

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

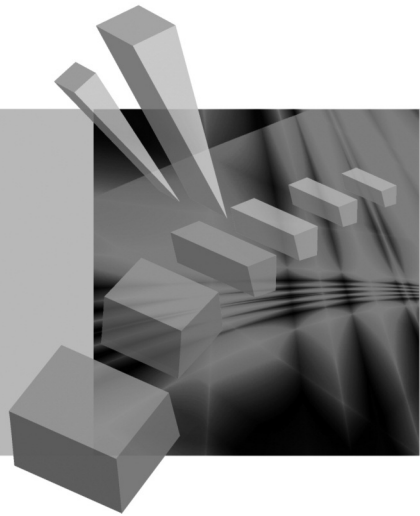
RESEARCH PAPERS

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 Pocięcha,
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](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

Kopowanie i powielanie w jakiegokolwiek formie
wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu
Wrocław 2012

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

Wstęp	13
Stanisława Bartosiewicz , Jeszcze raz o skutkach subiektywizmu w analizie wielowymiarowej	17
Andrzej Sokolowski , Q uniwersalna miara odległości	22
Eugeniusz Gatnar , Jakość danych w systemach statystycznych banków centralnych (na przykładzie NBP)	31
Marek Walesiak , Pomiar odległości obiektów opisanych zmiennymi mierzonymi na skali porządkowej – strategię postępowania.....	39
Krzysztof Jajuga, Marek Walesiak , XXV lat konferencji taksonomicznych – fakty i refleksje	47
Józef Pocięcha, Barbara Pawelek , Model SEM w analizie zagrożenia bankructwem przedsiębiorstw w świetle koniunktury gospodarczej – problemy teoretyczne i praktyczne	50
Paweł Lula , Uczące się systemy pozyskiwania informacji z dokumentów tekstowych	58
Ewa Roszkowska , Zastosowanie metody TOPSIS do wspomagania procesu negocjacji.....	68
Andrzej Młodak , Sąsiedztwo obszarów przestrzennych w ujęciu fizycznym oraz społeczno-ekonomicznym – podejście taksonomiczne	76
Andrzej Bąk , Modele kategorii nieuporządkowanych w badaniach preferencji	86
Jacek Kowalewski , Zintegrowany model optymalizacji badań statystycznych.....	96
Jan Paradysz, Karolina Paradysz , Obszary bezrobocia w Polsce – problem benchmarkowy.....	106
Tomasz Szubert , W co grać, aby jak najmniej przegrać? Próba klasyfikacji systemów gry w zakładach bukmacherskich.....	116
Izabela Szamrej-Baran , Klasyfikacja krajów UE ze względu na ubóstwo energetyczne	126
Sylwia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski , Analiza dojazdów do pracy za pomocą modelu grawitacji.....	135
Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz , Minimum egzystencji a czynniki warunkujące skłonność do korzystania z pomocy społecznej. Klasyfikacja gospodarstw domowych	144
Hanna Dudek , Subiektywne skale ekwiwalentności – analiza na podstawie danych o satysfakcji z osiągniętych dochodów	153

Joanicjusz Nazarko, Ewa Chodakowska, Marta Jaročka , Segmentacja szkół wyższych metodą analizy skupień <i>versus</i> konkurencja technologiczna ustalona metodą DEA – studium komparatywne.....	163
Ewa Chodakowska , Wybrane metody klasyfikacji w konstrukcji ratingu szkół.....	173
Bartosz Soliński , Sektor energetyki odnawialnej w krajach Unii Europejskiej – klasyfikacja w świetle strategii zarządzania zmianą.....	182
Krzysztof Szwarz , Klasyfikacja powiatów województwa wielkopolskiego ze względu na sytuację demograficzną.....	192
Elżbieta Gołata, Grażyna Dehnel , Rejestry administracyjne w analizie przedsiębiorczości.....	202
Katarzyna Chudy, Marek Sobolewski, Kinga Stępień , Wykorzystanie metod taksonomicznych w prognozowaniu wskaźników rentowności banków giełdowych w Polsce.....	212
Katarzyna Dębowska , Modelowanie upadłości przedsiębiorstw przy wykorzystaniu metod dyskryminacji i regresji.....	222
Alina Bojan , Wykorzystanie metod wielowymiarowej analizy danych do identyfikacji zmiennych wpływających na atrakcyjność wybranych inwestycji.....	231
Justyna Brzezińska , Analiza logarytmiczno-liniowa w badaniu przyczyn umieralności w krajach UE.....	240
Aneta Rybicka, Bartłomiej Jefmański, Marcin Pelka , Analiza klas ukrytych w badaniach satysfakcji studentów.....	247
Bartłomiej Jefmański , Pomiar opinii respondentów z wykorzystaniem elementów teorii zbiorów rozmytych i środowiska R.....	256
Julita Stańczuk , Porównanie rezultatów wielostanowej klasyfikacji obiektów ekonomicznych z wykorzystaniem analizy dyskryminacyjnej oraz sieci neuronowych.....	265
Jerzy Krawczuk , Skuteczność metod klasyfikacji w prognozowaniu kierunku zmian indeksu giełdowego S&P500.....	275
Anna Czapkiewicz, Beata Basiura , Symulacyjne badanie wpływu zaburzeń na grupowanie szeregów czasowych na podstawie modelu Copula-GARCH.....	283
Radosław Pietrzyk , Ocena efektywności inwestycji funduszy inwestycyjnych z tytułu doboru papierów wartościowych i umiejętności wykorzystania trendów rynkowych.....	291
Aleksandra Witkowska, Marek Witkowski , Zastosowanie metody Panzara-Rosse’a do pomiaru poziomu konkurencji w sektorze banków spółdzielczych.....	306
Marcin Pelka , Podejście wielomodelowe z wykorzystaniem metody <i>boosting</i> w analizie danych symbolicznych.....	315
Justyna Wilk , Analiza porównawcza oprogramowania komputerowego w klasyfikacji danych symbolicznych.....	323

Tomasz Bartłomowicz, Justyna Wilk , Zastosowanie metod analizy danych symbolicznych w przeszukiwaniu dziedzinowych baz danych.....	333
Kamila Migdał-Najman , Propozycja hybrydowej metody grupowania opartej na sieciach samouczących	342
Dorota Rozmus , Porównanie dokładności taksonomii spektralnej oraz zagregowanych algorytmów taksonomicznych opartych na idei metody <i>bagging</i>	352
Krzysztof Najman , Grupowanie dynamiczne z wykorzystaniem samouczących się sieci GNG	361
Małgorzata Misztal , Wpływ wybranych metod uzupełniania brakujących danych na wyniki klasyfikacji obiektów z wykorzystaniem drzew klasyfikacyjnych w przypadku zbiorów danych o niewielkiej liczebności – ocena symulacyjna	370
Mariusz Kubus , Zastosowanie wstępnego uwarunkowania zmiennej objaśnianej do selekcji zmiennych.....	380
Barbara Batóg, Jacek Batóg , Wykorzystanie analizy dyskryminacyjnej do identyfikacji czynników determinujących stopę zwrotu z inwestycji na rynku kapitałowym	387
Katarzyna Wójcik, Janusz Tuchowski , Analiza porównawcza miar podobieństwa tekstów opartych na macierzy częstości i tekstów opartych na wiedzy dziedzinowej	396
Iwona Staniec , Analiza czynnikowa w identyfikacji obszarów determinujących doskonalenie systemów zarządzania w polskich organizacjach	406
Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawelczyk, Jerzy Kołodziej, Jerzy Błaszczyk , Analiza porównawcza wybranych technik eksploracji danych do klasyfikacji danych medycznych z brakującymi obserwacjami	416
Iwona Foryś , Wykorzystanie analizy log-liniowej do wyboru czynników determinujących atrakcyjność cenową mieszkań w obrocie wtórnym na przykładzie lokalnego rynku mieszkaniowego.....	426
Ewa Genge , Analiza skupień oparta na mieszankach uciętych rozkładów normalnych.....	436
Jerzy Korzeniewski , Ocena efektywności metody uśredniania zmiennych i metody Ichino selekcji zmiennych w analizie skupień	444
Andrzej Dudek , SMS – propozycja nowego algorytmu analizy skupień	451
Artur Mikulec , Metody oceny wyniku grupowania w analizie skupień.....	460
Małgorzata Machowska-Szewczyk , Algorytm klasyfikacji rozmytej dla obiektów opisanych za pomocą zmiennych symbolicznych oraz rozmytych	469
Artur Zaborski , Analiza PROFIT i jej wykorzystanie w badaniu preferencji	479
Karolina Bartos , Analiza skupień wybranych państw ze względu na strukturę wydatków konsumpcyjnych obywateli – zastosowanie sieci Kohonena	488

Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak , Klasyfikacja gospodarstw domowych ze względu na bodźce do zawierania umowy o ubezpieczenie z wykorzystaniem modeli zmiennych jakościowych .	496
Izabela Kurzawa , Zastosowanie modelu LA/AIDS do badania elastyczności cenowych popytu konsumpcyjnego w gospodarstwach domowych w relacji miasto–wieś	505
Aleksandra Łuczak, Feliks Wysocki , Metody porządkowania liniowego obiektów opisanych za pomocą cech metrycznych i porządkowych	513
Agnieszka Sompolska-Rzechuła , Porównanie klasycznej i pozycyjnej taksonomicznej analizy zróżnicowania jakości życia w województwie zachodniopomorskim	523
Joanna Banaś, Małgorzata Machowska-Szewczyk , Ocena intensywności wykorzystania skrzynek poczty elektronicznej za pomocą uporządkowanego modelu probitowego	532
Iwona Bąk , Segmentacja gospodarstw domowych emerytów i rencistów pod względem wydatków na rekreację i kulturę	541
Aneta Becker , Zastosowanie metody ANP do porządkowania województw Polski pod względem dynamiki wykorzystania ICT w latach 2008-2010	552
Katarzyna Dębowska , Klasyfikacja sektorów ze względu na ich kondycję finansową przy użyciu metod wielowymiarowej analizy statystycznej	562
Anna Domagała , Propozycja metody doboru zmiennych do modeli DEA (procedura kombinowanego doboru w przód).....	571
Henryk Gierszal, Karina Pawlina, Maria Urbańska , Analiza statystyczna w badaniach zapotrzebowania na usługi teleinformatyczne sieci łączności ruchomej	580
Hanna Gruchociak , Konstrukcja estymatora regresyjnego dla danych o strukturze dwupoziomowej.....	590
Tomasz Klimanek, Marcin Szymkowiak , Zastosowanie estymacji pośredniej uwzględniającej korelację przestrzenną w opisie niektórych charakterystyk rynku pracy	601
Jarosław Lira , Prognozowanie opłacalności produkcji żywca wieprzowego w Polsce	610
Christian Lis , Wykorzystanie metody klasyfikacji w ocenie konkurencyjności portów południowego Bałtyku	619
Beata Bieszk-Stolorz, Iwona Markowicz , Wykorzystanie wielomianowego modelu logitowego do oceny szansy podjęcia pracy przez bezrobotnych .	628
Lucyna Przezbórska-Skobiej, Jarosław Lira , Przestrzeń agroturystyczna Polski i ocena jej atrakcyjności.....	637
Paweł Ulman , Model rozkładu wydatków a funkcje popytu.....	646
Maria Urbańska, Tadeusz Mizera, Henryk Gierszal , Zastosowanie metod analizy statystycznej w badaniach mięczaków	655

Summaries

Stanisława Bartosiewicz , The effects of subjectivism in multivariate analysis revisited.....	21
Andrzej Sokółowski , Q universal distance measure	30
Eugeniusz Gatnar , Data quality in central banks' statistical systems (NBP example)	38
Marek Walesiak , Distance measures for ordinal data – strategies of proceedings.....	46
Krzysztof Jajuga, Marek Walesiak , XXV years of taxonomic conferences – some facts and remarks.....	49
Józef Pocięcha, Barbara Pawelek , General SEM model in researching corporate bankruptcy and business cycles – theoretical and practical problems.....	57
Paweł Lula , Learning-based systems of information extraction from textual resources	67
Ewa Roszkowska , The application of the TOPSIS method to support the negotiation process	75
Andrzej Młodak , Neighborhood of spatial areas in the physical and socio-economic context – a taxonomic approach.....	85
Andrzej Bąk , Models for unordered categories in preference analysis.....	95
Kowalewski Jacek , An integrated model of optimizing statistical surveys	105
Jan Paradysz, Karolina Paradysz , Areas of unemployment in Poland – benchmark problem	115
Tomasz Szubert , How to play to lose the least? Classification of systems in sports bets	125
Izabela Szamrej-Baran , Classification of EU member states in view of fuel poverty	134
Sylvia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski , An attempt to use the gravity model in the analysis of commuters.....	143
Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz , Subsistence minimum versus factors influencing tendency to benefit from social care. Classification of households	152
Hanna Dudek , Subjective equivalence scales – analysis based on data about satisfaction with incomes.....	162
Joanicjusz Nazarko, Ewa Chodakowska, Marta Jarocka , Segmentation of universities using cluster analysis versus technological competitors determined by the DEA method – a comparative study	172
Ewa Chodakowska , Selected methods of classification in schools' rating.....	181
Bartosz Soliński , Renewable energy sector in the European Union – classification in the light of change management strategy	191
Krzysztof Szwarc , Classification of Wielkopolska voivodeship due to the demographic situation	201

Elżbieta Gołata, Grażyna Dehnel , Administrative registers in business analysis.....	211
Katarzyna Chudy, Marek Sobolewski, Kinga Stępień , Application of taxonomic methods in forecasting the profitability ratios of listed banks in Poland.....	221
Katarzyna Dębowska , Modeling bankruptcy of firms by using discrimination and regression methods.....	230
Alina Bojan , Identification of variables which influence attractiveness of given investments with the usage of multivariate analysis.....	239
Justyna Brzezińska , Log-linear analysis in the study of mortality in EU.....	246
Aneta Rybicka, Bartłomiej Jefmański, Marcin Pelka , Latent class analysis in student satisfaction surveys.....	254
Bartłomiej Jefmański , The respondent's opinions measurement in the R program with an application of fuzzy sets theory.....	264
Julita Stańczuk , A comparison of the results of multistate classification of economic objects using discriminant analysis and artificial neural networks.....	274
Jerzy Krawczuk , Effectiveness of classification methods in S&P500 stock index direction changes forecasting.....	282
Anna Czapkiewicz, Beata Basiura , The simulation study of the utility of the Copula-GARCH models for clustering financial time series.....	290
Radosław Pietrzyk , Timing and selectivity in mutual funds performance measurement.....	305
Aleksandra Witkowska, Marek Witkowski , Use of the Panzar-Rosse method to assess of the competition level in the cooperative banks sector.....	314
Marcin Pelka , Ensemble learning with the application of <i>boosting</i> in symbolic data analysis.....	322
Justyna Wilk , Comparative study of symbolic data classification software.....	332
Tomasz Bartłomowicz, Justyna Wilk , Application of symbolic data analysis methods for domain database searching.....	341
Kamila Migdał-Najman , A proposal of hybrid clustering method based on self-learning networks.....	351
Dorota Rozmus , Comparison of accuracy of spectral clustering and cluster ensembles stability based on bagging idea.....	360
Krzysztof Najman , A dynamic grouping based on self-learning GNG networks.....	369
Małgorzata Misztal , Influence of data imputation methods on the results of object classification using classification trees in the case of small data sets – simulation assessment.....	379
Mariusz Kubus , The application of pre-conditioning of explanatory variable for feature selection.....	386
Barbara Batóg, Jacek Batóg , Application of discriminant analysis to the identification of factors determining the rate of return on the capital market.....	395

Katarzyna Wójcik, Janusz Tuchowski , Comparative analysis of text documents similarity measures based on frequency matrix and based on domain knowledge.....	405
Iwona Staniec , Factor analysis in the identification of areas that determine the improvement of management systems in Polish organizations.....	415
Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawełczyk, Jerzy Kołodziej, Jerzy Błaszczyk , Comparative analysis of selected data mining approaches to the classification of medical data with missing values (covariates).....	425
Iwona Foryś , The log-linear analysis using to select the factors determining the attractiveness of the price of flats on the secondary market on the example of local housing market.....	435
Ewa Genge , Trimming approach to the mixtures of normal distributions.....	443
Jerzy Korzeniewski , Efficiency assessment of Ichino method and mean value method of selecting variables in cluster analysis.....	450
Andrzej Dudek , SMS – proposal of new clustering algorithm.....	459
Artur Mikulec , Evaluation methods for the grouping result in cluster analysis.....	468
Małgorzata Machowska-Szewczyk , Fuzzy clustering algorithm for objects described by symbolic or fuzzy variables.....	478
Artur Zaborski , PROFIT analysis and its using in the research of preferences.....	487
Karolina Bartos , Cluster analysis of selected countries due to the structure of their citizens' consumer expenditures – the use of Kohonen networks.....	495
Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak , Classification of households according to the impulses of concluding the insurance contract by means of qualitative variable models.....	504
Izabela Kurzawa , The application of LA/AIDS model to examine price elasticities of demand of households in the urban-rural relationship.....	512
Aleksandra Luczak, Feliks Wysocki , Linear ordering methods of objects described by a set of metric and ordinal characteristics.....	522
Agnieszka Sompolska-Rzechuła , The comparison of the classical and positional taxonomic analysis of the quality of life differentiation in Zachodniopomorskie voivodeship.....	531
Joanna Banaś, Małgorzata Machowska-Szewczyk , Evaluation of intensity of mailboxes using with the ordered probit model.....	540
Iwona Bąk , Segmentation of pensioners and annuitants households in terms of expenditures on recreation and culture.....	551
Aneta Becker , Application of ANP method to organize Polish voivodships in terms of dynamics of the use of ICT in 2008-2010.....	561
Katarzyna Dębowska , The classification of sectors' financial situation using the methods of multivariate statistical analysis.....	570

Anna Domagała , Proposal of a new method for variable selection in DEA models (combined forward stepwise selection method).....	579
Henryk Gierszal, Karina Pawlina, Maria Urbańska , Statistical analysis in demand research of ICT services in mobile networks.....	589
Hanna Gruchociak , Construction of regression estimator for two-level data	600
Tomasz Klimanek, Marcin Szymkowiak , Application of spatial models in indirect estimation of some labor market characteristics	609
Jarosław Lira , Forecasting of hog livestock production profitability in Poland	618
Christian Lis , The utilization of taxonomic methods in the appraisal of competitiveness of south Baltic ports	627
Beata Bieszk-Stolorz, Iwona Markowicz , The application of the multinomial logit model in evaluating employment odds for the unemployed job seekers	636
Lucyna Przezbórska-Skobiej, Jarosław Lira , Agritourism space of Poland and its valuation.....	645
Paweł Ulman , Model of expenses distribution and demand functions.....	654
Maria Urbańska, Tadeusz Mizera, Henryk Gierszal , Methods of statistical analysis in research of molluscs	663

Marek Lubicz, Maciej Zięba

Politechnika Wrocławska

Adam Rzechonek, Konrad Pawelczyk, Jerzy Kołodziej

Akademia Medyczna we Wrocławiu

Jerzy Błaszczuk

Dolnośląskie Centrum Onkologii we Wrocławiu

ANALIZA PORÓWNAWCZA WYBRANYCH TECHNIK EKSPLOKACJI 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 (jednorażową 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 k NN 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 przewyższenia 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. k NN) 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:

$$p = \exp(\text{logit}_2) / (1 + \exp(\text{logit}_2))$$

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

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 środowiska 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 $W \times 1$), 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 $W \times 1$ raz 20,9 i 16,9% dla $W \times 2$. Opracowanie podejść prognostycznych o efektywności akceptowalnej przez klinicystów wymaga w związku z tym dalszych badań.

Literatura

- Bellazzi 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., Utlej 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.