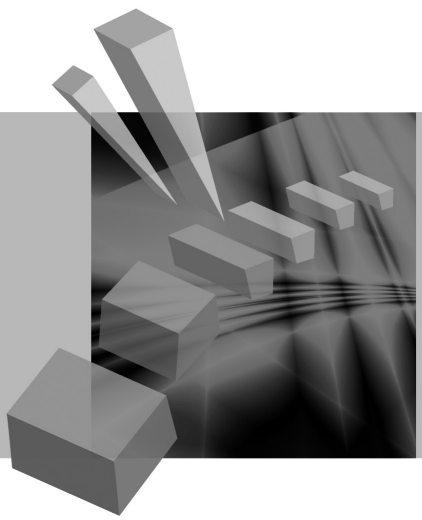
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

279

Taksonomia 21

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2013

Redaktor Wydawnictwa: Aleksandra Śliwka
Redaktor techniczny: Barbara Łopusiewicz
Korektor: Barbara Cibis
Łamanie: Małgorzata Czupryńska
Projekt okładki: Beata Dębska

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

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

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

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

Wersja pierwotna: publikacja drukowana
Druk: Drukarnia TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 279 • 2013

Taksonomia 21. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Małgorzata Misztal
Uniwersytet Łódzki

OCENA WPŁYWU WYBRANYCH METOD IMPUTACJI NA WYNIKI KLASYFIKACJI OBIEKTÓW W MODELACH DRZEW KLASYFIKACYJNYCH

Streszczenie: W przeciwieństwie do większości metod statystyki wielowymiarowej drzewa klasyfikacyjne należą do grupy algorytmów uczących, w których w oryginalny sposób rozwiązano problem występowania brakujących wartości w analizowanych zbiorach danych. W pracy zbadano wpływ wybranych metod imputacji danych na wyniki klasyfikacji obiektów z wykorzystaniem drzew klasyfikacyjnych CART i CRUISE oraz porównano procedury imputacji zaimplementowane bezpośrednio w obu algorytmach budowy drzewa. 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.

Słowa kluczowe: braki danych, imputacja, drzewa klasyfikacyjne.

1. Wstęp

W sytuacji występowania braków danych w analizowanych w praktyce zbiorach danych wymieniane są trzy sposoby postępowania (por. np. [Hastie i in. 2008; Song i in. 2008]): (1) odrzucenie obiektów z wartościami brakującymi, (2) tolerowanie braków danych (wykorzystanie algorytmu uczącego do rozwiązania problemu brakujących wartości w fazie uczenia) oraz (3) uzupełnianie braków danych (imputacja brakujących wartości przed zastosowaniem algorytmu uczącego).

Podejście (2) dotyczy tylko algorytmów opartych na metodzie rekurencyjnego podziału (drzewa decyzyjne). W przypadku innych algorytmów uczących stosowane są zwykle podejścia (1) i (3).

Prezentowany artykuł jest kontynuacją badań opisanych w pracy Misztal [2012], w której porównano kilka wybranych technik postępowania w sytuacji występowania braków danych oraz zbadano ich wpływ na wyniki klasyfikacji obiektów z wykorzystaniem drzewa klasyfikacyjnego CART [Breiman i in. 1984].

Celem głównym niniejszej pracy jest zbadanie wpływu wybranych, prostych metod imputacji danych na wyniki klasyfikacji obiektów z wykorzystaniem drzew klasyfikacyjnych CART i CRUISE [Kim, Loh 2001]. Do celów szczegółowych na-

leży porównanie algorytmów imputacji braków danych zaimplementowanych bezpośrednio w procedurach budowy drzew CART i CRUISE oraz znalezienie odpowiedzi na pytanie, w jaki sposób imputacja braków danych przed budową drzewa zmieni dokładność klasyfikacji.

2. Metody imputacji w algorytmach CART I CRUISE

Drzewa klasyfikacyjne należą do tej grupy algorytmów uczących, w których w oryginalny sposób rozwiązano problem występowania w zbiorze danych brakujących wartości.

W algorytmie CART [Breiman i in. 1984] opracowano procedurę wykorzystującą tzw. zmienne zastępcze (*surrogate variables/splitters*). Polega ona na użyciu do podziału w danym węźle zmiennej X* zamiast zmiennej X, która w tym obiekcie nie wystąpiła. Zmienna zastępcza X* wybierana jest w taki sposób, aby uzyskany podział w węźle był jak najbardziej zbliżony do tego, jaki daje zmienna X.

Szukając zmiennej do podziału, w danym węźle brane są pod uwagę tylko te obiekty, dla których nie ma braków danych w tej zmiennej. Po wyborze najlepszego predyktora i optymalnego punktu podziału tworzony jest ranking zmiennych zastępczych z optymalnymi dla nich punktami podziału. Jeśli dla zmiennej najlepszej do podziału występuje brak danych, to wykorzystywana jest pierwsza zmienna zastępcza, jeśli w niej także są braki danych – to druga itd.

Zmienne zastępcze wykorzystują korelacje między zmiennymi, zatem im wyższa korelacja między zmiennymi, tym mniejsza utrata informacji związana z wystąpieniem braku danych.

W algorytmie CRUISE [Kim, Loh 2001] przyjęto inne rozwiązanie. W przypadku braków danych w zbiorze uczącym wybór zmiennej do podziału i optymalnego punktu podziału oparty jest wyłącznie na dostępnych wartościach danej zmiennej w węźle (*available case solution*). Brakujące wartości są zastępowane wartością średnią lub modalną dla danej klasy. Następnie dokonywany jest podział obiektów w węźle, a imputowane wartości zostają usunięte.

Jeżeli braki danych występują w zbiorze testowym, stosowane jest podejście oparte na tzw. zmiennej alternatywnej (*alternate variable*). Jeżeli optymalna do podziału w danym węźle zmienna X nie wystąpiła dla klasyfikowanego obiektu, to na podstawie wartości drugiej w kolejności optymalnej zmiennej X* identyfikowana jest klasa, do której należy badany obiekt. Następnie brakujące wartości zmiennej X są zastępowane średnią lub modalną dla danej klasy w danym węźle. Jeżeli zmienna X* dla danego obiektu również nie występuje, to brakujące wartości zmiennej X są zastępowane średnią lub modalną z wartości w danym węźle, bez uwzględniania przynależności do klas. Po przydzieleniu obiektów do węzłów – potomków imputowane wartości zostają usunięte.

3. Założenia eksperymentu

W celu weryfikacji postawionej hipotezy badawczej wykorzystano 6 zbiorów danych empirycznych, pochodzących z repozytorium baz danych na Uniwersytecie Kalifornijskim w Irvine (UCI – por. [Blake, Keogh, Merz 1988]) oraz z badań własnych (BW). Podstawowe informacje dotyczące wykorzystanych zbiorów danych przedstawia tab. 1.

Każdy zbiór danych podzielono w sposób losowy na część uczącą i testową o podobnej liczebności.

W każdym zbiorze uczącym generowano braki danych według trzech rodzajów mechanizmu powstawania brakujących wartości – MCAR, MAR, NMAR (por. [Little, Rubin 2002]). Przyjęto ogólny wzorzec braków danych – braki danych mogły się pojawić w każdej zmiennej poza zmienną zależną Y.

Tabela 1. Charakterystyka wykorzystanych zbiorów danych.

Nazwa zbioru/źródło	Liczba obiektów	Liczba cech	Liczba klas
Glass Identification Database (UCI)	214	9	2
Iris Plants Database (UCI)	150	4	3
Wine Recognition Data (UCI)	178	13	3
Wisconsin Prognostic Breast Cancer (UCI)	194	12	2
Vertebral Column (UCI)	310	6	2
Atrial Fibrillation (BW)	300	9	2

Źródło: opracowanie własne.

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 wylosowanej zmiennej usuwano największe lub najmniejsze wartości.

W kolejnych eksperymentach usuwano 5, 10, 20, 30 i 40% danych w zbiorze uczącym; w zbiorze testowym braki danych nie występowały.

Zbiór uczący był wykorzystywany do budowy drzewa, a zbiór testowy do szacowania błędu klasyfikacji.

Wykorzystano 4 metody uzupełniania brakujących wartości: (1) zastępowanie średnią (*mean*), (2) imputację typu *hot deck* – zastępowanie braku wartością wylosowaną spośród obserwowanych wartości (*sample*), (3) zastępowanie metodą *predictive mean matching* (*pmm*) oraz (4) imputację z wykorzystaniem metody *missForest* (*mF*, por. [Stekhoven, Bühlmann 2012]). W przeprowadzonych badaniach zrezygnowano z metod imputacji wielokrotnej ze względu na trudności z oceną uzyskanych wyników – trudno ocenić, czy poprawa wyników jest efektem samej imputacji wielokrotnej czy też agregacji uzyskanych modeli.

W kolejnych krokach budowano drzewa klasyfikacyjne CART i CRUISE dla: (1) oryginalnego zbioru danych, (2) zbioru danych z brakującymi wartościami (z wykorzystaniem algorytmu zaimplementowanego w procedurze budowy drzewa – *tree*) oraz (3) zbiorów danych z uzupełnionymi brakami (4 metody uzupełniania: *mean*, *sample*, *pmm*, *mF*).

Każdy eksperyment powtarzano 1000 razy. Przyjęty schemat postępowania został przeprowadzony 2 razy – dokonano zamiany zbioru uczącego i testowego. Wyniki z obu „edycji" uśredniono.

Do obliczeń wykorzystano środowisko R (pakiety: `rpart`, `mice`, `missForest`) oraz program do budowy drzew klasyfikacyjnych CRUISE (ver. 3.6.3) udostępniony na stronie: http://www.stat.wisc.edu/~loh/cruise.html.

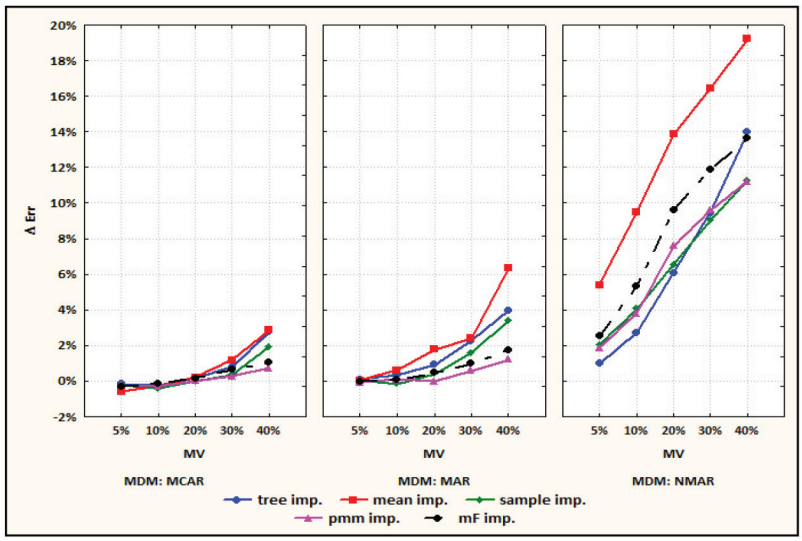
4. Wyniki

Uzyskane wyniki przeanalizowano z wykorzystaniem procedury zaproponowanej w pracach Twali [2009] oraz Twali, Jonesa i Handa [2008]. W pierwszej kolejności obliczono dla każdego zbioru danych przyrosty błędów:

$$\Delta Err = Err_I - Err_C,$$

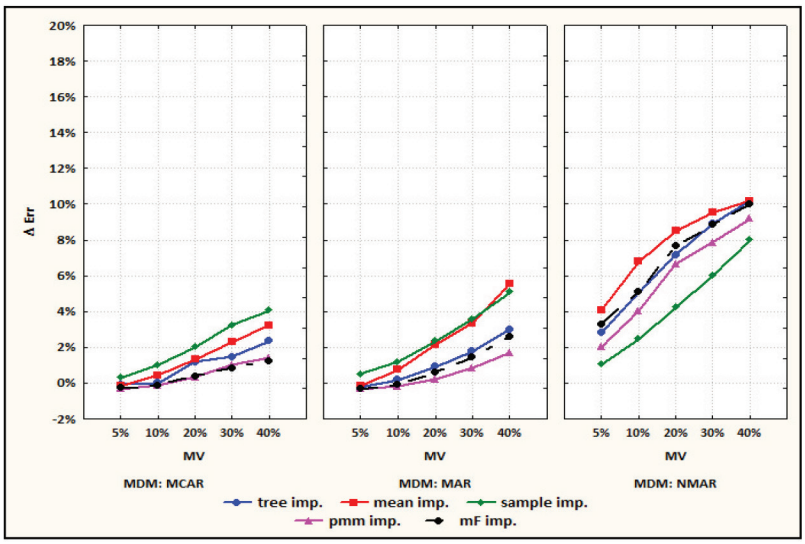
gdzie: Err_I – błąd klasyfikacji dla zbioru danych z uzupełnionymi brakami danych;
Err_C – błąd klasyfikacji dla kompletnego, oryginalnego zbioru danych.

Uśrednione wyniki dla 6 rozważanych zbiorów danych, 3 mechanizmów powstawania braków danych, 5 różnych odsetków braków danych oraz 5 metod uzupełniania brakujących wartości przedstawiono na rys. 1-2.



Rys. 1. Porównanie wyników – algorytm CART

Źródło: obliczenia własne.



Rys. 2. Porównanie wyników – algorytm CRUISE

Źródło: obliczenia własne.

Jak widać na rys. 1 i 2, w przypadku mechanizmów powstawania braków danych typu MCAR i MAR oraz przy niewielkim odsetku brakujących wartości (5-10%) oba drzewa klasyfikacyjne (CART i CRUISE) zastosowane do zbiorów danych z brakującymi wartościami, jak również wcześniejsza imputacja przed budową drzewa metodami *predictive mean matching* oraz *missForest* dają podobne wyniki, zdecydowanie najmniej różniące się od wyników uzyskanych dla kompletnego, oryginalnego zbioru danych. Przy większej liczbie braków (20% i więcej) błąd klasyfikacji wzrasta, przede wszystkim dla imputacji metodami *mean* oraz *sample*.

Jeżeli mechanizm powstawania braków danych jest nielosowy (NMAR), wyniki uzyskane z wykorzystaniem drzewa klasyfikacyjnego CRUISE wydają się bardziej stabilne, a uzyskane błędy klasyfikacji niższe niż w przypadku drzewa klasyfikacyjnego CART.

W celu dokładniejszej analizy uzyskanych rezultatów zastosowano analizę wariancji z powtarzanymi pomiarami. Jej wyniki podsumowano w tab. 2 oraz na rys. 3-8[1].

Jak wynika z tab. 2, istotne statystycznie są 2 efekty główne (efekt mechanizmu powstawania braków danych i efekt odsetka brakujących wartości), efekt powtarzanego pomiaru (metoda uzupełniania braków) oraz trzy efekty interakcji.

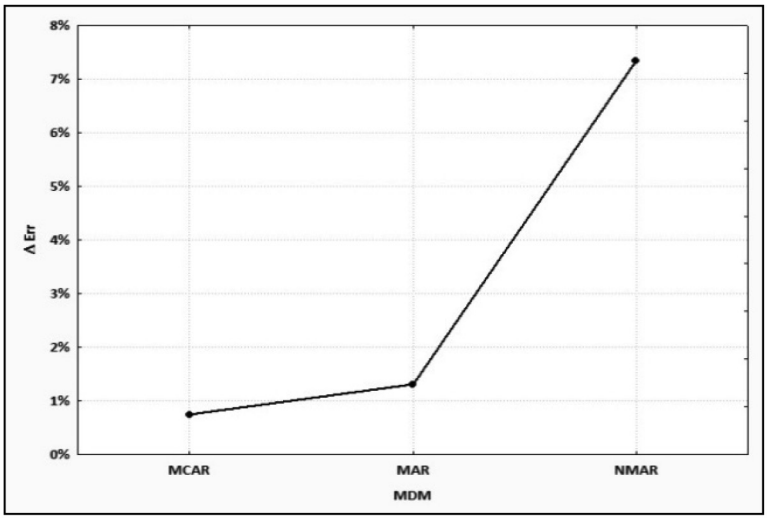
Analizując uzyskane wyniki, można zaobserwować (rys. 3), że w przypadku nielosowego mechanizmu powstawania braków (NMAR) otrzymano istotnie gorsze wyniki klasyfikacji w porównaniu do mechanizmów losowych (MCAR i MAR).

[1] Pominięto niektóre nieistotne statystycznie interakcje.

Tabela 2. Wyniki analizy wariancji

Czynnik	Poziom *p*
DRZEWO (zastosowany algorytm budowy drzewa)	0,5601
MDM (mechanizm powstawania braków danych)	**0,0000**
MV (odsetek braków danych)	**0,0000**
DRZEWO*MDM	0,1137
DRZEWO*MV	0,9473
MDM*MV	0,0509
DRZEWO*MDM*MV	0,9589
METODA (sposób imputacji brakujących wartości)	**0,0000**
METODA*DRZEWO	**0,0021**
METODA*MDM	**0,0000**
METODA*MV	0,1676
METODA*DRZEWO*MDM	**0,0000**
METODA*DRZEWO*MV	0,9003
METODA*MDM*MV	0,9990
METODA*DRZEWO*MDM*MV	1,0000

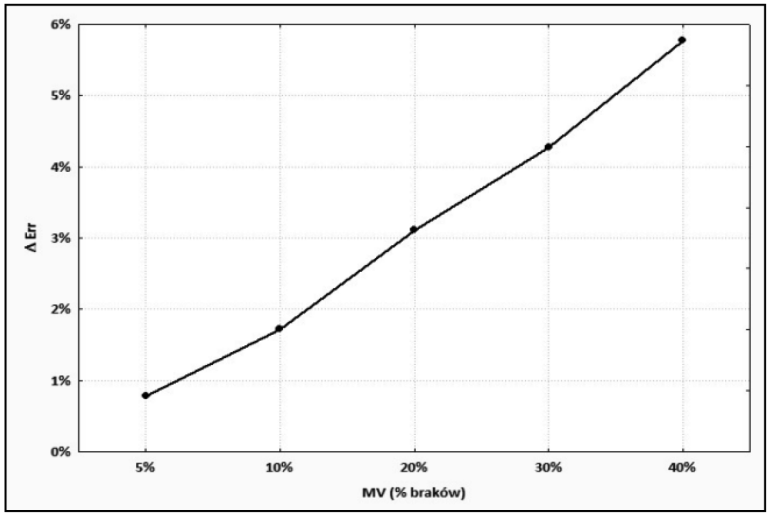
Źródło: obliczenia własne.



Rys. 3. Ocena wpływu mechanizmu powstawania braków

Źródło: obliczenia własne.

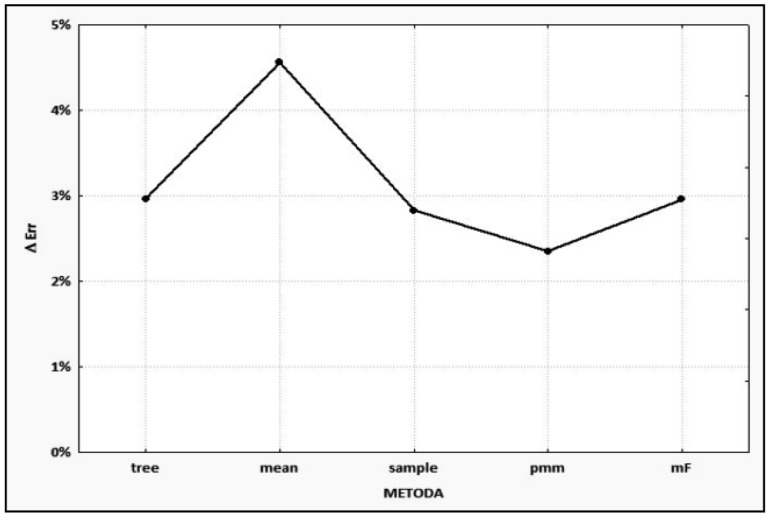
Błąd klasyfikacji rośnie ze wzrostem odsetka brakujących wartości w zbiorze danych (rys. 4), istotne różnice nie występują tylko między wynikami dla 5, 10 i 20% braków danych.



Rys. 4. Ocena wpływu odsetka brakujących danych

Źródło: obliczenia własne

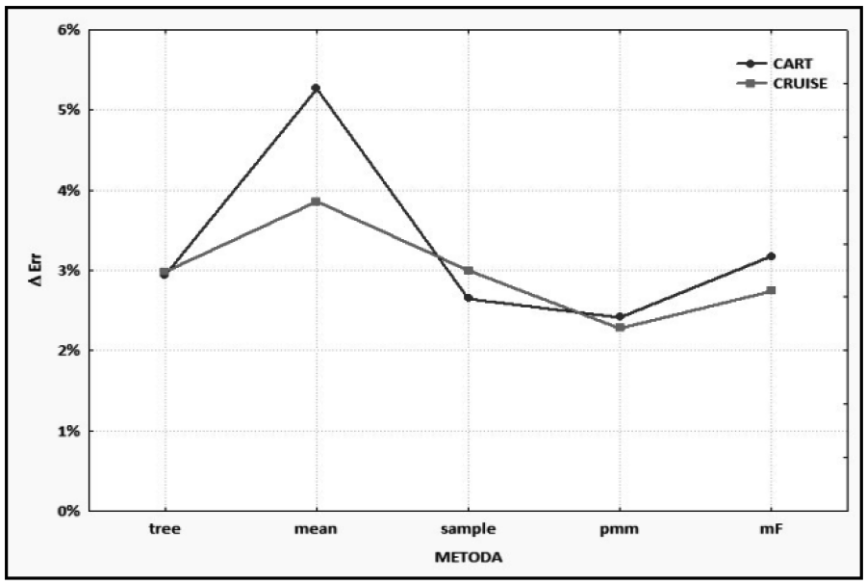
Wybór metody uzupełniania brakujących wartości wpływa na dokładność klasyfikacji (rys. 5). Wyniki uzyskane w przypadku zastosowania metody zastępowania średnią (*mean*) są istotnie gorsze od wyników dla pozostałych metod.



Rys. 5. Ocena wpływu metody uzupełniania brakujących wartości

Źródło: obliczenia własne.

W przypadku interakcji algorytmu budowy drzewa i metody imputacji braków danych (rys. 6) największe błędy klasyfikacji występują dla zastępowania średnią, przy czym dla algorytmu CART wyniki są zdecydowanie gorsze niż dla algorytmu CRUISE. Dla drzewa CART każda z metod uzupełniania brakujących wartości daje istotnie niższe błędy klasyfikacji niż procedura zastępowania średnią. Dla drzewa CRUISE z kolei istotną przewagę nad zastępowaniem średnią mają dwie metody – *predictive mean matching* (pmm) oraz *missForest* (mF).



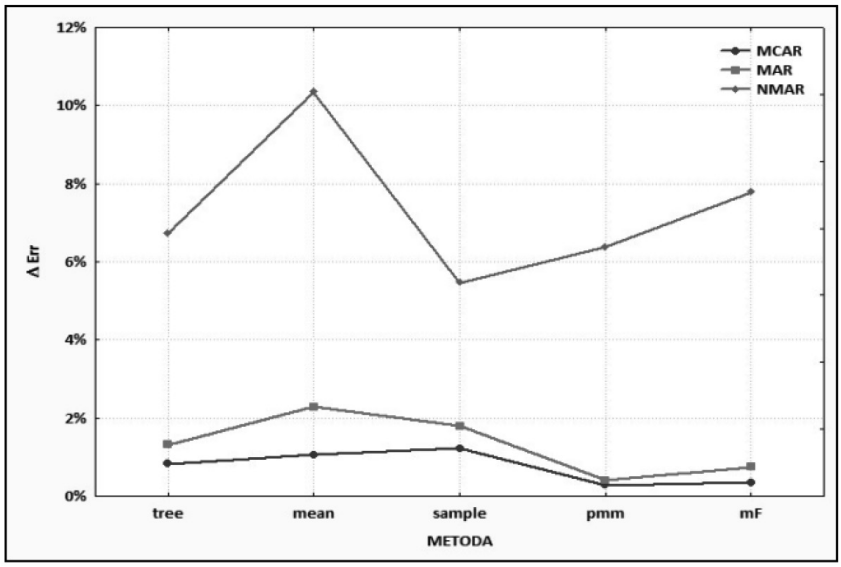
Rys. 6. Ocena wpływu interakcji algorytmu budowy drzewa i metody uzupełniania brakujących wartości

Źródło: obliczenia własne.

Analizując interakcję mechanizmu powstawania braków danych i metody postępowania (rys. 7), można zauważyć, że w sytuacji braków nielosowych (NMAR) następuje istotne pogorszenie dokładności klasyfikacji w porównaniu do mechanizmów losowych (MCAR i MAR).

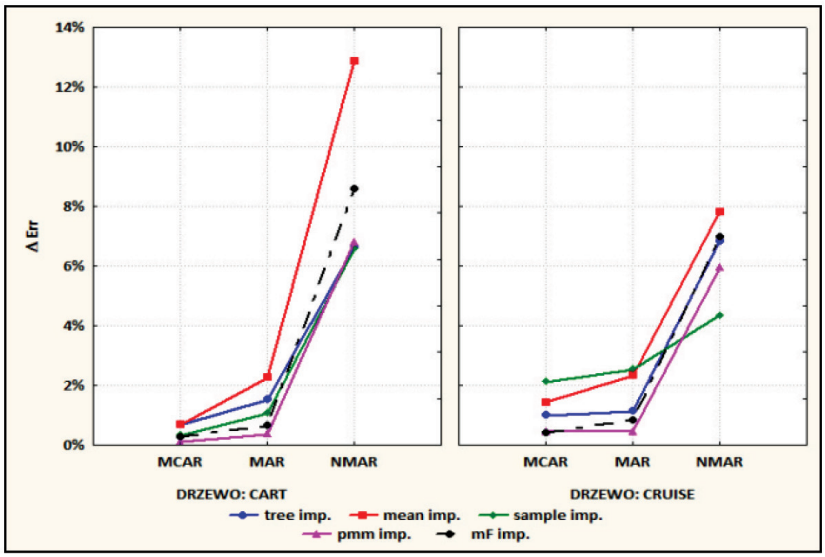
Badając interakcję algorytmu budowy drzewa, mechanizmu powstawania braków danych oraz metody uzupełniania brakujących wartości (rys. 8), stwierdzić należy, że największe błędy klasyfikacji występują przy brakach nielosowych (NMAR).

Dla algorytmu CART wszystkie metody uzupełniania brakujących wartości dają gorsze wyniki w sytuacji nielosowego mechanizmu powstawania braków niż dla mechanizmów losowych. Dodatkowo przy brakach nielosowych błąd klasyfikacji dla metody imputacji braków z wykorzystaniem wartości średniej (*mean*) jest istotnie wyższy w porównaniu do pozostałych metod imputacji.



Rys. 7. Ocena wpływu interakcji mechanizmu powstawania braków i metody uzupełniania brakujących wartości

Źródło: obliczenia własne.



Rys. 8. Ocena wpływu interakcji algorytmu budowy drzewa, mechanizmu powstawania braków i metody uzupełniania brakujących wartości

Źródło: obliczenia własne.

Podobnie dla algorytmu CRUISE przy nielosowym mechanizmie powstawania braków błędy klasyfikacji są wyższe w porównaniu do mechanizmów losowych dla każdej metody uzupełniania brakujących wartości poza zastępowaniem braku wartością wylosowaną spośród wartości obserwowanych (*sample*). Dodatkowo przy brakach nielosowych błąd klasyfikacji dla metody imputacji braków z wykorzystaniem wartości wylosowanej (*sample*) jest istotnie niższy w porównaniu do imputacji średnią (*mean*), metodą *missForest* (*mF*) oraz procedury zaimplementowanej w samym algorytmie CRUISE.

Przy obu algorytmach budowy drzewa najgorsze wyniki związane są z zastępowaniem braków wartością średnią (*mean*), przy czym dla mechanizmu nielosowego powstawania braków błąd klasyfikacji dla drzewa CART zdecydowanie przewyższa błąd klasyfikacji dla drzewa CRUISE.

5. Uwagi końcowe

Ze względu na niewielką liczbę zbiorów danych uwzględnionych w analizach przeprowadzone badanie należy uznać za wstępne i będące próbą oceny sensowności prowadzenia tego typu analiz.

Analizując uzyskane wyniki, można zauważyć wpływ mechanizmu powstawania braków danych na otrzymane rezultaty klasyfikacji. Przy niewielkiej liczbie brakujących wartości (5-10%) wszystkie sposoby postępowania dają podobne wyniki. Przy większej liczbie braków (≥20%) zaobserwowano niewielką przewagę wykorzystania imputacji metodą *predictive mean matching* nad pozostałymi metodami. Trudno wskazać zwycięzcę wśród stosowanych metod uzupełniania brakujących wartości; najgorsze wyniki uzyskano dla metody zastępowania braków wartością średnią (*mean*). Wreszcie warto zauważyć, że w przypadku braków nielosowych (NMAR) algorytm CRUISE mniej obciąża wyniki klasyfikacji.

Prowadzone badania należałoby rozszerzyć uwzględniając większą liczbę zbiorów danych, zbiory o większej liczbie obserwacji oraz zbiory ze zmiennymi mierzonymi na różnych skalach pomiaru. Dodatkowo interesującym problemem byłaby weryfikacja hipotezy postawionej przez Loha i Kima [2001], którzy wskazują na obciążenie algorytmu CART w sytuacji występowania braków danych – zmienna z dużym odsetkiem brakujących wartości ma mniejsze szanse pełnić funkcję zmiennej zastępczej.

Literatura

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.

Breiman L., *Random forests*, "Machine Learning" 2001, vol. 45, no. 1, p. 5-32.
Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning. Data Mining, Inference and Prediction*, Springer, New York 2008.
Kim H., Loh W.-Y., *Classification trees with unbiased multiway splits*, "Journal of American Statistical Association" 2001, vol. 96, p. 598-604.
Little R. J. A., Rubin D. B., *Statistical Analysis with Missing Data*, Second Edition, Wiley, New Jersey 2002.
Misztal M., *Wpływ wybranych metod uzupełniania brakujących danych na wyniki klasyfikacji obiektów z wykorzystaniem drzew klasyfikacyjnych w przypadku zbiorów danych o niewielkiej liczebności – ocena symulacyjna*, [w:] Taksonomia 19, *Klasyfikacja i analiza danych – teoria i zastosowania*, red. K. Jajuga, M. Walesiak, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 242, Wydawnictwo UE we Wrocławiu, Wrocław, 2012, s. 371-379.
Stekhoven D.J., Bühlmann P., *MissForest – Nonparametric Missing Value Imputation for Mixed-Type Data*, "Bioinformatics" 2012, vol. 28, no. 1, p. 112-118.
Song Q., Shepperd M., Chen X., Liu J., *Can k-NN imputation improve the performance of C4.5 with small software project data sets? A comparative evaluation*, "Journal of System and Software" 2008, vol. 81, no. 12, p. 2361-2370.
Twala B., *An empirical comparison of techniques for handling incomplete data using decision trees*, "Applied Artificial Intelligence" 2009, vol. 23, p. 373-405.
Twala B., Jones M. C., Hand D. J., *Good methods for coping with missing data in decision trees,* "Pattern Recognition Letters" 2008, vol. 29, no. 7, p. 950-956.

ASSESSMENT OF THE INFLUENCE OF SELECTED IMPUTATION METHODS ON THE RESULTS OF OBJECT CLASSIFICATION USING CLASSIFICATION TREES

Summary: In contrast with most multivariate statistical analysis methods, classification tree is an example of the learning algorithm coping with missing values in special, original way. In the paper the influence of some selected missing data techniques on the results of object classification using CART and CRUISE classification trees was assessed. All the procedures were compared by artificially simulating different proportions and mechanisms of missing data using complete data sets mainly from the UCI repository of machine learning databases.

Keywords: missing values, imputation, classification trees.