PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 328

Taksonomia 23
Klasyfikacja i analiza danych – teoria i zastosowania

Redaktorzy naukowi
Krzysztof Jajuga, Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2014

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

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

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

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

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

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Summaries

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

Marek Lubicz, Maciej Zięba
Politechnika Wrocławska

Konrad Pawełczyk, Adam Rzechonek, Marek Marciniak, Jerzy Kołodziej
Uniwersytet Medyczny we Wrocławiu

INDUKCJA REGUŁ DLA DANYCH NIEKOMPLETNYCH I NIEZBALANSOWANYCH: MODELE KLASYFIKATORÓW I PRÓBA ICH ZASTOSOWANIA DO PREDYKCJI RYZYKA OPERACYJNEGO W TORAKOCHIRURGII[1]

Streszczenie: Artykuł dotyczy klasyfikacji obiektów w sytuacji łącznego występowania wielu niedoskonałości dostępnych danych, w szczególności: niekompletności i niezbalansowania danych. Zaproponowano zastosowanie podejścia wykorzystującego adaptacje wzmocnionych, wrażliwych na koszt klasyfikatorów SVM. Porównano efektywność podejścia z klasyfikatorami dla danych niezbalansowanych, dostępnymi w środowiskach uczenia maszynowego KEEL i WEKA. Rozważono też zagadnienie interpretowalności wyników klasyfikacji (indukcja reguł i drzew decyzyjnych z efektywnych modeli typu „czarna skrzynka"). Omówiono przykładowe zastosowanie do klasyfikacji zaktualizowanych baz danych medycznych z Wrocławskiego Ośrodka Torakochirurgii.

Słowa kluczowe: uczenie maszynowe, klasyfikacja, indukcja reguł, dane niezbalansowane, brakujące obserwacje, predykcja ryzyka operacyjnego.

1. Wstęp

Istotną trudnością przy klasyfikacji obiektów na podstawie danych rzeczywistych jest łączne występowanie wielu niedoskonałości dostępnych danych, w szczególności: niekompletności (brakujące obserwacje niektórych cech) oraz niezbalanso-

[1] Praca naukowa częściowo 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 ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

wania (znaczna przewaga liczebności jednej lub kilku klas). Takie sytuacje występują m.in. w problemach klasyfikacji danych medycznych, w tym w rozważanym przez autorów problemie analizy ryzyka operacyjnego w torakochirurgii (krótko- lub długookresowa predykcja zgonu po operacji). Specyfika danych utrudnia dobór efektywnego klasyfikatora z wykorzystaniem standardowych metod uzupełniania brakujących obserwacji i metod redukcji niezbalansowania. W literaturze przedmiotu wymienione problemy są zwykle rozpatrywane oddzielnie [np. Gatnar 2001; Marshall i in. 2010; Galar i in. 2012], a próby zastosowań praktycznych w rozważanej dziedzinie dają jak dotychczas niezadowalające rezultaty [Poullis i in. 2013]. Dodatkowe trudności przy doborze efektywnego podejścia są związane z wymaganiem interpretowalności przebiegu i wyników klasyfikacji w przypadkach ich praktycznego wykorzystania: efektywne klasyfikatory, takie jak sieci neuronowe czy metoda wektorów nośnych (SVM) [Gatnar 2008], są modelami typu „czarna skrzynka" – w odróżnieniu od mniej efektywnych klasyfikatorów regułowych lub opartych na drzewach decyzyjnych [Barakat, Bradley 2010]. Celem pracy jest analiza efektywności podejścia wykorzystującego adaptacje wzmocnionych, wrażliwych na koszt klasyfikatorów SVM [Zięba 2013] w porównaniu z klasyfikatorami dla danych niezbalansowanych, dostępnymi w podstawowych niekomercyjnych środowiskach uczenia maszynowego KEEL [Alcalá-Fdez i in. 2011] i WEKA [Witten i in. 2011], omówienie podejścia do indukcji reguł i drzew decyzyjnych [Zięba i in. 2014] oraz wstępnych wniosków z jego zastosowania na przykładach wykorzystujących bazy danych medycznych z Wrocławskiego Ośrodka Torakochirurgii.

2. Założenia badawcze i wyniki poprzednich badań

Problem predykcji ryzyka operacyjnego w torakochirurgii sformułowano jako zadanie klasyfikacji obiektów, którymi są pacjenci oddziału chirurgii klatki piersiowej z diagnozą pierwotnego raka płuca, operowani w latach 2007-2012, do jednej z dwóch klas: pacjenci, którzy przeżyli bądź też nie przeżyli ustalonego okresu obserwacji. Zależnie od przyjętego horyzontu czasowego (30 dni, 1 rok, 5 lat od operacji) jedna klasa jest klasą zdominowaną (mało liczną) w porównaniu z drugą – klasą dominującą. W niniejszym artykule ograniczono się do oceny ryzyka zgonu w ciągu 1 roku po operacji, a klasą zdominowaną jest podgrupa pacjentów, którzy zmarli w tym okresie. W rozważaniach skupiono się na metodach klasyfikacji wzorcowej, zaliczanych do metod analizy dyskryminacyjnej (uczenia z nauczycielem, uczenia nadzorowanego [Gatnar 1998]). Istotą tych metod jest określanie reguł klasyfikacji na podstawie informacji zawartych w ciągu uczącym, tj. zbiorze obiektów, których przynależność do klasy jest znana. W rozważanym przypadku obserwowanymi cechami są zmienne (z reguły nominalne) określające stan pacjenta lub zdarzenia diagnostyczno-terapeutyczne podczas hospitalizacji i po jej zakończeniu (przedoperacyjne, okołooperacyjne, histopatologia, pooperacyjne).

W badaniach wykorzystano zaktualizowane bazy danych, utworzone na podstawie danych Wrocławskiego Ośrodka Torakochirurgii, uzupełnione o rejestry Dolnośląskiego Centrum Onkologii i Dolnośląskiego Oddziału Wojewódzkiego NFZ. Podstawowa baza zawiera dane o 1384 pacjentach opisanych 239 cechami (odpowiednio: 114 przedoperacyjnymi, 32 okołooperacyjnymi, 63 histopatologicznymi, 30 pooperacyjnymi [Lubicz i in. 2010]). W danych źródłowych stwierdzono różnorodne niedoskonałości, m.in.:

- braki danych: po uzupełnieniu podstawowych braków pozostało 161 pacjentów z niekompletnymi danymi (0,23% - 65% obiektów z brakiem danych, zależnie od cechy),
- niezbalansowanie danych: w analizowanej próbie wystąpiło 237 zgonów do 1 roku od operacji (klasa zdominowana), co daje wskaźnik niezbalansowania 4,84 (iloraz liczebności klasy dominującej do liczebności klasy zdominowanej).

Efektywność klasyfikacji oceniano na podstawie prawidłowości predykcji wystąpienia zgonu pacjenta w ciągu roku od operacji (P – klasa pozytywna) i predykcji przeżycia 1 roku (N – klasa negatywna). Zastosowano następujące wskaźniki jakości klasyfikacji:

TPR – czułość klasyfikacji: $TPR = TP / (TP + FN)$,

TNR – swoistość klasyfikacji: $TNR = TN / (FP + TN)$,

FPR – odsetek błędów I rodzaju: $FPR = FP / (FP + TN)$,

ACC – dokładność (odsetek poprawnych klasyfikacji): $ACC = (TP + TN) / NN$

oraz szczególnie istotne przy klasyfikacji danych niezbalansowanych [Galar i in. 2012]:

GM – współczynnik średniej geometrycznej jakości predykcji: $GM = \sqrt{TPR * TNR}$,

AUC – wskaźnik reprezentujący pole powierzchni pod krzywą ROC (Receiver Operating Characteristic): $AUC = 0,5 * (1 + TPR - FPR)$,

gdzie poszczególne symbole oznaczają: *TP*, *TN* – liczby prawidłowo sklasyfikowanych pacjentów z klasy – odpowiednio: pozytywnej i negatywnej oraz – analogicznie: *FP*, *FN* – liczby błędnych klasyfikacji; *NN* – łączna liczba klasyfikowanych.

Przy wyznaczaniu metodyki badań uwzględniono wyniki poprzednich badań, obejmujące:

- w roku 2010: analizę literatury przedmiotu z zakresu oceny ryzyka operacyjnego i doświadczeń zastosowania metod statystycznych i technik eksploracji danych w torakochirurgii [Lubicz i in. 2010]; wykazano m.in. niejednoznaczność w ocenie istotnych czynników ryzyka pomimo prowadzenia wieloośrodkowych badań na dużych liczbach pacjentów oraz brak przewagi którejkolwiek z najczęściej stosowanych metod (regresja logistyczna, sieci neuronowe, SVM, drzewa decyzyjne [Santos-Garcia i in. 2004; Ferguson i in. 2008];

- w roku 2011: analizę porównawczą wybranych podejść (usunięcie przypadków, zastąpienie wartością średnią lub medianą, imputacja z zastosowaniem różnych algorytmów) do wstępnego przetwarzania przy brakach danych, które w przypadku oceny ryzyka rocznego zapewniały dokładność (ACC) rzędu 75-85% i wysoką swoistość (TNR) rzędu 94-96% przy nikłej czułości (TPR) klasyfikacji wystąpienia zgonu rzędu 10-30% [Lubicz i in. 2012];
- w roku 2012: analizę porównawczą wybranych podejść do klasyfikacji danych niezbalansowanych [Lubicz i in. 2013], umożliwiających znaczne zrównoważenie precyzji wykrywania klasy pozytywnej i negatywnej: czułość (TPR) rzędu 50-60% przy swoistości (TNR) rzędu 60-80% i dokładności (ACC) rzędu 60%.

W obecnym etapie badań przyjęto założenia metodyczne obejmujące:

- aktualizację analizy literatury przedmiotu z zakresu modelowania ryzyka operacyjnego; nie stwierdzono znaczącego postępu w zakresie wyznaczania czynników i modelowania ryzyka operacyjnego pacjentów z rakiem płuca [Bradley i in. 2012; Poullis i in. 2013; Qadri i in. 2013; Mediratta i in. 2014], także w odniesieniu do nielicznych przykładów zastosowania technik eksploracji danych (np. w [Rivo i in. 2012] jako sukces podaje się model regresji logistycznej z wysoką dokładnością i swoistością klasyfikacji rzędu 95-100% przy czułości 9,68%, co nie dziwi przy wskaźniku niezbalansowania 16,7);
- zastosowanie rozszerzonej metody uczenia klasyfikatorów typu SVM dla danych niezbalansowanych opisanej w [Zięba 2013] i opracowanie efektywnego klasyfikatora hybrydowego (wielomodelowego) BoostingCSVM dla danych niezbalansowanych, łączącego uczenie wrażliwe na koszt z algorytmem boosting [Zięba i in. 2014];
- analizę porównawczą klasyfikatora BoostingCSVM oraz algorytmów klasyfikacji danych niezbalansowanych, z modułu Imbalanced Learning pakietu KEEL i pakietu WEKA;
- wykorzystanie klasyfikatora BoostingCSVM i klasycznego algorytmu regułowego (Ripper) do wyznaczenia zestawu reguł decyzyjnych [Zięba i in. 2014] oraz próbę ich merytorycznej interpretacji przez specjalistów klinicznych.

3. Efektywny klasyfikator BoostingCSVM: model i analiza porównawcza

Kluczowym elementem koncepcji wzmacnianego klasyfikatora BoostingCSVM dla danych niezbalansowanych [Zięba 2013] jest procedura uczenia klasyfikatorów bazowych, polegająca na minimalizacji szczególnie sformułowanej ważonej funkcji błędu z niesymetrycznymi kosztami błędnych klasyfikacji. W zaproponowanym podejściu jako model bazowy dla klasyfikatora wzmacnianego przyjęto klasyfikator typu SVM, który jest uczony z wykorzystaniem algorytmu SMO [Keerthi i in.

2001]. W kolejnym etapie, w celu poprawy skuteczności klasyfikacji i zniwelowania negatywnych skutków jednoczesnego niezbalansowania danych w ramach klas i pomiędzy klasami, uczenie wrażliwe na koszt jest połączone z algorytmem wzmacniania typu AdaBoost [Freund i in. 1996; Gatnar 2008]) w ramach wzmocnionego, wielomodelowego klasyfikatora BoostingCSVM, do którego – w oparciu o wyniki przedstawione w [Wang, Japkowicz 2010] – wybierane są jedynie klasyfikatory bazowe o najwyższych wartościach wskaźnika GM. Formalne omówienie zaproponowanego podejścia przedstawiono w [Zięba i in. 2014].

Efektywność klasyfikatora BoostingCSVM analizowano po zaimplementowaniu w module Imbalanced Learning pakietu KEEL w porównaniu z dostępnymi w tym środowisku klasyfikatorami wrażliwymi na koszt i wielomodelowymi dla danych niezbalansowanych. Najlepsze wyniki klasyfikacji (wysokie GM, zrównoważone TNR i TPR) na opisanych wyżej danych medycznych otrzymano dla algorytmów opartych na zasadzie eliminacji obiektów z klasy dominującej i wykorzystaniu algorytmu C45 jako klasyfikatora bazowego: RUSBoost-I (wzmacnianej eliminacji losowej [Seiffert i in. 2010]), UnderBagging-I (eliminacji z agregacją bootstrapową [Barandela i in. 2003]) i EasyEnsemble [Liu i in. 2009], wrażliwej na koszt wersji klasycznego klasyfikatora C45 (C45CS-I; [Ting 2002]) oraz klasyfikatora BoostingCSVM, dla którego otrzymano najwyższą efektywność klasyfikacji. Zwraca uwagę niska efektywność klasyfikacji dla algorytmu AdaBoost i klasyfikatorów opartych na popularnym w literaturze przedmiotu algorytmie SMOTE [Chawla i in. 2002] równoważenia rozkładu klas przez generowanie nowych obserwacji.

Tabela 1. Porównanie efektywności klasyfikatora BoostingCSVM oraz metod klasyfikacji wrażliwej na koszt (CS) i wielomodelowych (EN) dla danych niezbalansowanych

Grupa metod	Algorytm	Dane przedoperacyjne					Dane przed- i okołooperacyjne				
		ACC	TN	TPR	GM	AU	AC	TN	TPR	GM	AU
BoostingCSVM		0,69	**0,71**	**0,60**	**0,65**	0,65	0,69	**0,72**	**0,55**	**0,63**	0,64
CS	C45CS-I	0,60	**0,60**	**0,56**	**0,58**	0,58	0,68	0,73	0,42	0,56	0,58
EN	AdaBoost -I, M1,	0,84	0,94	0,11	0,32	0,52	0,88	1,00	0,00	0,00	0,50
	EasyEnsemble-I	0,64	**0,66**	**0,54**	**0,59**	0,60	0,84	0,92	0,30	0,53	0,61
	RUSBoost-I	0,72	**0,74**	**0,57**	**0,65**	0,65	0,52	**0,52**	**0,50**	**0,51**	0,51
	SMOTEBagging-I	0,86	0,94	0,25	0,48	0,59	0,87	0,98	0,02	0,13	0,50
	SMOTEBoost-I	0,81	0,88	0,27	0,49	0,58	0,71	0,77	0,21	0,41	0,49
	UnderBagging-I	0,68	**0,70**	**0,54**	**0,61**	0,62	0,85	0,92	0,29	0,51	0,60

Źródło: obliczenia własne w środowisku KEEL.

Wyniki obliczeń dla wszystkich danych (tabela 2) dostępnych w okresie prowadzenia badań wykazują, że klasyfikator BoostedCSVM jest efektywnym i stabilnym narzędziem klasyfikacji danych niezbalansowanych; jako jedyny z klasyfikatorów dla danych niezbalansowanych analizowanych w środowisku KEEL

Tabela 2. Porównanie zmian jakości klasyfikacji w miarę akumulacji wiedzy (zwiększenie liczby cech)

Dane	Przedoperacyjne			+Okołooperacyjne			+Histopatologia			+Pooperacyjne		
Algorytm	TNR	TPR	GM	TNR	TPR	GM	TNR	TPR	GM	TNR	TPR	GM
BoostingCSVM	0,71	0,60	**0,65**	0,72	**0,55**	**0,63**	0,75	**0,52**	**0,63**	0,78	0,50	**0,62**
C45CS-I	0,60	0,56	0,58	0,73	0,42	0,56	0,75	0,35	0,51	0,77	0,49	**0,61**
RUSBoost-I	0,74	0,57	**0,65**	0,52	0,50	0,51	0,63	0,46	0,54	0,47	**0,53**	0,50
UnderBagging-I	0,70	0,54	**0,61**	0,92	0,29	0,51	0,87	0,36	0,56	0,88	0,35	0,55

Źródło: obliczenia własne w środowisku KEEL.

zachował stabilność przy zwiększonej liczbie zmiennych objaśniających (kolejne trójki kolumn w tabeli 2 odpowiadają wektorowi cech zwiększonemu o kolejne grupy danych klinicznych) oraz z reguły wykazywał najlepszą (chociaż ze względów klinicznych wciąż niedostateczną) czułość klasyfikacji (TPR; zdolność prospektywnego stwierdzenia potencjalnie negatywnych zdarzeń przy kwalifikacji do zabiegu operacyjnego). Należy zauważyć, że przedmiotem obecnych badań nie były metody doboru zmiennych objaśniających (selekcji cech), co jest zagadnieniem istotnym [Gatnar 2008], szczególnie przy dużej wymiarowości problemu; przewiduje się dalsze badania w tym zakresie, szczególnie w związku z planowanym rozszerzeniem wektora cech o dane laboratoryjne [Warwick i in. 2014] i immunohistochemiczne [Zhu i in. 2009].

4. Indukcja interpretowalnych reguł klasyfikacyjnych: model i próba zastosowania

W literaturze z zakresu uczenia maszynowego podkreśla się, że reprezentacja wiedzy klasyfikacyjnej powinna cechować się: wysoką jakością klasyfikacji, odpornością na niedoskonałości danych wykorzystywanych w procesie uczenia i zrozumiałością, tzn. powinna być wyrażona w formie pozwalającej na zrozumienie, ocenę i dalsze użycie przez człowieka [Stefanowski 2001]. Postulat ten uwzględnia się w tych metodach, które tworzą symboliczne reprezentacje wiedzy, takie jak drzewa i reguły decyzyjne. Wadą klasyfikatorów charakteryzujących się zrozumiałą reprezentacją wiedzy jest najczęściej niska jakość klasyfikacji i niewielka odporność na złą jakość danych, dlatego też zachodzi konieczność wykorzystania klasyfikatorów o wysokiej jakości predykcyjnej do konstrukcji modeli o wysokim stopniu interpretowalności. Wyróżnia się trzy główne podejścia do problemu indukcji reguł z modeli, które charakteryzują się wysoką skutecznością predykcji (np. SVM, sieci neuronowe), ale nie są bezpośrednio interpretowalne [Tickle i in. 1998]:

– dekompozycyjne: dokonuje się indukcji reguł poprzez analizę struktury wyuczonego modelu, np. sieci neuronowej lub marginesu separującego wyznaczonego przez SVM,

- pedagogiczne, niewymagające analizy struktury modelu, działające niezależnie od stosowanego klasyfikatora, zakładające wygenerowanie nowej porcji danych, ich zaetykietowanie lub modyfikację etykiet obserwacji ze zbioru uczącego z wykorzystaniem wysokiej jakości metody klasyfikacji, wyuczonej na pierwotnym zbiorze uczącym oraz konstrukcję drzewa lub reguł decyzyjnych z zastosowaniem metod klasycznych, np. RIPPER lub C4.5,
- eklektyczne, łączące cechy dwóch poprzednich.

W niniejszej pracy autorzy zastosowali podejście drugiego typu dla klasyfikatora BoostingCSVM w środowisku KEEL, będącego połączeniem trudnych do interpretacji modeli: klasyfikatora wzmacnianego i modelu typu SVM, obejmującego [Zięba i in. 2014]:

- wyuczenie klasyfikatora BoostingCSVM na wejściowym zestawie danych po uprzednim dobraniu właściwych parametrów uczenia, np. z wykorzystaniem walidacji krzyżowej,
- reetykietyzację elementów ciągu uczącego na podstawie wyuczonego klasyfikatora,
- wykorzystanie zmodyfikowanego ciągu uczącego do wyznaczenia zestawu reguł decyzyjnych z zastosowaniem algorytmu regułowego Ripper (JRip; [Witten i in. 2011]).

Zastosowanie tego podejścia do omówionych wyżej danych medycznych (dane przedoperacyjne, ocena przeżycia 1 roku) daje w wyniku efektywność klasyfikacji porównywalną z pierwotną (GM = 0,62; ACC = 0,69) oraz 12 reguł decyzyjnych. Dla porównania wygenerowano również reguły decyzyjne z wykorzystaniem innych klasyfikatorów w środowiskach KEEL i WEKA (np. najlepszy w środowisku WEKA algorytm AdaBoost.M1 z funkcją bazową J48 dał 118 reguł pozytywnych, ujętych w drzewie decyzyjnym po ostatecznym przycięciu). Zestawy reguł decyzyjnych poddano interpretacji i próbie merytorycznej oceny specjalistów klinicznych, uzyskując potwierdzenie zasadności klinicznej części reguł, m.in. znacznego ryzyka zgonu pooperacyjnego dla pacjentów z oceną kliniczną sTNM dla T = 3, 4 lub N = 2, z przerzutami (m.in. do centralnego układu nerwowego, drugiego płuca,

Tabela 3. Przykładowe reguły dla przeżycia 1-rocznego na podstawie danych przedoperacyjnych

1	(cukrzyca insulinozależna) i (rozpoznanie nowotworu przed operacją) => R1Yr=T
2	(mężczyzna) i (przerzuty do innych narządów: nadnercza lub drugie płuco, lub wątroba) => R1Yr=T
3	(kobieta) i (brak cukrzycy insulinozależnej) => R1Yr=N
4	(mężczyzna) i (choroba wieńcowa lub zaburzenia krążenia mózgowego) => R1Yr=T
5	(mężczyzna) i (palenie tytoniu) i (rozpoznanie nowotworu przed operacją) => R1Yr=T
6	(mężczyzna) i (palenie tytoniu) i (pogorszenie sprawności) i (przewlekła obturacyjna choroba płuc) i (powiększone węzły N1 lub ocena według skali Zubroda =1 lub 2) => R1Yr=T

Konkluzja R1Yr = T oznacza wysokie, a R1Yr = N – niskie ryzyko zgonu w ciągu roku po operacji.

Źródło: obliczenia własne w środowiskach KEEL i WEKA.

nadnerczy, wątroby), wystąpienie zawału serca w ciągu 6 miesięcy przed operacją lub zespołu metabolicznego, a także istotnej wartości predykcyjnej parametrów spirometrycznych FEV, płci i wieku [por. Mediratta i in. 2014]. Inne przykładowe reguły zamieszczono w tabeli 3. Dalsze prace w powyższym zakresie będą dotyczyć wygenerowania kompletnych zestawów reguł dla klasyfikatorów o największej zdolności predykcyjnej i ich porównania z chirurgicznymi systemami oceny ryzyka [Bradley i in. 2012; Qadri i in. 2013].

Literatura

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

Barakat N., Bradley A.P., 2010, *Rule extraction from support vector machines: A review*, Neurocomputing, vol. 74(41277), s. 178-190.

Barandela R., Valdovinos R.M., Sánchez J.S., 2003, *New applications of ensembles of classifiers*, Pattern Analysis and Applications, vol. 6, s. 245-256.

Bradley A., Marshall A., Abdelaziz M., Hussain K., Agostini P., Bishay E., Kalkat M., Steyn R., Rajesh P., Dunn J., Naidu B., 2012, *Thoracoscore fails to predict complications following elective lung resection*, European Respiratory Journal, 40(6), 1496-1501.

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

Ferguson M.K., Siddique J., Karrison T., 2008, *Modeling major lung resection outcomes using classification trees and multiple imputation techniques*, European Journal of Cardio-Thoracic Surgery, vol. 34(5), s. 1085-1089.

Fernández A., García S., Luengo J., Bernadó-Mansilla E., Herrera F., 2010, *Genetics-Based Machine Learning for Rule Induction: State of the Art, Taxonomy, and Comparative Study*, IEEE Transactions on Evolutionary Computation, vol. 14(6), s. 913-941.

Freund Y., Schapire R.E., Hill M., 1996, *Experiments with a New Boosting Algorithm*, Proceedings of the 13th International Conference on Machine Learning, Morgan Kaufmann, s. 148-156.

Galar M., Fernández A., Barrenechea E., Bustince H., Herrera F., 2012, *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, vol. 42(4), s. 463-484.

García S., Fernández A., Herrera F., 2009, *Enhancing the effectiveness and interpretability of decision tree and rule induction classifiers with evolutionary training set selection over imbalanced problems*, Applied Soft Computing, vol. 9(4), s. 1304-1314.

Gatnar E., 1998, *Symboliczne metody klasyfikacji danych*, PWN, Warszawa.

Gatnar E., 2001, *Nieparametryczna metoda dyskryminacji i regresji*, WN PWN, Warszawa

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

Keerthi S.S., Shevade S.K., Bhattacharyya C., Murthy K.R.K., 2001, *Improvements to Platt's SMO algorithm for SVM classifier design*, Neural Computation, 13, s. 637-649.

Liu X-Y., Wu J., Zhou Z-H., 2009, *Exploratory undersampling for class-imbalance learning*, IEEE Transactions on Systems, Man, and Cybernetics, Part B, vol. 39(2), 539-550.

Lubicz M., Rzechonek A., Pawełczyk K., Kołodziej J., Adamiak J., 2010, *Knowledge extraction and surgical risk modelling: intelligent support for thoracic surgery*, [w:] *Applications of Systems Science*, A. Grzech i in. (red.), EXIT, Warszawa, s. 327-336.

Lubicz M., Zięba M., Pawełczyk K., Rzechonek A., Kołodziej J., 2013, *Modele eksploracji danych niezbilansowanych – procedury klasyfikacji dla zadania analizy ryzyka operacyjnego*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 278, Wrocław, s. 262-270.

Lubicz M., Zięba M., Rzechonek A., Pawełczyk K., Kołodziej J., Błaszczyk J., 2012, *Analiza porównawcza wybranych technik eksploracji danych do klasyfikacji danych medycznych z brakującymi obserwacjami*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 242, Wrocław, s. 416-425.

Marshall A., Altman D.G., Royston P., Holder R.L., 2010, *Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study*, BMC Medical Research Methodology, vol. 10(7).

Mediratta N., Shackcloth M., Page R., Woolley S., Asante-Siaw J., Poullis M., 2014, *Should males ever undergo wedge resection for stage 1 non-small-cell lung cancer? A propensity analysis*. European Journal of Cardio-Thoracic Surgery (w druku).

Poullis M., McShane J., Shaw M., Woolley S., Shackcloth M., Page R., Mediratta N., 2013, *Prediction of in-hospital mortality following pulmonary resections: improving on current risk models*, European Journal of Cardio-Thoracic Surgery, vol. 44, s. 238-243.

Qadri S.S.A., Jarvis M., Ariyaratnam P., Chaudhry M.A., Cale A.R.J., Griffin S., Cowen M.E., Loubani M., 2013, *Could Thoracoscore predict postoperative mortality in patients undergoing pneumonectomy?*, European Journal of Cardio-Thoracic Surgery (w druku).

Rivo E., De La Fuente J., Rivo A., García-Fontán E., Cañizares M.-A., Gil, P., 2012, *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, vol. 14(1), s. 73-79.

Santos-Garcıa G., Varela G., Novoa N., Jimenez M.F., 2004, *Prediction of postoperative morbidity after lung resection using an artificial neural network ensemble*, Artificial Intelligence in Medicine, 30(1), s. 61-69.

Seiffert C., Khoshgoftaar T., Van Hulse J., Napolitano A., 2010, *Rusboost: A hybrid approach to alleviating class imbalance, IEEE Transactions on Systems*, Man and Cybernetics, Part A, vol. 40(1), s. 185-197.

Stefanowski J., 2001, *Algorytmy indukcji reguł decyzyjnych w odkrywaniu wiedzy, rozprawa habilitacyjna*, Wydawnictwo Politechniki Poznańskiej, seria Rozprawy, nr 361, Poznań.

Tickle A.B., Andrews R., Golea M., Diederich J., 1998, *The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural networks*, IEEE Transactions on Neural Networks, vol. 9(6), s.1057-1068.

Ting K.M., 2002, *An instance-weighting method to induce cost-sensitive trees, IEEE Transactions on Knowledge and Data Engineering*, vol. 14(3), s. 659-665.

Wang B., Japkowicz N., 2010, *Boosting support vector machines for imbalanced datasets*, Knowledge and Information Systems, vol. 25, s. 1-20.

Warwick R., Mediratta N., Shackcloth M., Shaw M., McShane J., Poullis M., 2014, *Preoperative red cell distribution width in patients undergoing pulmonary resections for non-small-cell lung cancer*, European Journal of Cardio-Thoracic Surgery, vol. 45, s. 108-113.

Witten I.H., Frank E., Hall M.A., 2011, *Data Mining: Practical Machine Learning Tools and Techniques*, Elsevier, Amsterdam.

Zhu Z.-H., Sun B.-Y., Ma Y., Shao J.-Y., Long H., Zhang X., Fu J.-H., Zhang L.-J., Su X.-D., Wu Q.-L., Ling P., Chen M., Xie Z.-M., Hu Y., Rong T.-H., 2009, *Three Immunomarker Support Vector Machines–Based Prognostic Classifiers for Stage IB Non-Small-Cell Lung Cancer*, Journal of Clinical Oncology, vol. 27(7), s. 1091-1099.

Zięba M., 2013, *Opracowanie zespołów klasyfikatorów SVM dla danych niezbalansowanych na potrzeby wspomagania decyzji w systemach informatycznych*, rozprawa doktorska, Politechnika Wrocławska.

Zięba M., Tomczak J.M., Lubicz M., Świątek J., 2014, *Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients*, Applied Soft Computing Journal, vol. 14(A), s. 99-108.

CLASSIFICATION RULES EXTRACTION FOR MISSING AND IMBALANCE DATA: MODELS OF CLASSIFIERS AND INITIAL RESULTS IN THE RULES-BASED THORACIC SURGERY RISK PREDICTION

Summary: The classification problem of multi-faceted imperfect data, e.g. with missing values and at the same time with class imbalance, is considered. Aspects of the classification effectiveness and interpretability of the results through classification rules extraction for the ”black-box” like classifiers are discussed. An approach based on a boosted SVM classifier and an oracle-based decision rules extraction procedure is proposed and applied to a sample hospital data base of Wrocław Thoracic Surgery Centre. The research was performed using Imbalanced Learning Module of the KEEL Data Mining software package and WEKA Machine Learning environment.

Keywords: data mining, classification, rules extraction, class imbalance, missing values, surgical risk prediction.