników klasyfikacji dla metod z każdej grupy i przeżycia rocznego

| Grupa metod | Dane przedoperacyjne PR | | | | | Dane przed- i okołooperacyjne OP | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | TPR | TNR | GM | AUC | ACC | TPR | TNR | GM | AUC |
| Próbkowanie + bagging | **0,71** | 0,35 | **0,79** | 0,50 | 0,57 | **0,73** | 0,33 | **0,82** | 0,50 | **0,58** |
| Hybrydowe | 0,57 | **0,57** | 0,58 | 0,57 | 0,57 | 0,59 | **0,58** | 0,59 | 0,58 | **0,58** |
| C_SVM | 0,65 | 0,52 | 0,68 | **0,59** | **0,60** | 0,65 | 0,48 | 0,68 | 0,57 | **0,58** |
| Wrażliwe na koszt złożone | 0,68 | 0,34 | 0,76 | 0,47 | 0,55 | 0,70 | 0,27 | 0,80 | 0,42 | 0,53 |
| Próbkowanie + boosting | 0,68 | 0,40 | 0,75 | 0,54 | 0,57 | 0,72 | 0,31 | 0,81 | 0,49 | 0,56 |

Źródło: obliczenia własne.

**Tabela 4.** Porównanie średnich wyników klasyfikacji dla metod z każdej grupy i przeżycia 5-letniego

| Grupa metod | Dane przedoperacyjne PR | | | | | Dane przed- i okołooperacyjne OP | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | TPR | TNR | GM | AUC | ACC | TPR | TNR | GM | AUC |
| Próbkowanie + bagging | 0,74 | 0,47 | 0,78 | 0,57 | 0,63 | 0,78 | 0,71 | 0,79 | 0,73 | 0,75 |
| Hybrydowe | 0,60 | 0,65 | 0,59 | 0,62 | 0,62 | 0,71 | **0,90** | 0,68 | **0,78** | **0,79** |
| C_SVM | 0,72 | **0,75** | 0,72 | **0,73** | **0,73** | **0,84** | 0,60 | **0,88** | 0,73 | 0,74 |
| Wrażliwe na koszt złożone | **0,79** | 0,25 | **0,88** | 0,46 | 0,57 | 0,82 | 0,42 | **0,88** | 0,61 | 0,65 |
| Próbkowanie + boosting | 0,74 | 0,45 | 0,79 | 0,57 | 0,62 | 0,80 | 0,61 | 0,84 | 0,70 | 0,72 |

Źródło: obliczenia własne.

wanej w pracy implementacji algorytmu C-SVM. Wśród pozostałych algorytmów dobre wyniki dają także, jak poprzednio, niektóre algorytmy z grupy próbkowanie/eliminacja + bagging (AUC i GM rzędu 0,70). Po rozszerzeniu zbioru uczącego o dane okołooperacyjne wszystkie analizowane algorytmy wykazują znaczne polepszenie jakości klasyfikacji, przy czym najlepsze wyniki daje zastosowanie algorytmu UnderBagging; wskaźniki GM i AUC rzędu 0,83; TPR i TNR odpowiednio 0,99 i 0,65. Najważniejszą cechą zastosowanych algorytmów klasyfikacji danych niezbilansowanych jest znaczne zrównoważenie czułości i swoistości klasyfikacji (TPR i TNR) przy zachowaniu poprawności klasyfikacji na stosunkowo wysokim poziomie (0,75-0,84), co oznacza istotną poprawę wyników otrzymanych w poprzednim etapie badań, ale wskazuje na konieczność dalszych prac nad zwiększeniem zdolności predykcyjnej, szczególnie dla ryzyka krótkookresowego. Jednym z kierunków prac będzie próba opracowania uogólnienia algorytmu C-SVM, pozwalającego utrzymać zbilansowanie wyników dla większej wymiarowości problemu (rozszerzone zakresy danych OP, HP) przy jednoczesnym polepszeniu wskaźników jakości klasyfikacji AUC i GM.

## Literatura

Alcalá-Fdez J., Fernandez A., Luengo J., Derrac J., García S., Sánchez L., Herrera F., *KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework,* "Journal of Multiple-Valued Logic and Soft Computing" 2011, vol. 17(2-3), s. 255-287.

Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P., *SMOTE: Synthetic Minority Oversampling TEchnique*, "Journal of Artificial Intelligence Research" 2002, vol. 16, s. 321-357.

Ferguson M.K., Siddique J., Karrison T., *Modeling major lung resection outcomes using classification trees and multiple imputation techniques*, "European Journal of Cardio-Thoracic Surgery" 2008, vol. 34, s. 1085-1089.

Galar M., Fernández A., Barrenechea E., Bustince H., Herrera F., *A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches*, IEEE Transactions On Systems, Man and Cybernetics-Part C: Applications and Reviews 2012, vol. 42(4), s. 463-484.

Gatnar W., *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*, WN PWN, Warszawa 2008.

He H., Garcia E.A., *Learning from Imbalanced Data*, IEEE Transactions on Knowledge and Data Engineering 2009, vol. 21(9), s. 1263-1284.

Lubicz M., Zięba M., Rzechonek A., Pawełczyk K., Kołodziej J., Błaszczyk J., *Analiza porównawcza wybranych technik eksploracji danych do klasyfikacji danych medycznych z brakującymi obserwacjami*, [w:] K. Jajuga, M. Walesiak (red.), Taksonomia 19, *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, 2012, s. 416-425.

Rivo E., De La Fuente J., Rivo A., García-Fontán E., Cañizares M.-A., Gil, P., *Cross-Industry Standard Process for data mining is applicable to the lung cancer surgery domain, improving decision making as well as knowledge and quality management*, "Clinical and Translational Oncology" 2012, vol. 14(1), s. 73-79.

Sun Y., Kamel M., Wong A., Wang Y., *Cost-sensitive boosting for classification of imbalanced data,* "Pattern Recognition" 2007, vol. 40, s. 3358-3378.

Tang Y., Zhang Y-Q., Chawla N.V., Krasser S., *SVMs Modeling for Highly Imbalanced Classification*, IEEE Transactions On Systems, Man and Cybernetics-Part B: Cybernetics 2009, vol. 39(1), s. 281-288.

Zięba M., *Ensemble decision trees for customer classification in service oriented systems*, Wydział Informatyki i Zarządzania Politechniki Wrocławskiej (niepublikowana praca magisterska), 2011.

## MODELLING CLASS IMBALANCE PROBLEMS: COMPARING CLASSIFICATION APPROACHES FOR SURGICAL RISK ANALYSIS

**Summary:** In classification tasks based on real-world data, for instance when analyzing medical data, it is quite often necessary to deal with problems related to the nature of data, in particular with class imbalance, when the number of examples that represent one class is much lower than the ones of the other classes. The aim of this paper is to perform comparative analysis of selected classification approaches, designed for imbalanced data sets. The research was performed using Imbalanced Learning Module of the KEEL Data Mining software package and WEKA Machine Learning environment. The source data was extracted from updated hospital data base of surgical lung cancer patients treated at Wrocław Thoracic Surgery Centre in the period 2000-2011.

**Keywords:** data mining, classification, class imbalance, missing values, medical data.