

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS

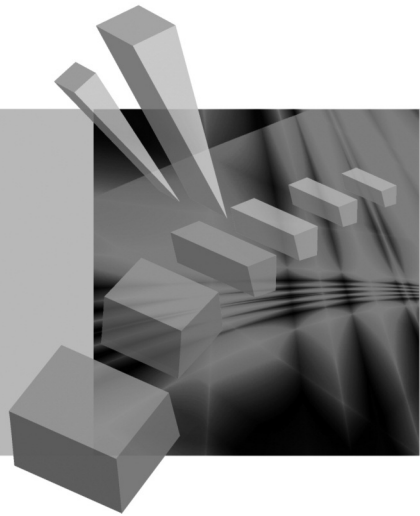
of Wrocław University of Economics

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 wspomaganie 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 Pawełczyk, 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

Małgorzata Misztal

Uniwersytet Łódzki

WPLYW WYBRANYCH METOD UZUPELNIANIA BRAKUJĄCYCH DANYCH NA WYNIKI KLASYFIKACJI OBIEKTÓW Z WYKORZYSTANIEM DRZEW KLASYFIKACYJNYCH W PRZYPADKU ZBIORÓW DANYCH O NIEWIELKIEJ LICZEBNOŚCI – OCENA SYMULACYJNA

Streszczenie: Drzewa klasyfikacyjne należą do tych algorytmów uczących, które mogą być wykorzystane w sytuacji występowania braków danych w zbiorze danych. W pracy porównano kilka wybranych technik postępowania w sytuacji występowania braków danych. Wykorzystano podejście symulacyjne, generując różne proporcje i mechanizmy powstawania braków danych w zbiorach danych pochodzących z repozytorium baz danych na Uniwersytecie Kalifornijskim w Irvine oraz z badań własnych. Celem badań była ocena wpływu wybranych metod imputacji danych na wyniki klasyfikacji obiektów z wykorzystaniem drzew klasyfikacyjnych w przypadku zbiorów danych o niewielkiej liczebności.

Słowa kluczowe: brakujące wartości, imputacja jednostkowa i wielokrotna, drzewa klasyfikacyjne.

1. Wstęp

Problem występowania w zbiorach danych brakujących wartości często pojawia się w praktycznych zastosowaniach metod statystycznych. Wśród sposobów postępowania w takich sytuacjach wymienia się odrzucenie obiektów z wartościami brakującymi, wykorzystanie algorytmu uczącego do rozwiązania problemu brakujących wartości w fazie uczenia oraz imputację brakujących wartości przed zastosowaniem algorytmu uczącego.

Drzewa klasyfikacyjne należą do grupy tych algorytmów uczących, w których w oryginalny sposób rozwiązano problem braków danych, zarówno w zbiorze uczącym, jak i w zbiorze testowym.

W pracy [Misztal 2011] podjęto próbę znalezienia odpowiedzi na pytanie, czy imputacja brakujących wartości przed zastosowaniem metody rekurencyjnego po-

działu daje dokładniejsze wyniki (w sensie dokładności predykcji) niż przyjęte w algorytmach drzew klasyfikacyjnych metody postępowania z brakami danych.

Uzyskane wówczas wyniki prowadzą do konkluzji, że imputacja brakujących wartości przed zastosowaniem metody rekurencyjnego podziału nie poprawia znacznie wyników klasyfikacji w przypadku zbiorów danych o dużej liczebności.

Można natomiast postawić hipotezę, że w przypadku zbiorów danych z małą liczbą obiektów oraz wysokim odsetkiem braków danych wykorzystanie metod imputacji przed zastosowaniem metody rekurencyjnego podziału wpływa istotnie na jakość klasyfikacji. Weryfikacja tej hipotezy jest celem głównym niniejszej pracy.

Podjęta w pracy tematyka jest stosunkowo rzadko spotykana w literaturze przedmiotu, szczególnie brak jest prac analizujących wyłącznie zbiory danych o niewielkiej liczbie obiektów. Opublikowane dotychczas wyniki badań, np. w pracach, takich jak [Ding, Simonoff 2010; Saar-Tsechansky, Provost 2007; Twala 2009; Twala, Jones, Hand 2008], nie zawierają jednoznacznych konkluzji i wskazówek dotyczących sposobów postępowania w sytuacji występowania braków danych w zbiorach danych analizowanych z wykorzystaniem drzew klasyfikacyjnych.

2. Podstawowe pojęcia

Imputacja jest metodą polegającą na zastąpieniu brakujących danych konkretnymi wartościami w celu uzyskania kompletnego zbioru danych.

Wyróżnia się imputację jednostkową (*Single Imputation – SI*) oraz wielokrotną (*Multiple Imputation – MI*). W przypadku imputacji jednostkowej uzupełnianie brakujących wartości przeprowadza się tylko raz, natomiast w imputacji wielokrotnej brakujące dane są uzupełniane kilka razy, analiza statystyczna przeprowadzana jest dla każdego z uzyskanych kompletnych zbiorów, a następnie uzyskane wyniki łączone są w jeden wynik końcowy.

Mechanizmy powstawania braków danych (*missing data mechanisms*) można podzielić (por. [Little, Rubin 2002]) na losowe (MCAR – *Missing Completely At Random* i MAR – *Missing At Random*) oraz nielosowe (NMAR – *Not Missing At Random*).

W przypadku danych typu MCAR brakujące wartości rozłożone są losowo wśród wszystkich wartości – prawdopodobieństwo wystąpienia brakującej wartości dla zmiennej X nie zależy od wartości samej zmiennej X ani od wartości pozostałych zmiennych.

Mechanizm typu MAR polega na tym, że brakujące wartości danej zmiennej nie zależą od wartości tej zmiennej, tylko od wartości pozostałych zmiennych w zbiorze danych, inaczej mówiąc – prawdopodobieństwo wystąpienia brakującej wartości dla zmiennej X nie zależy od wartości samej zmiennej X , tylko od obserwowanych wartości pozostałych zmiennych w zbiorze danych.

W przypadku mechanizmu typu NMAR brakujące wartości danej zmiennej zależą od czynników czy zdarzeń, których badacz nie jest w stanie zmierzyć, lub też wiążą się z samą wartością tej zmiennej.

W literaturze znaleźć można wiele różnorodnych technik uzupełniania brakujących danych (por. np. [Allison 2002; Little, Rubin 2002; Molenberghs, Kenward 2007]). Wśród częściej stosowanych metod wymienia się zastępowanie średnią/dominantą (*mean/mode imputation*), imputację regresyjną (*regression imputation*), imputację typu *hot deck* (*hot deck imputation*), metody modelowe (metoda największej wiarygodności, algorytm EM), metodę *predictive mean matching* czy imputację typu *k* najbliższych sąsiadów (*kNN imputation*).

Jak już wspomniano, w drzewach klasyfikacyjnych zastosowano oryginalne metody postępowania w sytuacji występowania braków danych. Szczególnie w algorytmie CART (por. [Breiman i in. 1984]) korzysta się z tzw. zmiennych zastępczych. Do podziału w danym węźle, zamiast zmiennej x_m , która w danym obiekcie nie wystąpiła, wykorzystywana jest zmienna zastępcza x^* , wybierana w taki sposób, aby uzyskany podział w węźle był jak najbardziej zbliżony do tego, jaki daje zmienna x_m . W każdym kroku analizy budowany jest ranking zmiennych zastępczych. Obiekt z brakującą wartością zmiennej wykorzystanej do podziału jest klasyfikowany z wykorzystaniem pierwszej w rankingu zmiennej zastępczej, a jeśli dla niej także występuje brak danych, to uwzględniana jest następna zmienna zastępcza itd.

3. Założenia eksperymentu

W celu weryfikacji postawionej wcześniej hipotezy badawczej wykorzystano 10 zbiorów danych empirycznych o niewielkiej liczebności, pochodzących z repozytorium baz danych na Uniwersytecie Kalifornijskim w Irvine (UCI) – por. [Blake, Keogh, Merz 1988], oraz z badań własnych.

Tabela 1. Charakterystyka wykorzystanych zbiorów danych

Nazwa zbioru	Id	Źródło	Liczba obserwacji	Liczba zmiennych objaśniających	Liczba klas
Protein Localization Sites	E.coli	UCI	336	5	8
Glass Identification Database	glass	UCI	214	9	2
Haberman's Survival Data	haberman	UCI	306	3	2
Iris Plants Database	iris	UCI	150	4	3
Breast Tissue	breastT	UCI	106	9	6
Wine Recognition Data	wine	UCI	178	13	3
Wisconsin Prognostic Breast Cancer	wpbc	UCI	194	12	2
Kredyty indywidualne	cred	B.W.	100	6	2
Narkomani	nark	B.W.	60	5	2
Zespół metaboliczny	zm	B.W.	86	21	2

Źródło: opracowanie własne.

Podstawowe informacje dotyczące wykorzystanych zbiorów danych przedstawia tab. 1.

W każdym zbiorze danych generowano braki danych według trzech rodzajów mechanizmu powstawania brakujących wartości – MCAR, MAR, NMAR. Przyjęto ogólny wzorzec braków danych – braki danych mogły się pojawić w każdej zmiennej poza zmienną zależną Y.

W przypadku mechanizmu typu MCAR braki danych pojawiają się losowo w całym zbiorze danych. Dla mechanizmu typu MAR przyjęto założenie, że występowanie braków danych jest bardziej prawdopodobne w określonych podgrupach wyróżnionych według zmiennej zależnej Y. Przy generowaniu braków typu NMAR dla danej zmiennej usuwano największe lub najmniejsze wartości. Przykładowy wynik zastosowanych metod generowania braków przedstawia rys. 1.

Oryginalne dane:

x1	x2	x3	x4	x5	x6	y
36	702	7	850	76.34	11	wind
32	5000	7	1100	241.49	35	wind
31	5000	13	1340.21	292.58	24	wind
35	7710	24	2446	356.53	35	wind
32	7679	4	1750	370.88	35	wind
23	2800	3	560	205.5	18	wind
28	1147	4	676.9	114.91	12	wind
20	1300	5	1950	77.36	23	wind
36	1320	93	1011.3	142.24	11	wind
31	1612	22	594	93.94	23	wind
46	2776	58	1034.3	272.4	12	splac
41	2199	95	1752.24	154.7	18	splac
23	4300	8	1450	195.31	35	splac
40	3018	137	2416.23	137.08	35	splac
24	4950	20	1038.91	224.83	35	splac
50	1000	34	1163.4	98.13	12	splac
42	4742	231	1293	215.39	35	splac
32	3268	105	1940	375.15	10	splac
41	3000	214	1539.08	350.55	10	splac
43	3000	71	1433.7	323.28	11	splac

MCAR: BRAKI 10%

x1	x2	x3	x4	x5	x6	y
36	702	7	NA	76.34	11	wind
32	5000	7	1100	241.49	35	wind
31	5000	13	1340.21	292.58	24	wind
35	7710	NA	2446	356.53	35	wind
32	7679	4	1750	NA	35	wind
23	NA	3	560	205.5	18	wind
28	1147	NA	676.9	114.91	12	wind
20	1300	5	1950	77.36	23	wind
36	1320	93	1011.3	142.24	11	wind
31	1612	22	594	93.94	NA	wind
46	2776	58	NA	272.4	12	splac
41	NA	95	1752.24	154.7	18	splac
23	4300	8	1450	195.31	35	splac
40	3018	137	2416.23	137.08	NA	splac
24	4950	20	1038.91	224.83	35	splac
50	1000	34	1163.4	98.13	12	splac
42	4742	231	1293	215.39	35	splac
NA	3268	105	1940	375.15	10	splac
41	3000	214	1539.08	350.55	NA	splac
43	3000	71	1433.7	NA	11	splac

MAR: BRAKI 10%

x1	x2	x3	x4	x5	x6	y
36	702	7	850	76.34	11	wind
32	NA	7	1100	241.49	35	wind
31	NA	13	1340.21	NA	24	wind
35	7710	24	2446	356.53	35	wind
NA	7679	4	1750	370.88	35	wind
23	2800	3	NA	205.5	18	wind
28	1147	4	676.9	114.91	12	wind
20	1300	5	NA	NA	23	wind
36	1320	93	1011.3	142.24	11	wind
31	1612	22	594	93.94	NA	wind
46	2776	58	1034.3	272.4	12	splac
NA	2199	95	1752.24	154.7	18	splac
23	4300	8	1450	195.31	35	splac
NA	3018	NA	2416.23	137.08	35	splac
NA	NA	20	1038.91	224.83	35	splac
50	1000	34	1163.4	98.13	12	splac
42	4742	231	1293	215.39	35	splac
32	3268	105	1940	375.15	10	splac
41	3000	214	1539.08	350.55	10	splac
43	3000	71	1433.7	323.28	11	splac

NMAR: BRAKI 10%

x1	x2	x3	x4	x5	x6	y
36	702	7	850	76.34	11	wind
32	NA	7	1100	241.49	35	wind
31	NA	13	1340.21	292.58	24	wind
35	NA	24	2446	356.53	35	wind
32	NA	4	1750	370.88	35	wind
23	NA	3	560	205.5	18	wind
28	1147	4	676.9	114.91	12	wind
20	1300	5	1950	77.36	23	wind
36	NA	93	1011.3	142.24	11	wind
31	NA	22	594	93.94	23	wind
46	2776	58	1034.3	272.4	12	splac
NA	2199	95	1752.24	154.7	18	splac
NA	4300	8	1450	195.31	35	splac
NA	3018	137	2416.23	137.08	35	splac
NA	4950	20	1038.91	224.83	35	splac
50	1000	34	1163.4	98.13	12	splac
42	4742	231	1293	215.39	35	splac
NA	3268	105	1940	375.15	10	splac
41	3000	214	1539.08	350.55	10	splac
43	3000	71	1433.7	323.28	11	splac

Rys. 1. Przykładowe wyniki generowania braków danych

Źródło: opracowanie własne.

W kolejnych eksperymentach usuwano 5, 10, 20, 30 i 40% danych.

Wykorzystano 4 metody uzupełniania brakujących wartości: (1) zastępowanie średnią (SI – *mean*), (2) imputację typu *hot deck* – zastępowanie braku wartością wylosowaną spośród obserwowanych wartości (SI – *sample*), (3) zastępowanie metodą *predictive mean matching* (SI – *pmm*) oraz (4) imputację wielokrotną metodą *predictive mean matching* (MI – *pmm*).

Następnie budowano drzewa klasyfikacyjne CART (z wykorzystaniem procedury *rpart* w pakiecie R) dla:

- oryginalnego zbioru danych;
- zbioru danych z brakującymi wartościami;
- zbioru danych po usunięciu obiektów z brakującymi wartościami (*complete case analysis*);
- zbiorów danych z uzupełnionymi brakami (4 metody uzupełniania).

Każdy eksperyment powtarzano 1000 razy. Błąd klasyfikacji szacowano za pomocą metody sprawdzania krzyżowego 10-CV.

Obliczenia wykonano w środowisku R z wykorzystaniem pakietów: *rpart*, *ipred*, *mice*.

4. Wyniki

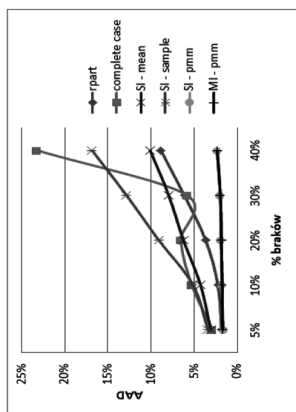
Na podstawie uzyskanych wyników obliczono odchylenia od wartości „wzorcowej” – błędu klasyfikacji dla oryginalnego zbioru danych:

$$|Err_I - Err_C|,$$

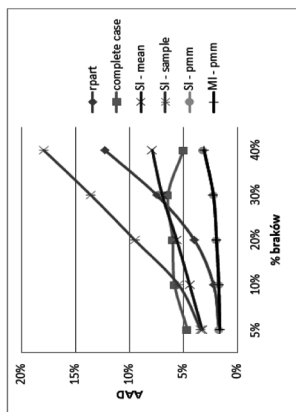
gdzie: Err_I – błąd klasyfikacji dla zbioru danych z uzupełnionymi brakami danych;
 Err_C – błąd klasyfikacji dla kompletnego, oryginalnego zbioru danych.

Na rysunkach 2-4 przedstawiono uśrednione wyniki (średnie odchylenie bezwzględne – AAD – *Average Absolute Deviation*) dla 10 rozważanych zbiorów danych, 3 mechanizmów powstawania braków danych, 5 różnych odsetków braków danych oraz 5 metod postępowania w sytuacji wystąpienia brakujących wartości (*Complete Case Analysis*) i 4 metody imputacji.

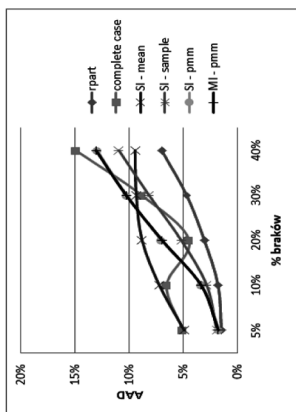
Jak widać na rysunkach, w przypadku mechanizmów powstawania braków danych typu MCAR i MAR i przy niewielkim odsetku braków danych (5-10%) zarówno drzewo klasyfikacyjne CART (*rpart*), zastosowane do zbiorów danych z brakującymi wartościami, jak i imputacja metodą *predictive mean matching* przed budową drzewa (jednostkowa: SI – *pmm* i wielokrotna: MI – *pmm*) dają podobne wyniki, zdecydowanie najmniej różniące się od wyników uzyskanych dla kompletnego, oryginalnego zbioru danych. Przy większej liczbie braków (20% i więcej) zaobserwowano przewagę wykorzystania imputacji metodą *predictive mean matching* nad pozostałymi metodami.



Rys. 2. Porównanie wyników w przypadku braków typu MCAR
Źródło: obliczenia własne.



Rys. 3. Porównanie wyników w przypadku braków typu MAR
Źródło: obliczenia własne.



Rys. 4. Porównanie wyników w przypadku braków typu NMAR
Źródło: obliczenia własne.

Odmierna sytuacja występuje, gdy mechanizm powstawania braków danych jest nielosowy (NMAR) – wyniki uzyskane z wykorzystaniem drzewa klasyfikacyjnego CART (*rpart*) dla zbiorów danych z brakującymi wartościami są zawsze dokładniejsze w porównaniu z pozostałymi metodami postępowania.

W celu dokładniejszej analizy uzyskanych rezultatów zastosowano trójczynnиковą analizę wariancji (ANOVA). Jej wyniki podsumowano w tab. 2 oraz na rys. 5-9¹. Jak wynika z tab. 2, wszystkie efekty główne oraz trzy efekty interakcji są istotne statystycznie.

Tabela 2. Wyniki analizy wariancji

Czynnik	Poziom <i>p</i>
MDM (mechanizm powstawania braków danych)	0,0014
MV% (odsetek braków danych)	0,0000
Metoda (sposób postępowania w sytuacji wystąpienia braków danych) ²	0,0000
Interakcja: MDM*MV%	0,6297
Interakcja: MDM*Metoda	0,0000
Interakcja: MV%*Metoda	0,0003
Interakcja: MDM*MV%*Metoda	0,0300

Źródło: obliczenia własne.

Można zaobserwować (rys. 5), że w przypadku nielosowego mechanizmu powstawania braków (NMAR) otrzymano znacznie mniej dokładne wyniki klasyfikacji w porównaniu z mechanizmami losowymi (MCAR i MAR).

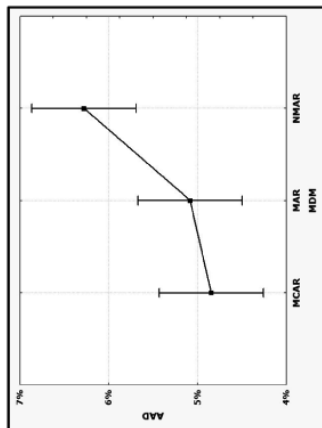
Błąd klasyfikacji rośnie ze wzrostem odsetka brakujących wartości w zbiorze danych (rys. 6), istotne różnice nie występują tylko między wynikami dla 5 i 10% braków danych.

Wybór metody postępowania z brakami danych wpływa na jakość klasyfikacji (rys. 7). Wyniki uzyskane w przypadku zastosowania imputacji typu *hot – deck* (SI – *sample*) oraz zastępowania średnią (SI – *mean*) są istotnie gorsze od wyników dla pozostałych metod.

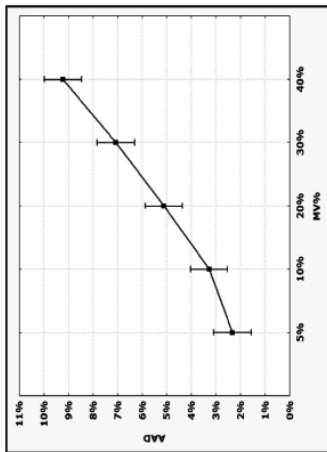
Analizując interakcję mechanizmu powstawania braków danych i metody postępowania (rys. 8), można zauważyć, że procedura zmiennych zastępczych zaimplementowana w algorytmie CART (*rpart*) daje podobne wyniki zarówno przy losowym, jak i nielosowym mechanizmie powstawania braków danych. Z drugiej strony, widać, że przy imputacji metodą *predictive mean matching* (SI – *pmm* i MI – *pmm*) w sytuacji braków nielosowych (NMAR) następuje zdecydowane pogorszenie jakości klasyfikacji.

¹ Pominięto prezentację graficzną dla interakcji trzech czynników.

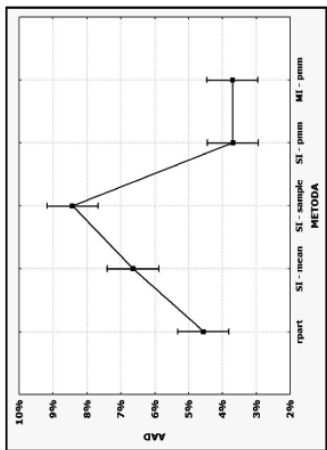
² Nie uwzględniono metody usuwania obiektów z co najmniej jedną brakującą wartością (*complete case analysis*).



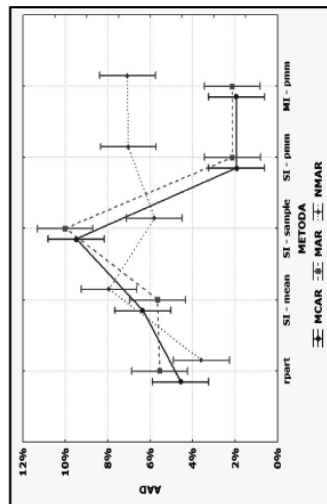
Rys. 5. Ocena wpływu mechanizmu powstawania braków
Źródło: obliczenia własne.



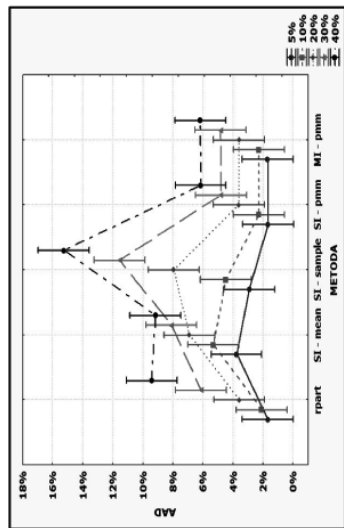
Rys. 6. Ocena wpływu odsetka brakujących danych
Źródło: obliczenia własne.



Rys. 7. Ocena wpływu metody postępowania z brakami danych
Źródło: obliczenia własne.



Rys. 8. Ocena wpływu interakcji mechanizmu powstawania braków i metody postępowania
Źródło: obliczenia własne.



Rys. 9. Ocena wpływu interakcji odsetka brakujących danych i metody postępowania
Źródło: obliczenia własne.

Badając interakcję odsetka brakujących danych i metody postępowania (rys. 9), można stwierdzić, że wszystkie metody dają podobne wyniki przy niskim (5-10%) odsetku braków; różnice między metodami postępowania są wyraźnie widoczne przy wyższych odsetkach brakujących wartości.

Szczegółowa analiza wyników dla interakcji wszystkich trzech czynników potwierdza wnioski przedstawione powyżej dla efektów głównych i interakcji par czynników.

5. Uwagi końcowe

Analizując uzyskane wyniki, można zauważyć wyraźny wpływ mechanizmu powstawania braków danych na rezultaty klasyfikacji. W przypadku MCAR i MAR zastosowanie imputacji jednostkowej lub wielokrotnej metodą *predictive mean matching* najmniej zniekształca uzyskane wyniki (uzyskano najniższe odchylenia od rzeczywistego wyniku). W przypadku NMAR wszystkie metody imputacji dają gorsze wyniki niż zaimplementowana w CART procedura postępowania z brakami danych.

Przy niewielkiej liczbie braków (5-10%) wszystkie sposoby postępowania dają podobne wyniki.

Reasumując, należy stwierdzić, że korzyści z zastosowania metod imputacji przed budową drzewa klasyfikacyjnego są na tyle mało znaczące, że nie rekompensują kosztów zastosowania tych metod (pracochłonność, dodatkowy czas obliczeń, wymagany dodatkowy wkład pracy od badacza itp.). Co istotne, procedury zaimplementowane w algorytmach budowy drzew nie stawiają wymagań co do mechanizmu powstawania braków danych ani odsetka brakujących wartości.

Z drugiej jednak strony, przeprowadzone eksperymenty symulacyjne obejmują zaledwie fragment złożonej problematyki imputacji brakujących wartości.

W kolejnych badaniach należałoby uwzględnić: (1) inne metody imputacji danych (np. algorytm k-NN, algorytm EM), (2) inne algorytmy budowy drzew klasyfikacyjnych, typu CRUISE czy QUEST, w których problem braków danych rozwiązano w odmienny sposób niż w algorytmie CART, (3) dodatkowe wzorce braków danych (poza zastosowanym wzorcem ogólnym, także np. wzorec monotoniczny).

Dodatkowo interesującym zagadnieniem byłoby także urealnienie mechanizmu generowania braków danych przez jednoczesne uwzględnienie w zbiorze danych braków losowych i nielosowych, a także zmiana sposobu oceny jakości imputacji.

Literatura

- Allison P.D., *Missing Data*, Series: Quantitative Applications in the Social Sciences 07-136, SAGE Publications, Thousand Oaks, London, New Delhi 2002.
- Blake C., Keogh E., Merz C.J., *UCI Repository of Machine Learning Datasets*, Department of Information and Computer Science, University of California, Irvine 1988.

- Breiman L., Friedman J., Olshen R., Stone C., *Classification and Regression Trees*, CRC Press, London 1984.
- Ding Y., Simonoff J.S., *An investigation of missing data methods for classification trees applied to binary response data*, „Journal of Machine Learning Research” 2010, no 11.
- Little R.J.A., Rubin D.B., *Statistical Analysis with Missing Data*, Second Edition, Wiley, New Jersey 2002.
- Misztal M., *Próba oceny wpływu wybranych metod imputacji danych na wyniki klasyfikacji obiektów z wykorzystaniem drzew klasyfikacyjnych*, [w:] Taksonomia 18, *Klasyfikacja i analiza danych – teoria i zastosowania*, red. K. Jajuga, M. Walesiak, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 176, Wydawnictwo AE, Wrocław 2011.
- Molenberghs G., Kenward M.G., *Missing Data in Clinical Studies*, Wiley, England 2007.
- Saar-Tschansky M., Provost F., *Handling missing values when applying classification models*, „Journal of Machine Learning Research” 2007, no 8.
- Twala B., *An empirical comparison of techniques for handling incomplete data using decision trees*, „Applied Artificial Intelligence” 2009, no 23.
- Twala B., Jones M.C., Hand D.J., *Good methods for coping with missing data in decision trees*, „Pattern Recognition Letters” 2008, no 29(7).

INFLUENCE OF DATA IMPUTATION METHODS ON THE RESULTS OF OBJECT CLASSIFICATION USING CLASSIFICATION TREES IN THE CASE OF SMALL DATA SETS – SIMULATION ASSESSMENT

Summary: Classification tree is an example of the learning algorithm coping with missing values. In the paper some selected missing data techniques are compared by artificially simulating different proportions and mechanisms of missing data using complete data sets mainly from the UCI repository of machine learning databases. The goal of the paper is to assess the influence of these techniques on the results of object classification by means of classification trees in the case of small data sets.

Keywords: missing values, single and multiple imputation, classification trees.