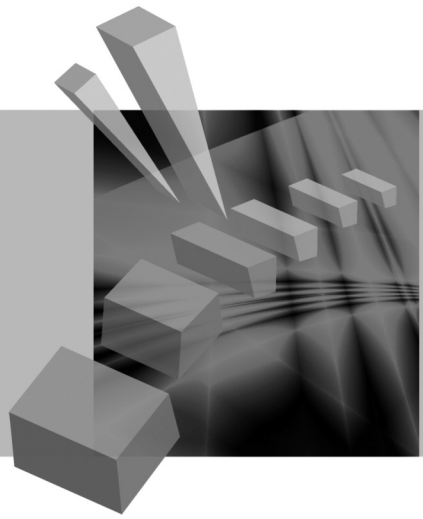
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

242

Taksonomia 19.

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2012

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

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

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

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
oraz w The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon http://kangur.uek.krakow.pl/
bazy_ae/bazekon/nowy/index.php

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

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

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM
Nakład: 320 egz.

Spis treści

Summaries

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

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

Małgorzata Misztal
Uniwersytet Łódzki

WPŁYW WYBRANYCH METOD UZUPEŁNIANIA 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
40	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

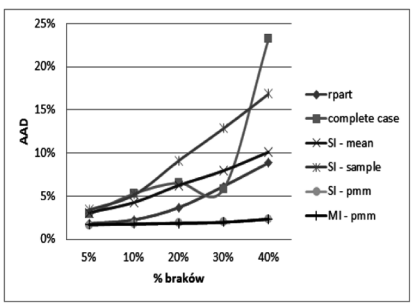
Na podstawie uzyskanych wyników obliczono odchylenia od wartości „wzorcowej" – błędu klasyfikacji dla oryginalnego zbioru danych:

$$| Err_I - Err_C |,$$

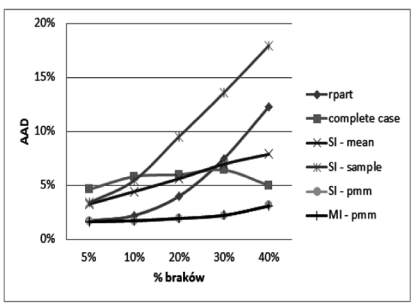
gdzie: Err_I – błąd klasyfikacji dla zbioru danych z uzupełnionymi brakami danych;
Err_C – błąd klasyfikacji dla kompletnego, oryginalnego zbioru danych.

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.

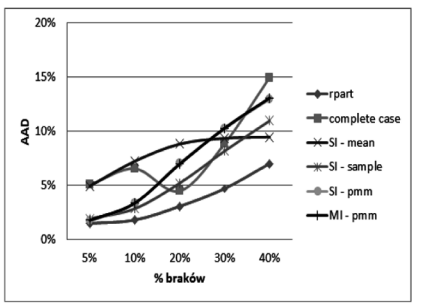
Jak widać na rysunkach, w przypadku mechanizmów powstawania braków danych typu MCAR i MAR i przy niewielkim odsetku braków danych (5-10%) zarówno drzewo klasyfikacyjne CART (*rpart*), zastosowane do zbiorów danych z brakującymi wartościami, jak i imputacja metodą *predictive mean matching* przed budową 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.

Odmienna 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 trójczynnikową analizę wariancji (ANOVA). Jej wyniki podsumowano w tab. 2 oraz na rys. 5-9[1]. Jak wynika z tab. 2, wszystkie efekty główne oraz trzy efekty interakcji są istotne statystycznie.

Tabela 2. Wyniki analizy wariancji

Czynnik	Poziom *p*
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.

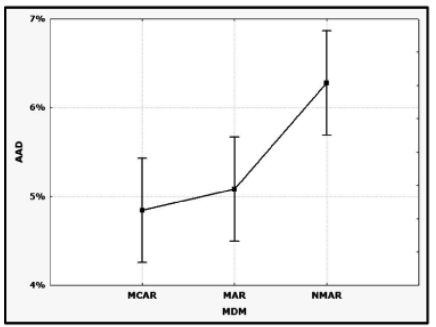
Wybór metody postępowania z brakami danych wpływa na jakość klasyfikacji (rys. 7). Wyniki uzyskane w przypadku zastosowania imputacji typu *hot – deck* (SI – *sample*) oraz zastępowania średnią (*SI – mean*) są istotnie gorsze od wyników dla pozostałych metod.

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.

---

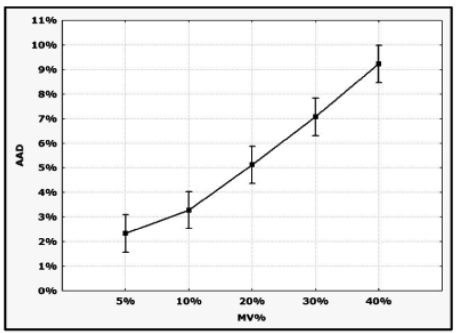
[1] Pominięto prezentację graficzną dla interakcji trzech czynników.

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



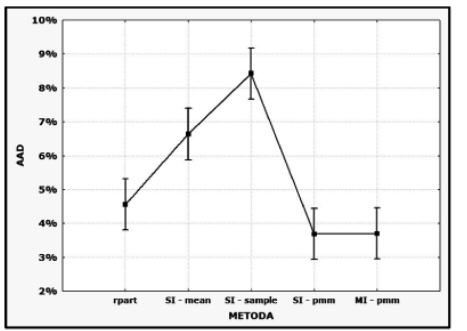
Rys. 5. Ocena wpływu mechanizmu powstawania braków

Źródło: obliczenia własne.



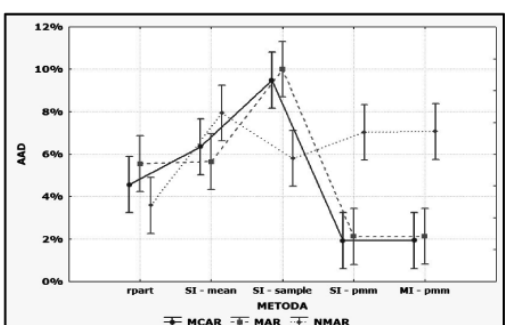
Rys. 6. Ocena wpływu odsetka brakujących danych

Źródło: obliczenia własne.



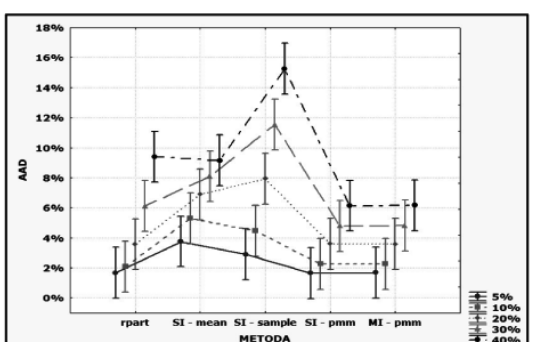
Rys. 7. Ocena wpływu metody postępowania z brakami danych

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. wzorzec 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-Tsechansky 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.