PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 426

Taksonomia 26
Klasyfikacja i analiza danych – teoria i zastosowania

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2016

Redaktor Wydawnictwa: Agnieszka Flasińska

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

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

Informacje o naborze artykułów i zasadach recenzowania
znajdują się na stronach internetowych
www.pracenaukowe.ue.wroc.pl
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
e-ISSN 2392-0041
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Zamówienia na opublikowane prace należy składać na adres:
Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
ul. Komandorska 118/120, 53-345 Wrocław
tel./fax 71 36 80 602; e-mail:econbook@ue.wroc.pl
www.ksiegarnia.ue.wroc.pl

Druk i oprawa: TOTEM

Spis treści

Wstęp

W dniach 14–16 września 2015 r. w Hotelu Novotel Gdańsk Marina w Gdańsku odbyła się XXIV Konferencja Naukowa Sekcji Klasyfikacji i Analizy Danych PTS (XXIX Konferencja Taksonomiczna) „Klasyfikacja i analiza danych – teoria i zastosowania", zorganizowana przez Sekcję Klasyfikacji i Analizy Danych Polskiego Towarzystwa Statystycznego oraz Katedrę Statystyki Wydziału Zarządzania Uniwersytetu Gdańskiego. Przewodniczącymi Komitetu Organizacyjnego konferencji byli prof. dr hab. Mirosław Szreder oraz dr hab. Krzysztof Najman, prof. nadzw. UG, sekretarzami naukowymi dr hab. Kamila Migdał-Najman, prof. nadzw. UG oraz dr hab. Anna Zamojska, prof. nadzw. UG, a sekretarzem organizacyjnym Anna Nowicka z Fundacji Rozwoju Uniwersytetu Gdańskiego.

Konferencja Naukowa została dofinansowana ze środków Narodowego Banku Polskiego.

Zakres tematyczny konferencji obejmował takie zagadnienia, jak:

a) teoria (taksonomia, analiza dyskryminacyjna, metody porządkowania liniowego, metody statystycznej analizy wielowymiarowej, metody analizy zmiennych ciągłych, metody analizy zmiennych dyskretnych, metody analizy danych symbolicznych, metody graficzne),

b) zastosowania (analiza danych finansowych, analiza danych marketingowych, analiza danych przestrzennych, inne zastosowania analizy danych – medycyna, psychologia, archeologia, itd., aplikacje komputerowe metod statystycznych).

Zasadniczymi celami konferencji SKAD były prezentacja osiągnięć i wymiana doświadczeń z zakresu teoretycznych i aplikacyjnych zagadnień klasyfikacji i analizy danych. Konferencja stanowi coroczne forum służące podsumowaniu obecnego stanu wiedzy, przedstawieniu i promocji dokonań nowatorskich oraz wskazaniu kierunków dalszych prac i badań.

W konferencji wzięło udział 81 osób. Byli to pracownicy oraz doktoranci następujących uczelni i instytucji: AGH w Krakowie, Politechniki Łódzkiej, Politechniki Gdańskiej, Politechniki Opolskiej, Politechniki Wrocławskiej, Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie, Szkoły Głównej Handlowej w Warszawie, Uniwersytetu im. Adama Mickiewicza w Poznaniu, Uniwersytetu Ekonomicznego w Katowicach, Uniwersytetu Ekonomicznego w Krakowie, Uniwersytetu Ekonomicznego w Poznaniu, Uniwersytetu Ekonomicznego we Wrocławiu, Uniwersytetu Gdańskiego, Uniwersytetu Jana Kochanowskiego w Kielcach, Uniwersytetu Łódzkiego, Uniwersytetu Mikołaja Kopernika w Toruniu, Uniwersytetu Przyrodniczego w Poznaniu, Uniwersytetu Szczecińskiego, Uniwer-

sytetu w Białymstoku, Wyższej Szkoły Bankowej w Toruniu, a także przedstawiciele NBP i PBS Sp. z o.o.

W trakcie dwóch sesji plenarnych oraz trzynastu sesji równoległych wygłoszono 58 referatów poświęconych aspektom teoretycznym i aplikacyjnym zagadnienia klasyfikacji i analizy danych. Odbyła się również sesja plakatowa, na której zaprezentowano 14 plakatów. Obradom w poszczególnych sesjach konferencji przewodniczyli profesorowie: Józef Pociecha, Eugeniusz Gatnar, Tadeusz Trzaskalik, Krzysztof Jajuga, Marek Walesiak, Barbara Pawełek, Feliks Wysocki, Ewa Roszkowska, Andrzej Sokołowski, Andrzej Bąk, Tadeusz Kufel, Mirosław Krzyśko, Krzysztof Najman, Małgorzata Rószkiewicz, Mirosław Szreder.

Teksty 25 recenzowanych artykułów naukowych stanowią zawartość prezentowanej publikacji z serii „Taksonomia" nr 26. Pozostałe recenzowane artykuły znajdują się w „Taksonomii" nr 27.

W pierwszym dniu konferencji odbyło się posiedzenie członków Sekcji Klasyfikacji i Analizy Danych Polskiego Towarzystwa Statystycznego, któremu przewodniczył prof. dr hab. Józef Pociecha. Ustalono plan przebiegu zebrania obejmujący następujące punkty:

A. Sprawozdanie z działalności Sekcji Klasyfikacji i Analizy Danych PTS.
B. Informacje dotyczące planowanych konferencji krajowych i zagranicznych.
C. Organizacja konferencji SKAD PTS w latach 2016 i 2017.
D. Wybór przedstawiciela Rady Sekcji SKAD PTS do IFCS.
E. Dyskusja nad kierunkami rozwoju działalności Sekcji.

Prof. dr hab. Józef Pociecha otworzył posiedzenie Sekcji SKAD PTS. Sprawozdanie z działalności Sekcji Klasyfikacji i Analizy Danych PTS przedstawiła sekretarz naukowy Sekcji dr hab. Barbara Pawełek, prof. nadzw. UEK. Poinformowała, że obecnie Sekcja liczy 231 członków. Przypomniała, że na stronie internetowej Sekcji znajdują się regulamin, a także deklaracja członkowska. Poinformowała, że zostały opublikowane zeszyty z serii „Taksonomia" nr 24 i 25 (PN UE we Wrocławiu nr 384 i 385). W „Przeglądzie Statystycznym" (zeszyt 4/2014) ukazało się sprawozdanie z ubiegłorocznej konferencji SKAD, która odbyła się w Międzyzdrojach, w dniach 8–10 września 2014 r. Prof. Barbara Pawełek przedstawiła także informacje dotyczące działalności międzynarodowej oraz udziału w ważnych konferencjach członków i sympatyków SKAD.

W konferencji Międzynarodowego Stowarzyszenia Towarzystw Klasyfikacyjnych (IFCS – International Federation of Classification Societies) w dniach 6–8 lipca 2015 r. w Bolonii, zorganizowanej przez Università di Bologna, udział wzięło 19 osób z Polski (w tym 17 członków Sekcji), które wygłosiły 15 referatów (wkład członków SKAD – 79,0%). Ponadto prof. Józef Pociecha był członkiem Komitetu Naukowego Konferencji z ramienia SKAD, członkiem Międzynarodowego Komitetu Nagród IFCS oraz organizatorem i przewodniczącym sesji nt. „Classification models for forecasting of economic processes".

W konferencji „European Conference on Data Analysis" (Colchester, 2–4 września 2015 r.) zorganizowanej przez The German Classification Society (GfKl) we współpracy z The British Classification Society (BCS) i Sekcją Klasyfikacji i Analizy Danych PTS (SKAD) udział wzięło 18 osób z Polski (w tym 14 członków Sekcji), które wygłosiły 15 referatów (wkład członków SKAD – 66,0%). Ponadto profesorowie Krzysztof Jajuga oraz Józef Pociecha byli członkami Komitetu Naukowego konferencji, prof. Andrzej Dudek został poproszony przez organizatorów o przygotowanie referatu i wygłoszenie na Sesji Plenarnej „Cluster analysis in XXI century, new methods and tendencies", prof. Krzysztof Jajuga był przewodniczącym sesji plenarnej, przewodniczącym sesji nt. „Finance and economics II" oraz organizatorem i przewodniczącym sesji nt. „Data analysis in finance", prof. Józef Pociecha był organizatorem i przewodniczącym sesji nt. „Outliers in classification procedures – theory and practice", prof. Andrzej Dudek był przewodniczącym sesji nt. „Machine learning and knowledge discovery II".

Kolejny punkt posiedzenia Sekcji obejmował zapowiedzi najbliższych konferencji krajowych i zagranicznych, których tematyka jest zgodna z profilem Sekcji. Prof. dr hab. Józef Pociecha poinformował o dwóch wybranych konferencjach krajowych (były to XXXIV Konferencja Naukowa „Multivariate Statistical Analysis MSA 2015", Łódź, 16–18 listopada 2015 r. i X Międzynarodowa Konferencja Naukowa im. Profesora Aleksandra Zeliasia nt. „Modelowanie i prognozowanie zjawisk społeczno-gospodarczych", Zakopane, 10–13 maja 2016 r.) oraz o trzech wybranych konferencjach zagranicznych. Konferencja „European Conference on Data Analysis" odbędzie się na Uniwersytecie Ekonomicznym we Wrocławiu w dniach 26–28 września 2017 r. W przeddzień tej konferencji, tj. 25.09.2017 r., odbędzie się Niemiecko-Polskie Sympozjum nt. „Analizy danych i jej zastosowań GPSDAA 2017". Następna konferencja Międzynarodowego Stowarzyszenia Towarzystw Klasyfikacyjnych (IFCS) odbędzie się w 2017 r. w Tokio. W 2019 r. Niemiecko-Polskie Sympozjum nt. „Analizy danych i jej zastosowań GPSDAA 2019" organizuje prof. Andreas Geyer-Schultz w Karlsruhe.

W następnym punkcie posiedzenia podjęto kwestię organizacji kolejnych konferencji SKAD. SKAD 2016 zorganizuje Katedra Metod Statystycznych Wydziału Ekonomiczno-Socjologicznego Uniwersytetu Łódzkiego.

W kolejnej części zebrania dokonano wyboru przedstawiciela Rady Sekcji SKAD PTS do IFCS na kadencję 2016–2019. Powołano Komisję Skrutacyjną, której przewodniczącym został prof. Tadeusz Kufel, a członkami dr hab. Iwona Konarzewska i dr Dominik Rozkrut. Profesor Józef Pociecha poprosił zebranych o proponowanie kandydatur zgłaszając jednocześnie prof. Andrzeja Sokołowskiego. Wobec braku następnych kandydatur listę zamknięto. Komisja Skrutacyjna przeprowadziła głosowanie tajne. W głosowaniu uczestniczyło 41 członków Sekcji. Profesor Andrzej Sokołowski został przedstawicielem Rady Sekcji SKAD PTS do

IFCS na kadencję 2016–2019, uzyskując następujący wynik: 39 głosów na „tak", 1 głos na „nie", 1 głos był nieważny.

W ostatnim punkcie zebrania dyskutowano nad kierunkami rozwoju działalności Sekcji obejmującymi następujące problemy: udział w międzynarodowym ruchu naukowym (wspólne granty, publikacje), umiędzynarodowienie konferencji SKAD (uczestnicy zagraniczni, dwujęzyczność konferencji), wydawanie własnego czasopisma.

Profesor Józef Pociecha zamknął posiedzenie Sekcji SKAD.

Krzysztof Jajuga, Marek Walesiak

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 426 • 2016

Taksonomia 26 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania e-ISSN 2392-0041

Marek Lubicz
Politechnika Wrocławska
e-mail: marek.lubicz@pwr.edu.pl

PROBLEMY DOBORU ZMIENNYCH OBJAŚNIAJĄCYCH W KLASYFIKACJI DANYCH MEDYCZNYCH

FEATURE SELECTION AND ITS IMPACT ON CLASSIFIER EFFECTIVENESS – CASE STUDY FOR MEDICAL DATA

DOI: 10.15611/pn.2016.426.09

Streszczenie: Artykuł dotyczy zagadnienia doboru zmiennych objaśniających w modelach klasyfikacji obiektów z uczeniem dla danych niezrównoważonych. Zbadano wpływ metod doboru zmiennych w połączeniu z metodami wstępnego przetwarzania danych na jakość klasyfikacji dla wybranych grup klasyfikatorów. W komputerowej analizie porównawczej wykorzystano dane z Wrocławskiego Ośrodka Torakochirurgii. Obliczenia przeprowadzono w pakiecie uczenia maszynowego WEKA.

Słowa kluczowe: uczenie maszynowe, klasyfikacja, selekcja zmiennych, dane niedoskonałe, ryzyko operacyjne, zarządzanie szpitalem.

Summary: The article concerns the problems of feature selection in supervised classification models for incomplete and imbalanced data. We compared the results of the application of feature selection methods implemented in the WEKA and STATISTICA machine learning environments. The impact of particular feature selection methods applied in conjunction with the pre-processing methods of missing and imbalanced date on the effectiveness and efficiency of selected single and ensemble classifiers was analyzed. The comparative analysis used updated data from the Wrocław Centre for Thoracic Surgery, on patients operated between 2006 and 2013 due to lung cancer. Sets of rules relating to hospital clinical and managerial decisions have been extracted for selected feature selection and classification methods, and for data relating to preoperative risk assessment.

Keywords: machine learning, classification, feature selection, imperfect data, surgical risk, hospital management.

1. Wstęp

Jakość i efektywność klasyfikacji nadzorowanej (ze zbiorem uczącym) zależą w zasadniczym stopniu od jakości danych źródłowych. Dostępne dane źródłowe w rzeczywistych problemach decyzyjnych zaliczane są często do kategorii danych niedoskonałych (*imperfect data*; [Hartono, Hashimoto 2007]), określanych też jako dane zaszumione (*noisy data*; [Zhu, Wu 2004; Sáez i in. 2013]). W literaturze przedmiotu z zakresu eksploracji danych i uczenia maszynowego [Witten i in. 2011; Kelleher i in. 2015] wyróżnia się dwa podstawowe rodzaje wad danych źródłowych: dotyczące oznaczenia klasy obiektów i dotyczące wartości cech charakteryzujących obiekty. Wśród wad pierwszego rodzaju wymienia się m.in.:

- niepewność oznaczenia klasy w zbiorze uczącym [Frénay, Verleysen 2014], co może być związane z błędami w zbiorze danych lub istotą problemu (np. brak ostrych granic między klasami) i dodatkowo zmieniać się w czasie (np. okresowa aktualizacja długości przeżycia pooperacyjnego),
- niezrównoważenie klas [López i in. 2013]: nadreprezentacja obiektów z jednej klasy, wynikająca z istoty modelowanego zjawiska (np. liczba zgonów po operacji) lub zastosowanej metody budowy zbioru uczącego (doboru próby).

Wady związane z wartościami charakterystyk obiektów mogą dotyczyć np.:

- niekompletności zbiorów danych źródłowych – braków wartości niektórych zmiennych dla pewnych obiektów [García-Laencina i in. 2009],
- błędów w określaniu wartości poszczególnych charakterystyk lub technologii gromadzenia i przechowywania danych [Sáez i in. 2016].

W poprzednich pracach, omówionych w [Lubicz i in. 2014], rozważano metody wstępnej obróbki danych niekompletnych i niezrównoważonych na przykładzie danych medycznych. Zastosowanie tych metod istotnie wpływa na liczność (redukcja obiektów z brakami danych, równoważenie klas metodami próbkowania) lub zawartość (imputacja brakujących danych, redukcja cech z dużą ilością braków i słabą korelacją z etykietą klasy) zbioru uczącego. Jednocześnie brak zastosowania równoważenia klas znacznie zniekształca wyniki klasyfikacji [He, Garcia 2009; Galar i in. 2012; López i in. 2013] i może powodować ignorowanie przez klasyfikator obiektów z klasy mniejszościowej (pozytywnej). W niniejszym artykule rozważono problem doboru zmiennych objaśniających (selekcji zmiennych) do modelu klasyfikacji. Z przyczyn wymienionych powyżej, problem selekcji zmiennych jest rozważany w warunkach konieczności jednoczesnego zastosowania technik obsługi danych niezrównoważonych. Empiryczną podstawą analizy porównawczej są dane medyczne dotyczące przedoperacyjnej oceny ryzyka [Lubicz i in. 2014].

2. Selekcja zmiennych

Selekcja zmiennych dla klasyfikacji, określana m.in. jako dobór zmiennych objaśniających [Gatnar 2008], selekcja cech [Koronacki, Ćwik 2005], selekcja informacji [Sobczak, Malina 1978]; dobór cech [Hand i in. 2005], oznacza proces wykrycia istotnych i odrzucenia nieistotnych lub redundantnych cech klasyfikowanych obiektów, kończący się wyznaczeniem podzbioru cech, który spełnia określone kryteria jakości, właściwe dla rozpatrywanego problemu klasyfikacji [Guyon, Elisseeff 2006]. Selekcję wykonuje się m.in. w celu zmniejszenia wymiarowości zadania przy dużej liczbie zmiennych objaśniających, a także w sytuacjach, w których utrudnione lub kosztowne jest ustalenie wartości niektórych zmiennych dla wszystkich klasyfikowanych obiektów. Z formalnego punktu widzenia potrzebę selekcji zmiennych uzasadnia się [Gatnar 2008; Larose 2008] m.in. dążeniem do uniknięcia nadmiernego dopasowania się modelu do danych i ułatwieniem interpretacji wyników klasyfikacji.

Metody selekcji zmiennych można klasyfikować według różnych perspektyw. Najprostszy podział uwzględnia [Bolon-Canedo i in. 2015] metody zwracające indywidualne oceny ważności poszczególnych cech oraz metody wyznaczające podzbiory cech, oparte na określonych strategiach przeszukiwania przestrzeni rozwiązań. Problem wyznaczenia optymalnego podzbioru cech można sformułować jako złożony problem kombinatoryczny, do którego wykorzystuje się podejścia znane z badań operacyjnych [Bertolazzi i in. 2016], często oparte na technikach inspirowanych naturą (algorytmy genetyczne [Hira, Gillies 2015], PSO [Mangat, Vig 2014]). Inny podział dotyczy podejścia do oceny ważności cech i wykorzystania tych ocen w procesie klasyfikacji. Wyróżnia się zwykle [Duch 2006]: metody wbudowane w mechanizmy klasyfikacji danych, np. w modele sieci neuronowych (*embedded methods*); metody związane z konkretnymi algorytmami klasyfikacji, które wykorzystywane są przy ocenie podzbiorów cech (metody opakowane – *wrapper methods*) oraz metody niezależne od algorytmów klasyfikacji, filtrujące zbiór cech z zastosowaniem określonej miary ważności, np. opartej na korelacji lub zawartości informacyjnej. I. Guyon i A. Elisseeff [2006] omawiają także inne charakterystyki metod selekcji cech, za szczególnie istotną uznając uwzględnienie w ocenie wzajemnych zależności między cechami (ocena indywidualna, ocena wektorowa).

Kierując się wskazówkami literatury przedmiotu (m.in. [Vergara, Estevez 2014; Bolon-Canedo i in. 2013, 2015; Dessi, Pes 2015]) do analizy porównawczej wybrano wymienione w tab. 1 i 2 podstawowe, najczęściej stosowane metody selekcji cech w zagadnieniach uczenia maszynowego. Nazwy metod selekcji podano w konwencji stosowanej w pakiecie WEKA, przyjętym jako środowisko komputerowej analizy porównawczej w cyklu prac badawczych [Lubicz i in. 2014]. Podejścia z pierwszej grupy umożliwiają uporządkowanie zbioru cech względem ustalonego kryterium (*attribute evaluator*); technikę wyznaczenia optymal-

Tabela 1. Metody selekcji cech wykorzystane w analizie porównawczej: ocena indywidualna cech

Metody indywidualnej oceny cech	Źródło
Kryteria statystyczne: Chi-squared	[Liu, Setiono 1996]
Kryteria informacyjne: Information gain; Gain ratio; Symmetrical Uncertainty; FilteredAttribute	[Quinlan 1986]
Kryteria odległości (*n* najbliższych sąsiadów) ReliefF	[Kononenko 1994]
ClassifierAttributeEvaluation oparte na klasyfikatorach: Naive Bayes, JRip, Random Forest, AdaBoostM1 [J48]	[Witten i in. 2011]

Źródło: opracowanie własne.

Tabela 2. Metody selekcji cech wykorzystane w analizie porównawczej: ocena podzbioru cech

Metody oceny podzbioru cech (*attribute evaluator*)	Źródło	Techniki przeszukiwania przestrzeni podzbiorów cech (*search method*)	Źródło
CFS (Correlation-based feature selection)	[Hall 1999]	RankSearch	[Hall, Holmes 2003]
		algorytmy zachłanne (BestFirst, GreedyStepwise, LinearForward-Selection)	pakiet WEKA
Consistency-based Filter	[Dash, Liu 2003]		
ClassifierSubsetEvaluation oparte na: JRip (Ripper), Random Forest	pakiet WEKA	EvolutionarySearch	
		PSOSearch	[García-Nieto i in. 2009]
		GeneticSearch	[Goldberg 1989]
WrapperSubsetEvaluation oparte na: JRip (Ripper), J48	[Kohavi, John 1997]	ScatterSearch	[Lopez i in. 2006]

Źródło: opracowanie własne.

nego uporządkowania (*search method*) określono w WEKA jako Ranker. W podejściach drugiej grupy określa się metodę oceny podzbioru cech oraz technikę przeszukiwania przestrzeni podzbiorów cech dla danej metody oceny (dla każdej metody oceny cech wymienionej w tab. 2, w analizie porównawczej zastosowano alternatywnie wszystkie wymienione w tej tabeli techniki wyznaczania optymalnych podzbiorów).

3. Założenia badawcze i dane źródłowe

Podstawowy, utylitarny problem badawczy dotyczył predykcji ryzyka operacyjnego w torakochirurgii. Problem ten sformułowano jako zadanie klasyfikacji obiektów, którymi są operowani pacjenci, do jednej z dwóch klas, zdefiniowanych odrębnie dla każdego przyjętego kryterium oceny: R1 (R30) – wystąpienie zgonu pacjenta w ciągu 1 roku (30 dni) po operacji; CCIF – wystąpienie komplikacji medycznych podczas hospitalizacji pooperacyjnej; LOSp – przedłużenie hospitalizacji pooperacyjnej (ponad 10 dni). W badaniach wykorzystano zaktualizowane dane przedoperacyjne z Wrocławskiego Ośrodka Torakochirurgii, dotyczące 1711 pacjentów operowanych w latach 2006–2013 z powodu raka płuca. Aktualizacja dotyczyła danych opisanych

w [Lubicz i in. 2014] i obejmowała: dołączenie danych o pacjentach operowanych w roku 2013, weryfikację i uzupełnienie wartości parametrów oddechowych FVC, FEV1 oraz długości przeżycia pooperacyjnego. Dla wymienionych wyżej zmiennych objaśnianych (R1, R30, CCIF, LOSpop) wskaźniki niezrównoważenia, obliczane jako iloraz liczebności klasy większościowej do mniejszościowej, wynoszą odpowiednio 4,7; 34,7; 3,2; 8,8. Charakterystykę zmiennych objaśniających przedstawiono w tabeli 3. Po wyeliminowaniu cech niewykazujących zmienności w dostępnym zbiorze danych, dalszej analizie poddawano 84 cechy (liczby w nawiasie w drugiej kolumnie tab. 3). Ponieważ wykorzystywany w badaniach źródłowy system komputerowy automatycznie nadaje domyślne wartości zerowe zmiennym objaśniającym, uniemożliwiając rozróżnienie sytuacji braku danych od faktycznych braków objawu i rozpoznania, przeprowadzono dodatkową weryfikację danych o rozpoznaniach nowotworowych i chorobach współistniejących.

Tabela 3. Charakterystyka danych źródłowych, wykorzystywanych w analizie porównawczej

Grupa danych	Liczba cech	Opis	% braki	% $\neq 0$
Osoba	3 (3)	płeć, wiek (dzień oceny przedoperacyjnej), strona (L/P)	0	0
Rozpoznania nowotworowe	51 (29)	kody rozpoznań wg ICD10, istnienie przerzutów (binarne)	0	1
Inne rozpoznania	34 (33)	uznawane za istotne w rokowaniu (binarne)	0	5
Ocena kliniczna	8 (8)	wartości klinicznego TNM i ocena wg skali Zubroda (symboliczne); parametry oddechowe FVC, FEV1 (wartość, % normy)	0–57	35–100
Objawy przed operacją	7 (7)	ból; krwioplucie; duszność; kaszel; pogorszenie sprawności; powiększone węzły chłonne N1, N2 (binarne)	0	5–67
Czynniki ryzyka	3 (3)	życie w środowisku przemysłowym (binarne), palenie tytoniu (liczba lat)	0–55	5–45
Chemioterapia	3 (1)	chemioterapia przedoperacyjna (binarne)	0	6

Źródło: opracowanie własne.

Wynikiem każdej metody selekcji cech jest ograniczony zestaw cech lub ranking ważności cech, których potencjalną trafność można obiektywnie ocenić jedynie w procesie klasyfikacji. Plan badań obejmował zatem:

- etap 1A: właściwej selekcji cech metodami wymienionymi w tab. 1 i 2; wynikiem tego etapu były uporządkowane (dla oceny indywidualnej) lub zredukowane nieuporządkowane (dla oceny podzbiorów cech) zbiory cech, odrębnie dla każdego kryterium klasyfikacji pacjentów,
- etap 1B: ze względu na wysoki stopień niezrównoważenia klas powtórzono etap selekcji po uprzednim zrównoważeniu rozkładu klas często cytowaną w literaturze przedmiotu metodą SMOTE [Chawla i in. 2002],
- etap 2: oceny wyników selekcji z wykorzystaniem wybranych klasyfikatorów, o potwierdzonej w poprzednich etapach prac badawczych skuteczności klasyfi-

kacji danych niezrównoważonych (omówienie w [Lubicz i in. 2014]): regułowych (JRIP, PART), drzewowych (Random Forest RF), wielomodelowych (AdaBoostM1 na bazie Random Forest lub J48) i naiwnego klasyfikatora Bayesa (NB).

Kierując się wskazówkami literatury przedmiotu ([He, Garcia 2009; Galar i in. 2012; Sokolova, Lapalme 2009; Kelleher i in. 2015]) do oceny jakości klasyfikacji w sytuacji niezrównoważenia klas przyjęto, podobnie jak w poprzednich etapach prac badawczych ([Lubicz i in. 2014]), następujące wskaźniki jakości klasyfikacji:

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

κ – statystykę Kappa Cohena,

oraz szczególnie istotne przy klasyfikacji danych niezrównoważonych:

AUC = 0,5×(1 + TPR − FPR) – wskaźnik reprezentujący pole powierzchni pod krzywą ROC (Receiver Operating Characteristic),

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

gdzie: TPR = TP/(TP + FN) – czułość klasyfikacji, TNR = TN/(FP + TN) – swoistość klasyfikacji, FPR = FP/(FP + TN) – odsetek błędów I rodzaju; TP, TN (FP, FN) – liczby prawidłowo (błędnie) sklasyfikowanych pacjentów z klasy pozytywnej i negatywnej; NN – łączna liczba klasyfikowanych pacjentów.

4. Wyniki analizy porównawczej i wnioski

Pierwsze etapy badań obejmowały utworzenie 58 plików danych eksperymentalnych (uporządkowane lub zredukowane zestawy cech), będących wynikiem zastosowania metod selekcji cech wymienionych w tab. 1 i 2, odrębnie dla każdego kryterium oceny (etap 1A) oraz analogicznych plików po dodatkowym zastosowaniu równoważenia klas metodą SMOTE (etap 1B). Następnie w etapie 2 wykonano w programie WEKA klasyfikację danych eksperymentalnych dla każdego rozpatrywanego klasyfikatora, przy założeniu 10-krotnej walidacji krzyżowej oraz 10 powtórzeń każdego przebiegu. Fragment wyników obliczeń, odpowiadających najlepszym rozwiązaniom dla danych zrównoważonych, dla poszczególnych wskaźników jakości i kryteriów oceny, przedstawiono w tab. 4 i 5. Dla porównania w tabelach zawarto też wyniki klasyfikacji na danych pierwotnych, bez zastosowania selekcji cech.

Badania potwierdziły przede wszystkim konieczność stosowania równoważenia klas przed wykonaniem klasyfikacji dla danych o dużym stopniu niezrównoważenia, niezależnie od tego, czy stosuje się jakąkolwiek metodę selekcji cech. Jest to jasno widoczne dla zmiennych objaśnianych R30 i LOSp (tab. 4), dla których jakość klasyfikacji pacjentów z klasy mniejszościowej (κ, GM) jest bliska zera, a wysoki poziom dokładności (ACC) jest równy wskaźnikowi udziału pacjentów z klasy większościowej. Zastosowanie równoważenia klas (tab. 5) zdecydowanie poprawia jakość klasyfikacji przy nieznacznym pogorszeniu jej dokładności.

Tabela 4. Fragment zestawienia wyników klasyfikacji z wykorzystaniem metod selekcji cech – wyniki dla danych bez zrównoważenia klas

Kryterium	Metoda selekcji cech	L. cech	Klasyfi-kator	ACC	κ	AUC	GM
R1	ConsistencySubsetEval;BestFirst	36	AB[J48]	77,5	**0,099**	0,591	0,401
R1	**bez wstępnej selekcji cech**	84	AB[RF]	76,0	0,080	0,579	**0,402**
R1	ConsistencySubsetEval;BestFirst	36	RF	82,0	0,087	0,638	0,272
R1	InfoGain lub ReliefAttributeEval; Ranker	84	RF	**82,6**	0,058	**0,645**	0,180
R30	ClassifierSubsetEval[JRip];RankSearch	1	AB[J48]	**97,2**	0,000	0,499	0,000
R30	**bez wstępnej selekcji cech**	84	AB[RF]	96,2	0,002	0,543	0,018
R30	ClassifierSubsetEval[RF];RankSearch	81	RF	97,1	0,001	0,585	0,000
R30	ClassifierSubsetEval[RF];GeneticSearch	38	RF	97,1	0,002	**0,633**	0,000
CCIF	CfsSubsetEval;BestFirst	10	NB	75,0	0,082	0,620	0,316
CCIF	CfsSubsetEval;ScatterSearch	10	NB	75,0	0,082	0,620	0,316
CCIF	ClassifierAttributeEval[AB[J48]];Ranker	84	RF	70,4	0,112	0,598	**0,463**
CCIF	**bez wstępnej selekcji cech**	84	NB	70,4	0,112	0,600	**0,463**
LOSp	ConsistencySubsetEval;BestFirst	33	PART	86,7	0,018	0,523	0,178
LOSp	ConsistencySubsetEval;GreedyStepwise	1	PART	89,8	0,000	0,500	0,000
LOSp	**bez wstępnej selekcji cech**	84	NB	85,8	0,011	0,578	**0,193**
LOSp	SymmetricalUncertAttributeEval;Ranker	84	RF	89,8	0,000	0,550	0,000

Źródło: opracowanie własne.

Tabela 5. Fragment zestawienia wyników klasyfikacji z wykorzystaniem metod selekcji cech – wyniki dla danych po zrównoważeniu klas metodą SMOTE [Chawla i in. 2002]

Kryterium	Metoda selekcji cech	L. cech	Klasyfi-kator	ACC	κ	AUC	GM
R1	ConsistencySubsetEval;BestFirst	30	AB[J48]	78,1	0,460	0,773	**0,707**
R1	**bez wstępnej selekcji cech**	84	AB[RF]	80,7	0,497	0,783	0,700
R1	ConsistencySubsetEval;BestFirst	30	RF	82,8	**0,528**	**0,808**	0,686
R1	InfoGain lub ReliefAttributeEval; Ranker	84	RF	83,0	**0,528**	0,805	0,682
R30	ClassifierSubsetEval[JRip];RankSearch	80	AB[J48]	95,6	0,509	0,764	**0,669**
R30	**bez wstępnej selekcji cech**	84	AB[RF]	95,8	0,509	0,765	0,657
R30	ClassifierSubsetEval[RF];RankSearch	81	RF	**96,9**	**0,579**	0,806	0,656
R30	ClassifierSubsetEval[RF];GeneticSearch	38	RF	96,3	0,526	**0,823**	0,641
CCIF	CfsSubsetEval;BestFirst	10	NB	72,7	0,425	0,773	**0,709**
CCIF	CfsSubsetEval;ScatterSearch	10	NB	72,7	0,425	0,773	**0,709**
CCIF	ClassifierAttributeEval[AB[J48]];Ranker	84	RF	**77,9**	**0,499**	**0,776**	0,701
CCIF	**bez wstępnej selekcji cech**	84	NB	71,0	0,398	0,768	0,700
LOSp	ConsistencySubsetEval;BestFirst	41	PART	84,6	0,457	0,741	**0,683**
LOSp	ConsistencySubsetEval;GreedyStepwise	41	PART	84,6	0,457	0,741	**0,683**
LOSp	**bez wstępnej selekcji cech**	84	NB	82,6	0,418	0,740	0,680
LOSp	SymmetricalUncertAttributeEval;Ranker	84	RF	87,0	0,491	**0,758**	0,664

Źródło: opracowanie własne.

Badania dla analizowanego zbioru danych medycznych nie wykazały zasadniczej zmiany jakości klasyfikacji po zastosowaniu selekcji cech do danych zrównoważonych (porównywalne wartości wskaźników κ, GM; tab. 5); dla danych niezrównoważonych otrzymano nawet w większości przypadków zmniejszenie wartości wskaźnika GM. Jednocześnie, w odróżnieniu od spostrzeżeń w literaturze przedmiotu [Bolón-Canedo i in. 2013; Dessi Pes 2015; Vergara, Estevez 2014; Bertolazzi i in. 2016], z wyjątkiem zmiennej objaśnianej R30, analiza nie wykazała znacznej przewagi metod selekcji cech zależnych od klasyfikatorów nad metodami filtrującymi zbiór cech niezależnie od klasyfikatora (porównywalne wyniki klasyfikacji i czasy obliczeń).

W większości przypadków najwyższą jakość klasyfikacji otrzymywano dla metod wyznaczających optymalny podzbiór cech, przy jednoczesnym zastosowaniu klasyfikatora Random Forest lub (zmienne objaśniane R1, R30) klasyfikatora AdaBoostM1 z klasyfikatorem bazowym J48 lub Random Forest. Stosunkowo wysokie wartości wskaźników jakości (AUC, GM) otrzymywano dla selekcji na podstawie analizy wewnętrznej niezgodności zbioru cech (ConsistencySubsetEval), szczególnie z zastosowaniem metaheurystyk (np. algorytmów genetycznych lub innych algorytmów ewolucyjnych) do znalezienia optymalnego podzbioru cech. Najgorsze wyniki (κ, GM rzędu 0,0–0,1) otrzymano dla metod opartych na korelacji (CfsSubsetEval) i jednoczesnego zastosowania standardowych klasyfikatorów NB, JRip i PART.

Analiza porównawcza wykazała zróżnicowanie wyników dla poszczególnych kryteriów oceny: dla zmiennych objaśnianych R1, R30, CCIF, LOSp średnie dokładności klasyfikacji wynosiły odpowiednio 79, 96, 74 i 84% przy zbliżonych średnich wartościach pozostałych wskaźników dla większości rozwiązań (kappa 0,4–0,5; AUC 0,74–0,78; GM 0,68–0,71). Może to świadczyć o konieczności uwzględnienia w dalszych analizach innych, dotychczas niedostępnych danych medycznych.

Podsumowując, należy stwierdzić, że przeprowadzone badania nie potwierdziły wskazywanej w literaturze przedmiotu możliwości zwiększenia dokładności klasyfikacji dla analizowanego zbioru danych medycznych, odznaczającego się wysokim stopniem niezrównoważenia klas. W takich przypadkach postępowaniem z wyboru wydaje się być wyeliminowanie cech redundantnych oraz wykazujących niewielką zmienność z wykorzystaniem jednej z metod selekcji cech opartej na doborze podzbioru, wstępne zrównoważenie klas oraz dobór odpowiednio silnego klasyfikatora i próba kalibracji jego parametrów (zwiększenie liczby drzew decyzyjnych w klasyfikatorze Random Forest i dobór klasyfikatora bazowego dla klasyfikatora AdaBoostM1 miały większy pozytywny wpływ na wyniki klasyfikacji niż zastosowanie jednej z metod selekcji cech). Istotną kwestią może być także próba zwiększenia zawartości informacyjnej źródłowego zbioru cech; w przypadku analizowanego zbioru danych medycznych: o cechy dotyczące wyników badań diagnostycznych.

Literatura

Bertolazzi P., Felici G., Festa P., Fiscon G., Weitschek E., 2016, *Integer programming models for feature selection: New extensions and a randomized solution algorithm*, European Journal of Operations Research, vol. 250, no. 2, s. 389–399.

Bolón-Canedo V., Sánchez-Maroño N., Alonso-Betanzos A., 2013, *A review of feature selection methods on synthetic data*, Knowledge and Information Systems, vol. 34, no. 3, s. 483–519.

Bolón-Canedo V., Sánchez-Maroño N., Alonso-Betanzos A., 2015, *Recent advances and emerging challenges of feature selection in the context of big data*, Knowledge-Based Systems, vol. 86, s. 33–45.

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.

Dash M., Liu H., 2003, *Consistency-based search in feature selection*, Artificial Intelligence, vol. 151, no. 1-2, s. 155–176.

Dessi N., Pes B., 2015, *Similarity of feature selection methods: An empirical study across data intensive classification tasks*, Expert Systems with Applications, vol. 42, no. 10, s. 4632–4642.

Duch W., 2006, *Filter Methods*, [w:] I. Guyon, S. Gunn, M. Nikravesh, L.A. Zadeh (red.), *Feature Extraction: Foundations and Applications*, Springer, Berlin.

Frénay B., Verleysen M., 2014, *Classification in the presence of label noise: A survey*, IEEE Transactions on Neural Networks and Learning Systems, vol. 25, no. 5, s. 845–869.

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, no. 4, s. 463–484.

García-Laencina P.J., Sancho-Gómez J.L., Figueiras-Vidal A.R., 2009, *Pattern classification with missing data: A review*, Neural Computing and Applications, vol. 19, no. 2, s. 263–282.

García-Nieto J.M., Alba E., Jourdan L., Talbi E.-G., 2009, *Sensitivity and specificity based multiobjective approach for feature selection: Application to cancer diagnosis*, Information Processing Letters, vol. 109, no. 16, s. 887–896.

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

Goldberg D.E., 1989, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, Boston.

Guyon I., Elisseeff A., 2006, *An introduction to feature extraction*, [w:] I. Guyon, S. Gunn, M. Nikravesh, L.A. Zadeh (red), *Feature Extraction: Foundations and Applications*, Springer, Berlin.

Hall M.A., 1999, *Correlation-based Feature Subset Selection for Machine Learning*, PhD Thesis, The University of Waikato.

Hall M.A., Holmes G., 2003, *Benchmarking attribute selection techniques for discrete class data mining*, IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 3, s. 1437–1447.

Hand D., Mannila H., Smyth P., 2005, *Eksploracja danych*, WNT, Warszawa.

Hartono P., Hashimoto S., 2007, *Learning from imperfect data*, Applied Soft Computing, vol. 7, no. 1, s. 353–363.

He H., Garcia A., 2009, *Learning from Imbalanced Data*, IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, s. 1263–1284.

Hira Z.M., Gillies D.F., 2015, *A review of feature selection and feature extraction methods applied on microarray data*, Advances in Bioinformatics, vol. 2015, article ID 198363.

Kelleher J.D., Mac Namee B., D'Arcy A., 2015, *Fundamentals of Machine Learning for Predictive Data Analytics*, MIT Press.

Kohavi R, John G.H., 1997, *Wrappers for feature subset selection*, Artificial Intelligence, vol. 97, no. 1/2, s. 273–324.

Kononenko I., 1994, *Estimating attributes: Analysis and extensions of Relief*, [w:] European Conference on Machine Learning: ECML-94, Springer, s. 171–182.

Koronacki J., Ćwik J., 2005, *Statystyczne systemy uczące się*, WNT, Warszawa.

Larose D.T., 2008, *Metody i modele eksploracji danych*, Wydawnictwo Naukowe PWN, Warszawa.

Liu H., Setiono R., 1996, *A probabilistic approach to feature selection – a filter solution*, [w:] 13th International Conference on Machine Learning, s. 319–327.

López F.G., Torres M.G., Batista B.M., Moreno Pérez J.A., Moreno-Vega J.M., 2006, *Solving feature subset selection problem by a parallel scatter search*, European Journal of Operational Research, vol. 169, no. 2, s. 477–489.

López V., Fernández A., García S., Palade V., Herrera F., 2013, *An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics*, Information Sciences, vol. 250, s. 113–141.

Lubicz M., Zięba M., Pawełczyk K., Rzechonek A., Marciniak M., Kołodziej J., 2014, *Indukcja reguł dla danych niekompletnych i niezbalansowanych: modele klasyfikatorów i próba ich zastosowania do predykcji ryzyka operacyjnego w torakochirurgii*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 328, s. 146–155.

Mangat V., Vig R., 2014, *Novel associative classifier based on dynamic adaptive PSO: Application to determining candidates for thoracic surgery*, Expert Systems with Applications, vol. 41, no. 18, s. 8234–8244.

Quinlan J.R., 1986, *Induction of decision trees*, Machine Learning, vol. 1, no. 1, s. 81–106.

Sáez J.A., Galar M., Luengo J., Herrera F., 2013, *Tackling the problem of classification with noisy data using Multiple Classifier Systems: Analysis of the performance and robustness*, Information Sciences, vol. 247, s. 1–20.

Sáez J.A., Luengo J., Herrera F., 2016, *Evaluating the classifier behavior with noisy data considering performance and robustness: The equalized loss of accuracy measure*, Neurocomputing, vol. 176, s. 26–35.

Sobczak W., Malina W., 1978, *Metody selekcji informacji*, WNT, Warszawa.

Sokolova M., Lapalme G., 2009, *A systematic analysis of performance measures for classification tasks*, Information Processing and Management, vol. 45, no. 4, s. 427–437.

Vergara J.R., Estevez P.A., 2014, *A review of feature selection methods based on mutual information*, Neural Computing and Applications, vol. 24, no. 1, s. 175–186.

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

Zhu X.,Wu X., 2004, *Class noise vs. attribute noise: A quantitative study*, Artificial Intelligence Review, vol. 22, no. 3, s. 177–210.