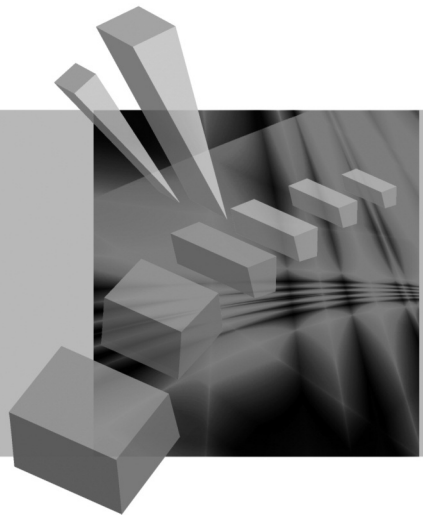
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

242

Taksonomia 19.

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pociecha,
Mirosław Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
oraz w 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

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

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM
Nakład: 320 egz.

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 242
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS 2012

Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Marek Lubicz, Maciej Zięba
Politechnika Wrocławska

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

Jerzy Błaszczyk
Dolnośląskie Centrum Onkologii we Wrocławiu

ANALIZA PORÓWNAWCZA WYBRANYCH TECHNIK EKSPLORACJI DANYCH DO KLASYFIKACJI DANYCH MEDYCZNYCH Z BRAKUJĄCYMI OBSERWACJAMI*

Streszczenie: W praktycznych zadaniach klasyfikacji, np. w analizie danych medycznych, stosunkowo często występuje konieczność wnioskowania na podstawie danych niekompletnych. Celem pracy jest porównanie efektywności wybranych podejść do rozwiązywania problemu klasyfikacji z brakującymi obserwacjami dla wybranych klasyfikatorów prostych i złożonych oraz dla różnych metod transformacji cech z brakującymi obserwacjami. W badaniach zastosowano implementacje technik eksploracji danych w środowisku STATISTICA Data Miner oraz systemie uczenia maszynowego WEKA. Jako dane do klasyfikacji wykorzystano bazę danych o pacjentach leczonych operacyjnie z powodu raka płuca we Wrocławskim Ośrodku Torakochirurgii w latach 2000-2011.

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

1. Wstęp

W artykule rozważono, występujące w wielu zastosowaniach analizy danych, zagadnienie wnioskowania na podstawie niekompletnych danych statystycznych. Porównano efektywność kilku podejść do rozwiązania problemu klasyfikacji z brakującymi obserwacjami, uwzględniając różne klasyfikatory i metody uzupełniania

* Praca naukowa finansowana ze środków budżetowych na naukę w latach 2010-2012 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".

braków danych. Obliczenia, np. wybranych zbiorów danych medycznych, przeprowadzono w środowisku Statistica Data Miner oraz w systemie uczenia maszynowego WEKA.

2. Klasyfikacja z brakującymi obserwacjami

W praktycznych zadaniach klasyfikacji danych medycznych często występuje konieczność wnioskowania na podstawie danych niekompletnych. Niekompletność, rozumiana jako brak znajomości wartości niektórych cech klasyfikowanych obiektów, jest jedną z form niepewności danych, wykorzystywanych do wnioskowania w konkretnym środowisku decyzyjnym. Do formalizacji wnioskowania w sytuacji niepewności stosuje się różnorodne podejścia, łączące probabilistykę, miękką matematykę stosowaną, inteligencję obliczeniową czy techniki eksploracji danych. Do walidacji podejść formalnych z konieczności przyjmuje się porządkujące założenia o charakterze analizowanego zjawiska (np. dotyczące parametrów modelu statystycznego) lub korzysta ze standardowych zbiorów danych uczących, co nie prowadzi do ułatwienia w podejmowaniu decyzji w rzeczywistym środowisku, w którym upraszczające założenia formalne lub odsetek poprawnej klasyfikacji rzędu 80% są nieakceptowalne. Pogląd ten wydają się potwierdzać wyniki Misztal [2011], w których niezależnie od zastosowanych formalizmów błędy klasyfikacji dla danych rzeczywistych są istotnie większe od błędów dla standardowych zbiorów danych z repozytorium UCI Machine Learning Repository [mlr.cs.umass.edu/ml/].

Zadanie klasyfikacji, w którym przypisuje się obiektowi, opisanemu przez wektor cech, jeden numer klasy, jest zwykle realizowane w dwóch etapach: uczenia z wykorzystaniem zbioru obiektów z nadanymi (np. przez ekspertów) numerami klas i właściwej klasyfikacji obiektów nieoznaczonych numerem klasy. Wektory wartości cech obiektów, zarówno w zbiorze uczącym, jak i przy właściwej klasyfikacji, mogą zawierać braki wynikające z różnych przyczyn (np. błędny sposób gromadzenia danych, brak wyników niektórych badań). Wyróżnia się: nielosowy brak danych, w których nie jest możliwe wnioskowanie o brakującej obserwacji jedynie na podstawie pozostałych danych w zbiorze uczącym, oraz całkowicie lub częściowo losowy brak danych, w których jest uzasadnione wykorzystanie dostępnych danych do wnioskowania o danych brakujących. W pierwszym przypadku stosuje się specyficzne problemowo-zorientowane podejścia, dla drugiego przypadku opracowano podejścia formalne, dotyczące z reguły sytuacji, w których rozkład braków nie jest związany z przynależnością do klasy. Łącznie z brakiem obserwacji mogą wystąpić inne anomalie domniemanych warunków klasyfikacji [Zięba 2011]: niezrównoważony podział na klasy (większość obiektów należy do klasy dominującej) lub niezrównoważone skutki błędnej klasyfikacji (koszt błędnej klasyfikacji do jednej klasy jest znacznie większy od kosztu błędnej klasyfikacji do innych klas).

Metody klasyfikacji z brakującymi obserwacjami można podzielić w zależności od sposobu postępowania z brakami danych oraz od wyboru techniki klasyfikacji na:

(a) eliminację braków danych; wyróżnia się eliminację przypadkami (jednorazową redukcję zbioru uczącego do kompletnych obserwacji) i eliminację parami (każdorazowe usuwanie z obliczeń przypadków z brakami danych dla wykorzystywanych zmiennych),

(b) uzupełnienie brakujących obserwacji (imputację) na podstawie metod statystycznych (jedno- lub wielokrotne zastąpienie średnią, medianą lub wartością najczęściej występującą; estymacja z wykorzystaniem regresji) lub na podstawie technik uczenia maszynowego (np. sieci neuronowe lub algorytm *k*-średnich) i zastosowania standardowych klasyfikatorów,

(c) estymację funkcji gęstości danych wejściowych (łącznie dla obserwacji kompletnych i brakujących) oraz zastosowanie wnioskowania bayesowskiego (np. algorytm EM),

(d) wykorzystanie technik uczenia maszynowego, integrujących klasyfikację i przetwarzanie bez wcześniejszej estymacji brakujących danych; autorzy pracy [Garcia-Laencina i in. 2010] jako najbardziej reprezentatywne wymieniają: podejścia wielomodelowe, drzewa decyzyjne, rozmyte uogólnienia sieci neuronowych, algorytmu kNN i modeli opartych na metodzie wektorów nośnych. Wskazują oni, że dla klasyfikacji z wykorzystaniem danych rzeczywistych nie można wskazać metod najlepszych w sensie dokładności klasyfikacji i w konkretnej dziedzinie problemowej konieczne jest każdorazowo dobranie efektywnych metod.

3. Problemy klasyfikacji i eksploracji danych w torakochirurgii

Techniki klasyfikacji i eksploracji danych są od lat stosowane w analizie danych medycznych [Bellazi, Zupan 2008], w dążeniu do przezwyciężenia ograniczeń podejść klasycznych i zwiększenia efektywności wnioskowania, chociaż czasem z dyskusyjnymi rezultatami [Schwarzer i in. 2000]. Także w wybranej przez autorów jako obszar badawczy dziedzinie chirurgii klatki piersiowej klasyczne modelowanie statystyczne jest uzupełniane złożonymi technikami prognostycznymi [Jefferson i in. 1997; Esteva i in. 2007]. Problem badawczy najczęściej dotyczy analizy przeżycia (po operacji lub całkowitego przeżycia w chorobach onkologicznych), a wyniki klasycznych modeli Coksa i Kaplana-Meiera są porównywane z oszacowaniami ryzyka otrzymanymi przy zastosowaniu różnych podejść parametrycznych (np. SVM), nieparametrycznych (np. kNN) lub złożonych (np. [Santos-Garcia i in. 2004]). Najwygodniejsze do interpretacji klinicznej są drzewa decyzyjne i inne podejścia regułowe, a także modele klasyczne, np. podany przez Berrisforda i in. [2005] model regresji logistycznej, określający ryzyko *p* zgonu w ciągu 30 dni po operacji:

$$p = \exp(\text{logit2}) / (1 + \exp(\text{logit2}))$$

$$\text{logit}_2 = -5{,}8858 + (0{,}0501 \times \text{WIEK}) - (0{,}0218 \times \text{ppoFEV1\%}),$$

przy czym parametry równania wyznaczono na podstawie danych o 5 cechach (wiek, rodzaj operacji, stan pacjenta (ASA), stopień duszności (MRC), ppoFEV1%) dla

1753 operowanych pacjentów. Walidacja analogicznych modeli w środowiskach klinicznych, innych niż te, z których pochodzą dane źródłowe, jest utrudniona przede wszystkim różnorodnością systemów gromadzenia danych klinicznych (i definicjami zmiennych). W większości prac nie analizuje się wpływu jakości dostępnych danych źródłowych na wyniki oszacowania ryzyka operacyjnego (por. [Gallivan 2005]), szczególnie pomija się występowanie niekompletności danych źródłowych. Jedną z nielicznych prac dotyczących ryzyka operacyjnego i uwzględniających braki danych jest [Ferguson i in. 2008], w której zauważa się, że udział brakujących obserwacji wynosi niekiedy 15-75% i proponuje się zastosowanie wielokrotnej imputacji opartej na algorytmie CART przed opracowaniem modelu regresji logistycznej. Jak widać, warunki klasyfikacji na rzeczywistych danych medycznych bywają drastycznie odmienne od sytuacji, w których dane lub ich parametry są generowane podczas eksperymentów komputerowych.

4. Dane źródłowe i założenia analizy porównawczej

Celem badań było porównanie efektywności podejść do rozwiązywania problemu klasyfikacji z brakującymi obserwacjami dla wybranych klasyfikatorów prostych i złożonych oraz dla różnych metod transformacji cech z brakującymi obserwacjami w zastosowaniu do rzeczywistych danych medycznych: danych o pacjentach leczonych operacyjnie z powodu raka płuca we Wrocławskim Ośrodku Torakochirurgii (WTO) w latach 2000-2011. Z perspektywy medycznej pytania badawcze dotyczyły m.in. modelowania ryzyka operacyjnego: przeżycia 30 dni, 1 roku i *n* lat (przedziały jednoroczne) po operacji. Celem perspektywicznym było określenie przesłanek do uniknięcia zabiegów niewnoszących istotnego polepszenia rokowania (ryzyko długookresowe) i dostosowania postępowania do przewidywanego ryzyka operacji (ryzyko krótkookresowe).

Perspektywa ilościowa dotyczyła możliwości określenia efektywnych podejść do klasyfikacji pacjentów dla trzech zmiennych objaśnianych (binarne zmienne Risk30, Risk1Yr; nominalna zmienna PopSur) w sytuacji występowania znacznych braków danych. Pierwszy, najbardziej czasochłonny etap badań obejmował zgromadzenie, integrację, anonimizację i wielokrotną weryfikację danych źródłowych z kilku niekompatybilnych systemów szpitalnych WTO i spoza WTO (Regionalny Rejestr Nowotworów, Narodowy Fundusz Zdrowia). Efektem prac było utworzenie badawczych baz danych o różnej szczegółowości:

– W1: szczegółowe dane o pacjentach z resekcjami płuc z powodu pierwotnego raka płuca (2007-2011, 1203 obiekty, 137 zmiennych objaśniających, 3% braków danych),
– W2: ograniczony zestaw podstawowych danych o wszystkich pacjentach, którym wykonano resekcje płuc w latach 2000-2011 (5599 obiektów, 15 zmiennych objaśniających, średnio 26% braków danych).

W drugim etapie zdefiniowano metody postępowania z brakami obserwacji. Ponieważ badania prowadzono w środowisku Statistica Data Miner oraz systemie uczenia maszynowego WEKA, do analizy porównawczej przyjęto techniki zaimplementowane w tych pakietach: MVE: usunięcie przypadków zawierających brakujące wartości cech; MVR: zastąpienie brakujących danych wartością średnią (zmienne ilościowe) lub najczęściej występującą wartością (zmienne jakościowe); IKN: imputacja brakujących danych z wykorzystaniem algorytmu KNN. Dodatkowo dla środowiska WEKA zaimplementowano procedurę MAA, w której brakujące obserwacje traktowane są jako dodatkowe wartości odpowiedniej cechy oraz procedury IJ4, IJR, INB, IML imputacji brakujących danych z wykorzystaniem algorytmów – odpowiednio – J48 (klasyczny C4.5), JRip (reguły decyzyjne), NB (naiwny algorytm Bayesa), MLP (perceptron wielowarstwowy).

W trzecim etapie dla każdej bazy, każdej zmiennej objaśnianej oraz każdej dostępnej techniki przetwarzania braków danych utworzono pliki eksperymentalne (24) i poddano je przetwarzaniu w modułach Przepisy Data Miner (Statistica) i WEKA Experimenter z wykorzystaniem następujących podstawowych klasyfikatorów (dla środowisku WEKA dodatkowo NB): metoda wektorów nośnych (SVM, SMO), sieci neuronowe (MLP), drzewa decyzyjne (CART i drzewa wzmacniane), las losowy (Random Forest). W każdym przypadku zbudowano 20-procentowe próby testowe, a jako metodę walidacji wybrano 10-częściowy sprawdzian krzyżowy. Analizę porównawczą wyników klasyfikacji dla różnych metod przetwarzania braków i różnych klasyfikatorów przeprowadzono na podstawie następujących wskaźników: TPR – czułość klasyfikacji (TP/(TP+FN)); TNR – swoistość klasyfikacji (TN/(FP+TN)); PPV – dodatnia zdolność predykcyjna (TP/(TP+FP)); ACC – odsetek poprawnych klasyfikacji; JY – statystyka J Youdena (TPR+TNR-1); KAPP – statystyka Kappa Cohena (dla binarnej klasyfikacji TP, TN, FP, FN oznaczają odpowiednio liczby: prawidłowego wykrycia ryzyka, prawidłowego wykrycia braku ryzyka, błędnego wykrycia ryzyka – błąd I rodzaju, błędnego niewykrycia ryzyka – błąd II rodzaju).

W ostatnim etapie wykonano dodatkowe badania porównawcze w środowisku WEKA, przy zastosowaniu algorytmu wzmacniającego AdaBoostM1 porównano efektywność wyboru jako klasyfikatorów bazowych kilku algorytmów (SMO, NB, RF, CART), dla których w etapie 3 otrzymano najlepszą dokładność klasyfikacji.

5. Omówienie wyników badań i wnioski

Wybrane wyniki badań, uporządkowane według malejącej wartości statystyki J Youdena, przedstawiono w tab. 1-4. Zastosowano w nich omówione wyżej oznaczenia zbioru uczącego (W1x, W2x w kolumnie „Dane"), wskaźników jakości klasyfikacji (sześć ostatnich kolumn) i metody przetwarzania brakujących danych (kolumna MB). Trzeci znak symbolu w kolumnie „Dane" oznacza zmienną objaśnianą (3 – ryzyko 30-dniowe Risk30, 1 – ryzyko roczne Risk1Yr, S – ryzyko długookresowe PopSur).

Tabela 1. Porównanie wyników klasyfikacji dla analizowanych metod przetwarzania braków danych

Dane	MB	TPR	TNR	KP	PPV	ACC	JY	Dane	MB	TPR	TNR	KP	PPV	ACC	JY
W13	IKN	0,97	0,08	0,04	0,98	94,9	0,05	W23	MVR	0,99	0,07	0,09	0,98	96,9	0,06
W13	INB	0,97	0,07	0,04	0,98	95,0	0,04	W23	MLP	0,99	0,07	0,08	0,98	96,8	0,06
W13	MAA	0,97	0,07	0,04	0,98	95,0	0,04	W23	INB	0,99	0,07	0,07	0,98	96,5	0,06
W13	MVR	0,97	0,07	0,04	0,98	95,0	0,04	W23	IKN	0,99	0,07	0,09	0,98	97,2	0,06
W13	IJ4	0,97	0,07	0,04	0,98	95,0	0,04	W23	IJR	0,99	0,06	0,07	0,98	96,9	0,06
W13	IJR	0,97	0,07	0,04	0,98	95,0	0,04	W23	MAA	0,99	0,06	0,07	0,98	97,0	0,05
W11	IKN	0,94	0,13	0,08	0,81	77,2	0,07	W21	IKN	0,96	0,29	0,30	0,86	83,5	0,24
W11	MLP	0,94	0,12	0,08	0,80	77,2	0,06	W21	MLP	0,94	0,26	0,24	0,85	81,7	0,20
W11	MAA	0,94	0,12	0,07	0,80	77,0	0,06	W21	INB	0,94	0,22	0,18	0,85	80,7	0,15
W11	INB	0,94	0,12	0,07	0,80	77,1	0,06	W21	IJR	0,95	0,16	0,14	0,84	80,9	0,11
W11	IJ4	0,94	0,12	0,07	0,80	77,2	0,06	W21	MAA	0,94	0,17	0,12	0,84	80,1	0,11
W11	MVR	0,94	0,11	0,07	0,80	77,1	0,06	W21	MVR	0,96	0,13	0,11	0,83	80,8	0,09
W1S	MLP	0,40	0,91	0,26	0,43	43,7	0,31	W2S	IKN	0,41	0,96	0,33	0,62	49,3	0,37
W1S	MVR	0,40	0,90	0,24	0,42	42,6	0,30	W2S	INB	0,26	0,95	0,20	0,40	39,3	0,21
W1S	INB	0,37	0,90	0,20	0,40	39,9	0,27	W2S	IJ4	0,23	0,95	0,18	0,37	38,5	0,18
W1S	IJR	0,36	0,90	0,21	0,40	40,5	0,26	W2S	MVR	0,23	0,95	0,19	0,37	39,4	0,18
W1S	IJ4	0,36	0,90	0,21	0,40	40,3	0,26	W2S	IJR	0,23	0,95	0,17	0,36	37,9	0,18
W1S	IKN	0,35	0,91	0,20	0,43	39,7	0,26	W2S	MAA	0,19	0,96	0,16	0,39	36,9	0,15

Źródło: obliczenia własne.

Porównanie metod przetwarzania brakujących danych (tab. 1) nie wskazuje na zdecydowaną przewagę jednej z metod dla analizowanych zbiorów danych źródłowych. Wyniki klasyfikacji (TPR, PPV, ACC) dla większości zbiorów są porównywalne z lekką przewagą imputacji z wykorzystaniem algorytmu kNN i z najgorszym wynikiem dla eliminacji przypadków z brakami danych (MVE). Jednocześnie zwraca uwagę niska wartość statystyki J Youdena dla analizy ryzyka krótkookresowego (Wx3, Wx1) pomimo stosunkowo dobrej dokładności klasyfikacji (ACC). Odmienna sytuacja występuje dla analizy ryzyka długookresowego (WxS), gdzie znacznie spada dokładność klasyfikacji, głównie w związku ze zmniejszeniem dodatniej zdolności predykcyjnej, ale bardzo poprawia się ujemna zdolność predykcyjna, co wpływa na zwiększenie wartości statystyki J.

Porównanie klasyfikatorów (tab. 2) wskazuje na istotną przewagę najprostszego naiwnego klasyfikatora bayesowskiego dla większości analizowanych zbiorów danych. Powyższe wnioski potwierdza zestawienie najlepszych kombinacji klasyfi-

Tabela 2. Porównanie wyników klasyfikacji dla analizowanych klasyfikatorów

Dane	Klasyfikator	TPR	TNR	KP	PPV	ACC	JY
W13	NaiveBayes	0,91	0,32	0,11	0,98	89,84	0,23
	SMO	0,98	0,10	0,09	0,98	96,09	0,08
	JRip	1,00	0,04	0,05	0,98	97,22	0,03
	SimpleCart	1,00	0,01	0,01	0,97	97,35	0,01
W11	NaiveBayes	0,81	0,46	0,25	0,85	73,70	0,27
	SMO	0,92	0,26	0,21	0,83	77,97	0,18
	MultilayerPerceptron	0,94	0,17	0,13	0,81	77,86	0,11
	RandomForest	0,96	0,11	0,09	0,80	78,53	0,07
W1S	SMO	0,45	0,87	0,25	0,36	43,06	0,32
	SimpleCart	0,39	0,92	0,24	0,45	42,61	0,31
	NaiveBayes	0,43	0,86	0,23	0,34	39,36	0,29
	JRip	0,27	0,96	0,20	0,54	42,38	0,22
W23	JRip	1,00	0,09	0,15	0,98	97,63	0,09
	NaiveBayes	0,97	0,10	0,06	0,98	95,12	0,07
	RandomForest	1,00	0,02	0,02	0,98	97,51	0,01
	SimpleCart	1,00	0,01	0,02	0,98	97,54	0,01
W21	NaiveBayes	0,89	0,33	0,24	0,86	78,77	0,22
	RandomForest	0,95	0,22	0,20	0,84	81,24	0,16
	JRip	0,98	0,10	0,10	0,83	81,60	0,07
	SimpleCart	0,98	0,08	0,08	0,83	81,99	0,07
W2S	NaiveBayes	0,29	0,94	0,23	0,38	38,60	0,22
	RandomForest	0,29	0,92	0,26	0,33	41,77	0,21
	JRip	0,12	0,98	0,10	0,46	36,04	0,11

Źródło: obliczenia własne.

kator + metoda przetwarzania braków danych dla każdego zbioru uczącego i każdej zmiennej objaśniającej (tab. 3). Zestawienie potwierdza też zróżnicowane możliwości predykcyjne dla ryzyka krótko- (dokładność klasyfikacji 75-95%) i długookresowego (najwyższa dokładność klasyfikacji z reguły nie przekracza 50%).

Tabela 3. Najlepsze wyniki klasyfikacji dla poszczególnych zbiorów uczących i zmiennych objaśniających

Dane	MB	Klasyfikator	TPR	TNR	KP	PPV	ACC	JY
W13	IKN	NaiveBayes	0,90	0,41	0,12	0,98	88,4	0,3067
W13	MVR	NaiveBayes	0,91	0,33	0,10	0,98	89,2	0,2366
W11	IKN	NaiveBayes	0,81	0,47	0,27	0,85	74,1	0,2838
W11	MAA	NaiveBayes	0,81	0,46	0,26	0,85	74,0	0,2746
W1S	MLP	SimpleCart	0,52	0,90	0,33	0,46	48,8	0,4183
W1S	MVR	SimpleCart	0,52	0,90	0,31	0,45	47,2	0,4136
W23	MLP	JRip	1,00	0,12	0,19	0,98	97,7	0,1225
W23	INB	NaiveBayes	0,95	0,17	0,08	0,98	93,2	0,1161
W21	IKN	RandomForest	0,97	0,45	0,50	0,89	87,7	0,4237
W21	MLP	NaiveBayes	0,88	0,42	0,31	0,87	79,9	0,2994
W2S	IKN	RandomForest	0,57	0,96	0,54	0,61	63,7	0,5273
W2S	IKN	NaiveBayes	0,39	0,94	0,28	0,46	42,8	0,3309

Źródło: obliczenia własne.

Tabela 4. Efekty zastosowania metody wzmocnienia dla wybranych zbiorów uczących

Dane	MB	Klasyfikator	TPR	TNR	KP	PPV	ACC	JY
W13	MVR	AdaBoostM1.SimpleCart	1,00	0,08	0,11	0,98	97,3	0,0799
W13	MVR	SimpleCart	1,00	0,01	0,01	0,97	97,3	0,0057
W11	IKN	AdaBoostM1.SimpleCart	0,89	0,28	0,18	0,82	75,9	0,1631
W11	IKN	SimpleCart	1,00	0,01	0,00	0,79	79,1	0,0035
W1S	MLP	AdaBoostM1.SMO	0,46	0,87	0,28	0,38	45,1	0,3396
W1S	MLP	SMO	0,46	0,88	0,29	0,38	45,6	0,3368
W1S	MLP	AdaBoostM1.NaiveBayes	0,40	0,89	0,26	0,38	41,1	0,2956
W1S	MLP	NaiveBayes	0,40	0,89	0,26	0,38	41,2	0,2928

Źródło: obliczenia własne.

Dodatkowe obliczenia (etap 4) wykazały, że wskazane może być kontynuowanie badań z innymi modelami klasyfikacji, np. z wykorzystaniem klasyfikacji wzmocnionej (wstępne wyniki w tab. 4), pozwalające w wielu przypadkach na zwiększe-

nie szczególnie ujemnej zdolności predykcyjnej. Próby zastosowania dostępnych w środowisku WEKA klasyfikatorów wielomodelowych (np. Decorate, Random Committe) i metody bagging nie doprowadziły do otrzymania wyników lepszych od zestawionych w powyższych tabelach.

Wyniki analizy na dostępnych zbiorach uczących potwierdziły obserwacje literaturowe [Belazzi, Zupan 2008] o co najmniej porównywalnej z innymi klasyfikatorami zdolności predykcyjnej podejść klasycznych, takich jak naiwny klasyfikator Bayesa i algorytm k-NN zastosowany do imputacji braków danych w zbiorze uczącym. Stosunkowo wysoka dokładność klasyfikacji (75-95% dla Wx1), wyższa od opisywanej w pracach z zakresu modelowania ryzyka operacyjnego, wymaga jednak zestawienia ze specyficznymi cechami zbiorów danych: empiryczne ryzyko (częstość zgonów pooperacyjnych) dla zbiorów W1 i W2 wynosiło odpowiednio 3,2 i 2,4% dla Wx1 raz 20,9 i 16,9% dla Wx2. Opracowanie podejść prognostycznych o efektywności akceptowalnej przez klinicystów wymaga w związku z tym dalszych badań.

Literatura

Bellazi R., Zupan B., *Predictive data mining in clinical medicine: Current issues and guidelines*, „International Journal of Medical Informatics" 2008, vol. 77.

Berrisford R., Brunelli A., Rocco G., Treasure T., Utley M., *The European thoracic surgery database project: modelling the risk of in-hospital death following lung resection*, „European Journal of Cardio-Thoracic Surgery" 2005, vol. 28.

Esteva H., Núñez T.G., Rodríguez R.O., *Neural networks and artificial intelligence in thoracic surgery*, „Thoracic Surgery Clinics" 2007, vol. 17.

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.

Gallivan S., *Assessing mortality rates from dubious data − when to stop doing statistics and start doing mathematics*, „Health Care Management Science" 2005, vol. 8.

Garcia-Laencina P.J., Sancho-Gomez J-L., Figueiras-Vidal A.R., *Pattern classification with missing data: a review*, „Neural Computing & Applications" 2010, vol. 19.

Jefferson M.F., Pendleton N., Lucas S.B., Horan M.A., *Comparison of a Genetic algorithm neural network with logistic regression for predicting outcome after surgery for patients with nonsmall cell lung carcinoma*, „Cancer" 1997, vol. 79.

Misztal M., *Próba oceny wpływu wybranych metod imputacji danych na wyniki klasyfikacji obiektów z wykorzystaniem drzew klasyfikacyjnych*, [w:] K. Jajuga, M. Walesiak (red.), *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 176, UE, Wrocław 2011.

Santos-Garcia G., Varela G., Novoa N., Jimenez M.F., *Prediction of postoperative morbidity after lung resection using an artificial neural network ensemble*, „Artificial Intelligence in Medicine" 2004, vol. 30.

Schwarzer G., Vach W., Schumacher M., *On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology*, „Statistics In Medicine" 2000, vol. 19.

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

COMPARATIVE ANALYSIS OF SELECTED DATA MINING APPROACHES TO THE CLASSIFICATION OF MEDICAL DATA WITH MISSING VALUES (COVARIATES)

Summary: In implementation projects, in particular when analyzing medical data, it is quite often necessary to deal with tackle decision problems with missing values of specific variables (covariates). The aim of this paper is to perform a comparative analysis of selected data mining approaches, particularly simple and combined classifiers (ensembles) to solve classification tasks with missing data. The research was conducted using data mining techniques implemented in STATISTICA Data Miner and WEKA Machine Learning environments. The source data was extracted from a hospital data base of lung cancer patients treated surgically at Wrocław Thoracic Surgery Centre in the period 2000-2011.

Keywords: data mining, classification, missing values, medical data.