PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 176
Taksonomia 18. Klasyfikacja i analiza danych – teoria i zastosowania 2011

Małgorzata Misztal
Uniwersytet Łódzki

PRÓBA OCENY WPŁYWU WYBRANYCH METOD IMPUTACJI DANYCH NA WYNIKI KLASYFIKACJI OBIEKTÓW Z WYKORZYSTANIEM DRZEW KLASYFIKACYJNYCH*

Streszczenie: W praktycznych zastosowaniach metod statystycznych często pojawia się problem występowania w zbiorach danych brakujących wartości. W takiej sytuacji wymienić można trzy sposoby postępowania: (1) odrzucenie obiektów z wartościami brakującymi, (2) wykorzystanie algorytmu uczącego do rozwiązania problemu brakujących wartości w fazie uczenia, (3) imputację brakujących wartości przed zastosowaniem algorytmu uczącego. Celem głównym pracy jest ocena wpływu wymienionych metod na wyniki klasyfikacji obiektów za pomocą drzew klasyfikacyjnych w sytuacji występowania braków danych.

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

1. Wstęp

W praktycznych zastosowaniach metod statystycznych często pojawia się problem występowania w zbiorach danych brakujących wartości. Uniemożliwia to uogólnienie wyników na całą badaną zbiorowość. Zbyt duże braki danych obciążają końcowe wyniki analiz poprzez zmniejszenie rozmiaru próby, zwiększenie wariancji estymatorów, zniekształcenie rozkładów badanych zmiennych, a także zmniejszają zaufanie odbiorców do wyników badania i negatywnie wpływają na jakość danych statystycznych.

Imputacja jest metodą polegającą na zastąpieniu brakujących danych konkretnymi wartościami w celu uzyskania kompletnego zbioru danych [Paradysz, Szym-

* Pracę wykonano w ramach realizacji tematu badawczego nr 505/587/R pt. *Metody imputacji i ich zastosowania*, finansowanego z dotacji na badania własne w UŁ w roku 2010.

kowiak 2007]. W rezultacie każdej jednostce w miejsce brakujących danych przypisywana jest jakaś wartość.

Dokonując imputacji danych, należy pamiętać, że nie powinna ona prowadzić do obciążeń lub zmian rozkładów cech w zbiorze danych oraz do wzrostu wariancji stosowanych estymatorów. Proces imputacji w większym stopniu powinien być uzależniony od danych pochodzących z próby, niż odwoływać się do założeń co do natury brakujących danych, a do oszacowań uzyskanych na podstawie imputowanych danych należy podchodzić ostrożnie.

Little i Rubin [2002] dzielą mechanizmy powstawania braków danych na ignorowalne i nieignorowalne.

Wśród mechanizmów ignorowalnych wymienić można dane typu MCAR – *Missing Completely At Random* oraz typu MAR – *Missing At Random*.

W przypadku danych typu MCAR brakujące wartości rozłożone są losowo wśród wszystkich wartości − prawdopodobieństwo wystąpienia brakującej wartości dla zmiennej Y nie zależy od wartości samej zmiennej Y 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 Y nie zależy od wartości samej zmiennej Y, tylko od obserwowanych wartości pozostałych zmiennych w zbiorze danych.

W przypadku mechanizmu nieignorowalnego − NMAR – *Not Missing At Random* − brakujące wartości danej zmiennej zależą od czynników czy zdarzeń, których badacz nie jest w stanie zmierzyć.

Przy założeniu, że mechanizm powstawania brakujących wartości jest ignorowalny, Hastie, Tibshirani i Friedman [2008] wymieniają trzy sposoby postępowania w sytuacji występowania braków danych:

– odrzucenie obiektów z wartościami brakującymi;
– wykorzystanie algorytmu uczącego do rozwiązania problemu brakujących wartości w fazie uczenia;
– imputację brakujących wartości przed zastosowaniem algorytmu uczącego.

Celem głównym pracy jest ocena wpływu metod postępowania w sytuacji braku danych na wyniki klasyfikacji obiektów za pomocą drzew klasyfikacyjnych. Celem szczegółowym natomiast jest znalezienie odpowiedzi na pytanie, czy imputacja brakujących wartości przed zastosowaniem metody rekurencyjnego podziału nie da dokładniejszych wyników (w sensie dokładności predykcji) niż przyjęte w algorytmach drzew klasyfikacyjnych metody postępowania z brakami danych.

2. Metody postępowania w sytuacji braku danych

Tradycyjną metodą stosowaną w przypadku zbiorów danych z wartościami brakującymi jest odrzucenie obiektów (lub zmiennych) z choćby jedną brakującą wartością. Analizę przeprowadza się wówczas wyłącznie z wykorzystaniem obiektów, dla których mamy pełne informacje (tzw. *complete case analysis*).

Jeśli dane są typu MCAR, to zredukowana próba będzie podpróbą losową z oryginalnej próby, a zatem jeśli estymatory byłyby nieobciążone dla całej próby (bez braków danych), to będą również nieobciążone dla takiej próby zredukowanej. Z drugiej strony, ponieważ zmniejsza się liczebność próby, uzyskuje się większe błędy standardowe oraz mniejszą istotność testów.

Jeśli dane są typu MAR, uzyskamy obciążone estymatory. Może się także zdarzyć, że pozostałe po usunięciu obiekty przestaną być reprezentatywne.

Kolejnym sposobem postępowania z brakami danych jest wykorzystanie algorytmu uczącego do rozwiązania problemu brakujących wartości w fazie uczenia. Podejście to dotyczy stosunkowo niewielkiej liczby algorytmów uczących, przede wszystkich opartych na metodzie rekurencyjnego podziału.

W algorytmach tworzących drzewa klasyfikacyjne wypracowano odrębne sposoby postępowania w sytuacji braku danych, zarówno w zbiorze uczącym, jak i w zbiorze testowym. Najprostsze z nich to potraktowanie wartości brakującej jako kolejnej kategorii danej zmiennej lub zastąpienie brakujących obserwacji danej zmiennej odpowiednio dobranymi wartościami – np. w algorytmie QUEST zastępuje się brakujące wartości odpowiednią współrzędną środka ciężkości klasy, która leży najbliżej danego obiektu. Oryginalne procedury zastosowano w algorytmach CART [Breiman i in. 1984] i CRUISE [Kim, Loh 2001].

W algorytmie CART 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 brakuje danych, to uwzględniana jest następna zmienna zastępcza, itd.

W przypadku algorytmu CRUISE wybór zmiennej do podziału oparty jest wyłącznie na dostępnych wartościach danej zmiennej w tym węźle. Brakujące wartości są zastępowane wartością średnią lub modalną dla każdej klasy w węźle. Następnie dokonywany jest podział obiektów w węźle i imputowane wartości zostają usunięte.

Ostatnim sposobem postępowania w sytuacji braku danych jest imputacja brakujących wartości przed zastosowaniem algorytmu uczącego.

Wyróżnić można imputację jednostkową (*single imputation*), w której uzupełnianie brakujących wartości przeprowadza się tylko raz, oraz imputację wielokrotną (*multiple imputation*), w której brakujące dane są uzupełniane *m* razy. Analizę statystyczną przeprowadza się dla każdego z *m* uzyskanych kompletnych zbiorów, a następnie uzyskane wyniki łączy się w jeden końcowy wynik.

W pakietach statystycznych najczęściej spotyka się propozycję zastąpienia brakującej wartości średnią lub dominantą (*mean/mode imputation*) obliczoną na podstawie wszystkich dostępnych danych lub danych dotyczących logicznie wyodrębnionych segmentów (w grupach dochodu itp.). Metoda ta zmniejsza wariancję zmiennych i pomija relacje między nimi. Dodatkowo − jeśli dane nie są typu MCAR − można uzyskać średnie niereprezentatywne dla zbiorowości.

Kolejną stosowaną metodą jest imputacja regresyjna (*regression imputation*) – w sytuacji, gdy zmienna z brakującymi wartościami jest skorelowana z innymi zmiennymi, wykorzystuje się modele regresji wielu zmiennych i zastępuje brakujące wartości wartościami uzyskanymi z modelu regresji.

W przypadku imputacji wielokrotnej najczęściej wykorzystywane metody to imputacja typu *hot deck* (*hot deck imputation*) oraz imputacja metodą *predictive mean matching*. W imputacji typu *hot deck* brakujące wartości zastępowane są wartościami wylosowanymi z powtórzeniami spośród dostępnych, obserwowanych wartości. Natomiast w imputacji typu *predictive mean matching* wykorzystywany jest model regresji dla zmiennej z brakami danych względem pozostałych zmiennych. Następnie na podstawie tego modelu dokonuje się predykcji wartości zarówno dla brakujących, jak i kompletnych danych, szuka się zestawu przypadków kompletnych, dla których wartości oszacowane z linii regresji są zbliżone do wartości uzyskanej dla obiektu z brakującą wartością i z tego zestawu losuje się jeden, wykorzystywany do uzupełnienia brakującej wartości.

Inne metody imputacji szczegółowo omawiają np. Little i Rubin [2002].

3. Wyniki badań

Wpływ metod postępowania w sytuacji występowania braków danych na wyniki klasyfikacji obiektów za pomocą drzew klasyfikacyjnych oceniono z wykorzystaniem 10 zbiorów danych empirycznych pochodzących z repozytorium baz danych na Uniwersytecie Kalifornijskim w Irvine (UCI) (por. [Blake i in. 1988]) oraz z badań własnych.

Podstawowe informacje dotyczące wykorzystanych zbiorów danych zawiera tab. 1. Liczba klas jest wszędzie taka sama (równa 2), zbiory różnią się liczbą obiektów (od 75 do 16 438) i zmiennych (od 5 do 72), a przede wszystkim odsetkiem brakujących wartości (od 0,25% do 10,19%).

Tabela 1. Charakterystyka wykorzystanych zbiorów danych

Nazwa (id)	Źródło	Liczba obiektów	Liczba zmiennych objaśniających	Typ zmiennych	Odsetek brakujących wartości	Odsetek obiektów z brakującymi wartościami
Wisconsin Breast Cancer Database (wbc)	UCI	699	9	mierzalne	0,25	2,29
Mammographic Mass Data (mm)	UCI	961	5	mieszane	3,37	13,63
Ozone Level Detection (oz)	UCI	2 536	72	mierzalne	8,18	27,13
Census Income (cen)	UCI	16 438	11	mieszane	0,53	5,81
Hepatitis Domain (hep)	UCI	155	19	mieszane	5,67	48,39
1984 United States Congressional Voting Records Database (vot)	UCI	435	16	niemierzalne	4,14	46,67
Credit Approval (cred)	UCI	690	14	mieszane	0,63	5,36
Lung Cancer (lung)	B. wł.	75	14	mieszane	10,19	80,00
ICU Stay (icu)	B. wł.	337	14	mieszane	0,30	3,26
CABG (cabg)	B. wł.	1 121	19	mieszane	2,14	15,17

Źródło: opracowanie własne.

Dla każdego zbioru danych:

1) zbudowano drzewa klasyfikacyjne CART i CRUISE w oparciu wyłącznie o zestaw obiektów bez brakujących wartości (*complete case analysis*);

2) zbudowano drzewa klasyfikacyjne CART i CRUISE w oparciu o wszystkie dane, także z brakującymi wartościami, wykorzystując metody zaimplementowane w algorytmach;

3) wykorzystano metody imputacji wielokrotnej (MI) przed zbudowaniem drzewa klasyfikacyjnego:

– brakujące wartości uzupełniano metodą *predictive mean matching* (imputationMethod=”pmm”) oraz z wykorzystaniem imputacji typu *hot deck* (imputationMethod=”sample”);
– procedurę uzupełniania brakujących wartości przeprowadzono $m = 5$ razy;
– dla każdego z uzyskanych kompletnych zbiorów danych zbudowano drzewo klasyfikacyjne CART i CRUISE;
– wyniki zagregowano w jeden model, przydzielając obiekt do klasy, która była najczęściej wybierana przez kolejne drzewa.

Dla wszystkich zbiorów danych i dla każdej z zastosowanych metod obliczono błędy klasyfikacji metodą 10-częściowego sprawdzania krzyżowego (10-fold-CV).

Obliczenia wykonano z wykorzystaniem środowiska `R` − pakiety: `rpart`, `ipred` i `mice` (imputacja wielokrotna (por. [Buuren, Groothuis-Oudshoorn 2010]))

oraz udostępnionego przez autorów programu budującego drzewa klasyfikacyjne CRUISE [www.stat.wisc.edu/~loh/cruise].

Wyniki obliczeń podsumowano w tab. 2. W ostatnim wierszu podano uśredniony błąd klasyfikacji dla każdej z analizowanych metod.

Tabela 2. Błędy klasyfikacji (w %) w 10-częściowym sprawdzaniu krzyżowym dla analizowanych zbiorów danych

Zbiór	Metoda rpart			Metoda CRUISE			Complete case analysis	
	klasyczna	MI (pmm)	MI (sample)	klasyczna	MI (pmm)	MI (sample)	rpart	CRUISE
wbc	6,87	6,90	7,32	5,01	5,04	5,38	5,42	5,27
mm	17,59	18,31	18,11	17,90	17,75	17,32	15,78	15,78
oz	3,63	3,79	3,69	20,69	21,67	21,57	4,17	32,14
cen	15,33	15,33	15,33	15,16	15,02	15,05	15,97	15,36
hep	21,94	18,45	21,03	27,48	21,10	24,38	20,00	12,50
vot	4,61	4,05	4,83	4,37	3,08	4,18	3,02	3,02
cred	14,06	14,87	14,90	13,62	13,51	13,85	15,01	13,63
lung	37,33	35,73	40,80	40,00	33,33	39,47	93,33	40,00
icu	45,70	45,87	46,05	40,38	39,94	40,24	38,96	35,89
cabg	31,85	31,60	31,60	40,91	41,12	40,20	31,97	42,92
Średni wynik	19,89	19,49	20,37	22,55	21,16	22,16	24,36	21,65

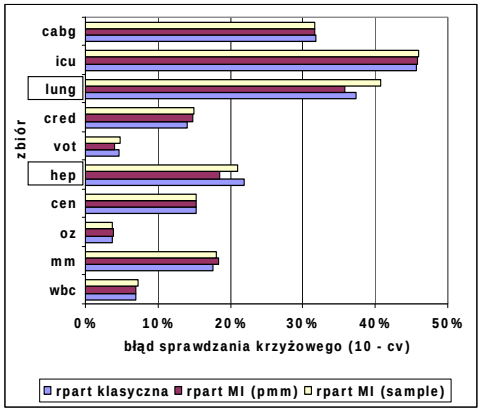
Źródło: obliczenia własne.

Dla większości zbiorów danych, uwzględniając tylko pełne obserwacje, uzyskujemy niższe błędy w sprawdzaniu krzyżowym, zarówno w porównaniu z tradycyjnymi algorytmami, jak i metodami imputacji wielokrotnej. Ze względu na znaczne zmniejszenie liczebności zbiorów są to wyniki obciążone (optymistycznie zaniżone).

Analizując wyniki uśrednione, daje się zauważyć niewielki spadek wartości błędu klasyfikacji w przypadku imputacji wielokrotnej metodą *predictive mean matching* w stosunku do klasycznych algorytmów CART i CRUISE. Jednakże nie są to różnice aż tak znaczące, żeby można było rekomendować imputację brakujących wartości przed zastosowaniem algorytmu uczącego – niewielka poprawa wyników nie rekompensuje kosztów zastosowania imputacji wielokrotnej (pracochłonność obliczeń, dłuższy czas obliczeń itp.).

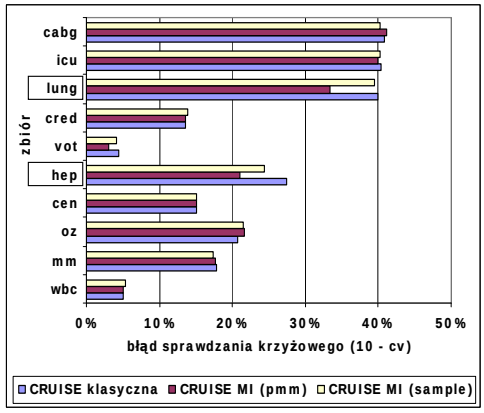
Na rysunkach 1 i 2 porównano wyniki zastosowania klasycznych algorytmów CART i CRUISE z wynikami uzyskanymi po zastosowaniu metod imputacji wielokrotnej dla każdego z 10 analizowanych zbiorów danych.

Szczegółowa analiza wyników pokazuje, że w przypadku dwóch zbiorów danych – Hepatitis Domain oraz Lung Cancer – zastosowanie imputacji wielokrotnej metodą *predictive mean matching* (pmm) – zmniejsza wyraźnie odsetek błędnych klasyfikacji.



Rys. 1. Błędy klasyfikacji (10-fold-CV; CART)

Źródło: opracowanie własne.



Rys. 2. Błędy klasyfikacji (10-fold-CV; CRUISE)

Źródło: opracowanie własne.

Oba zbiory charakteryzują się niewielką liczbą obserwacji oraz wysokim odsetkiem braków danych.

4. Uwagi końcowe

Tradycyjna metoda postępowania z brakami danych, polegająca na usuwaniu wszystkich tych obiektów, dla których występuje co najmniej jedna wartość brakująca, może znacznie zredukować liczebność próby, a także obciąża wyniki, jeśli pozostałe obserwacje nie są reprezentatywne. Należy o tym pamiętać, korzystając bezkrytycznie z metod proponowanych w komercyjnych pakietach statystycznych.

Wśród zalet drzew klasyfikacyjnych wymienia się między innymi możliwość korzystania w analizach ze zbiorów danych z wartościami brakującymi. Zaimplementowane w algorytmach drzew klasyfikacyjnych techniki postępowania w przypadku braku danych zwalniają badacza z konieczności samodzielnego uzupełniania brakujących wartości.

Imputacja brakujących wartości przed zastosowaniem metody rekurencyjnego podziału nie poprawia znacząco wyników klasyfikacji. Nieznaczna redukcja błędu klasyfikacji nie musi być wynikiem zastosowanej imputacji wielokrotnej, może wynikać również z agregacji drzew klasyfikacyjnych zbudowanych dla *m* kompletnych zbiorów danych.

Analizując wyniki szczegółowe, można postawić hipotezę, że w przypadku zbiorów danych z małą liczbą obiektów oraz wysokim odsetkiem braków danych zastosowanie metod imputacji przed zastosowaniem metody rekurencyjnego podziału poprawi jakość klasyfikacji (w sensie redukcji błędu klasyfikacji).

Wayman [2003] podkreśla, że metody imputacji wielokrotnej są odporne na odstępstwa od normalności rozkładu, a także dają dobre wyniki właśnie przy małej liczebności próby i wysokim odsetku brakujących danych.

Potwierdzenie tego przypuszczenia wymaga dalszych badań.

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.

Buuren S. van, Groothuis-Oudshoorn K., *MICE: Mulitvariate Imputation by Chained Equations in R*, „Journal of Statistical Software" 2010 (w druku).

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, 96, s. 598-604.

Little R.J.A., Rubin D.B., *Statistical Analysis with Missing Data*, Wiley, New Jersey 2002.

Paradysz J., Szymkowiak M., *Źródła danych ludnościowych*, „Metodologia Badań Demograficznych", Zeszyt nr 15 Sekcji Analiz Demograficznych, KND PAN, Warszawa 2007, 7-26.

Wayman J.C., *Multiple Imputation for Missing Data: What Is It and How Can I Use It?*, http://www.csos.jhu.edu/contact/staff/jwayman_pub/wayman_multimp_aera2003.pdf, 2003.

Źródło internetowe

www.stat.wisc.edu/~loh/cruise.

AN ATTEMPT TO ASSESS THE INFLUENCE OF SELECTED IMPUTATION METHODS ON THE CLASSIFICATION OF OBJECTS BASED ON CLASSIFICATION TREES

Summary: Incomplete data are quite common in practical applications of statistical methods. Dealing with data sets with missing values we can: (1) discard observations with missing values, (2) rely on the learning algorithm to deal with missing values in training phase or (3) impute all missing values before training. The main goal of the paper is to assess the influence of these strategies on the results of object classification by means of classification trees in the case of incomplete data.