PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1076 2005
TAKSONOMIA 12
Klasyfikacja i analiza danych – teoria i zastosowania

Mariusz Grabowski
Akademia Ekonomiczna w Krakowie

EFEKTYWNOŚĆ WYBRANYCH METOD UZUPEŁNIANIA INFORMACJI NIEKOMPLETNYCH A RÓŻNE MECHANIZMY POWSTAWANIA BRAKÓW DANYCH

1. Wstęp

Niekompletność danych, będących przedmiotem analizy, może stwarzać problemy wielu klasycznym metodom analizy danych. Problem ten, mimo istnienia wielu publikacji, nie został jednoznacznie rozwiązany [Grabiński, Wydmus, Zeliaś 1979; Little, Rubin 2002; Little 1988; Kordos 1998; Pawełek 1996; Shafer 1997].

Wyróżniamy dwa rodzaje badań analizy danych niekompletnych. Pierwszy z nich rozwija metody odporne na wystąpienie danych niekompletnych. Uważa się, że braki danych są integralną częścią zbioru danych, mają bowiem taki sam wpływ na zasób informacji jak dane istniejące. Z tego względu ich szacowanie uważa się za niewłaściwe – w myśl zasady: brak danych jest również informacją. Ten rodzaj badań kładzie silny nacisk na możliwość tworzenia takich narzędzi analizy danych, które dają statystycznie prawidłowe mechanizmy wnioskowania o parametrach rozkładów zmiennych określonych dla badanej zbiorowości. Drugi rodzaj jest reprezentowany przez ogólnie rozumiane metody podstawiania czy uzupełniania danych brakujących. Jego celem jest rozwój metod pozwalających na uzupełnienie danych w taki sposób, aby rozkład prawdopodobieństwa poszczególnych zmiennych z uzupełnionymi brakującymi danymi był zgodny z rozkładem dla określonych zmiennych w populacji.

Metody szacowania danych niekompletnych można również podzielić na dwie grupy: parametryczne – tzn. metody zakładające postać rozkładu w zbiorowości, z której pochodzi próbka – oraz metody nieparametryczne.

W niniejszym artykule zostaną porównane trzy metody: metoda MI (*multiple imputation*), metoda SI-EM (*single imputation*), wykorzystująca algorytm EM, oraz metoda SOM (*self-organized map*). Dwie pierwsze należą do metod parame-

trycznych, zakładających, że próbka pochodzi z populacji o wielowymiarowym rozkładzie normalnym, trzecia – do metod nieparametrycznych, w których nie zakłada się postaci rozkładu. Jako implementację procedury MI wykorzystano algorytm zawarty w pakiecie analitycznym SAS, procedury SI-EM – moduł MVA z pakietu SPSS, z kolei implementacja SOM pochodzi z pakietu opracowanego i udostępnionego przez zespół kierowany przez autora metody [Kohonen, Hynninen, Kangas 1995]. Artykuł ten stanowi kontynuację badań autora dotyczących oceny właściwości metody SOM w analizie danych niekompletnych o charakterze przekrojowym [Grabowski 1998; 1999; 2004].

2. Mechanizmy powstawania braków danych

W pracy [Little, Rubin 2002; Little 1988] zdefiniowano trzy typy mechanizmów powstawania braków danych w zbiorach o charakterze przekrojowym.

1. **MCAR** (*missing completely at random*) – ma on charakter czysto losowy. Pojawienie się braków danych jest niezależne zarówno od obserwowalnej części danych, jak i od tej części danych, której brakuje. Oznacza to, że należy się spodziewać, iż estymowane wartości parametrów rozkładu analizowanej zmiennej, z którego pochodzi dana próbka, są takie same dla danych pełnych i brakujących.
2. **MAR** (*missing at random*) – w mechanizmie tym braki danych zależą jedynie od obserwowalnej części danych i równocześnie nie są zależne od części nieobserwowanej. MAR jest założeniem, na którym opiera się większość z metod uzupełniania danych niekompletnych.
3. **NMAR** (*not missing at random, non-ignorable*) – w przeciwieństwie do MCAR i MAR, mechanizm tennie jest losowy. Braki danych zależą od wartości zmiennej, której te braki dotyczą. W NMAR prawdopodobieństwo wystąpienia braku danych jest znacznie większe dla pewnych wartości tej zmiennej. Na przykład, w badaniach społeczno-ekonomicznych, dotyczących dochodów ludności, dość powszechnym zjawiskiem jest zatajanie dochodów, gdy przekroczą one pewną wartość. NMAR to niewątpliwie najtrudniejszy mechanizm z metodologicznego punktu widzenia i nie ma jeszcze metod dających tutaj satysfakcjonujące rozwiązania.

Należy jednak zaznaczyć, że rozważania te mają charakter teoretyczny. Nie ma bowiem spójnego zestawu testów statystycznych, pozwalających ocenić charakter mechanizmu powstawania braków danych, tj. zakwalifikowania określonego zbioru danych do jednej z trzech grup: MCAR, MAR lub NMAR. Jedyny test, opisany w pracy [Little 1988], pozwala jedynie zweryfikować hipotezę o MCAR.

3. Test MCAR

Little [1988] zaproponował test statystyczny pozwalający zweryfikować hipotezy o wystąpieniu mechanizmu MCAR.

Macierz, która posłuży do konstrukcji statystyki testowej, składa się z p kolumn oraz J wierszy. Kolumny odpowiadają zmiennym z macierzy obserwacji, wiersze zaś stanowią wektory będące unikalnymi wzorcami brakujących danych. Liczba możliwych wzorców brakujących danych wynosi 2^p, m_j stanowi liczbę obserwacji, w których braki danych występują zgodnie z j-tym wzorcem, $\hat{\mu}$ i $\hat{\Sigma}$ są estymatorami największej wiarygodności p-wymiarowego rozkładu normalnego, bazującego na danych istniejących, $\hat{\mu}_{obs,j}$ i $\hat{\Sigma}^{-1}_{obs,j}$ są podzbiorami parametrów, odpowiadającymi obserwacjom pełnym dla wzorca j (odpowiednio: wektor o wymiarze p_j oraz macierz o wymiarach $p_j \times p_j$), a $\overline{y}_{obs,j}$ oznacza p_j-wymiarowy wektor średnich, obliczony na danych posiadających wzorzec braków j. Statystyka testowa ma postać:

$$d^2 = \sum_{j=1}^{J} d_j^2 = \sum_{j=1}^{J} m_j (\overline{y}_{obs,j} - \hat{\mu}_{obs,j}) \hat{\Sigma}^{-1}_{obs,j} (\overline{y}_{obs,j} - \hat{\mu}_{obs,j})^T. \quad (1)$$

Ma ona asymptotyczny rozkład χ^2 o liczbie stopni swobody wyrażonej wzorem:

$$\sum_{j=1}^{J} p_j - p. \quad (2)$$

Na podstawie tak zdefiniowanej statystyki weryfikujemy hipotezę:

H_0: mechanizm braków danych jest typu MCAR,

H_1: mechanizm braków danych nie jest typu MCAR.

Jeśli wartość prawdopodobieństwa testowego *p-value* nie przekroczy założonej wartości α (poziomu istotności), to należy odrzucić hipotezę H_0. Graficzną interpretację testu MCAR można znaleźć w pracy [Hesterberg 1999]. Jest ona szczególnie interesująca w przypadku odrzucenia H_0, gdyż ukazuje powody odrzucenia H_0.

4. Eksperyment symulacyjny

W celu przeprowadzenia eksperymentu symulacyjnego, mającego na celu zbadanie efektywności wymienionych metod uzupełniania danych niekompletnych w różnych mechanizmach powstawania braków danych, wygenerowano 4-wymiarowe zbiory danych przekrojowych będących próbkami z populacji o różnych rozkładach. Następnie losowo, zgodnie z odpowiednim mechanizmem powstawania braków danych, usunięto z nich część danych i zachowano dane oryginalne, aby mieć podstawę do późniejszego określenia błędu szacunku.

Po wygenerowaniu poszczególnych zbiorów testowych różnicowano je pod względem takich przypadków, jak:

- typ rozkładu – wielowymiarowy normalny oraz wielowymiarowy dwumodalny.
- procent braków danych – 3, 10 i 20%.
- mechanizmy braków danych – MCAR, MAR, NMAR.

W rezultacie otrzymano 18 zbiorów testowych. Poniżej opisano poszczególne przypadki zbiorów danych. Każdy z nich poprzedzono identyfikatorem użytym w badaniach:

- NU_4 – próbka 150 obserwacji pochodzących ze standardowego 4-wymiarowe-go rozkładu normalnego N(**0**;**I**),
- NMD24 – próbka 150 obserwacji pochodzących z mieszanki dwóch rozkładów normalnych N(**5**;**I**) oraz N(-**5**;**I**); po 75 obserwacji z każdej z podpopulacji.

Dla każdego z wymienioych przypadków wygenerowano 9 zbiorów danych (usuwano odpowiednio 3, 10 i 20% braków danych, uwzględniając mechanizmy MCAR, MAR i NMAR)

Poszczególne mechanizmy braków danych realizowano w opisany sposób:

1. Realizując mechanizm MCAR w przypadku zarówno jednomodalnym, jak i dwumodalnym, dane usuwano losowo. Korzystano z pseudolosowego generatora rozkładu równomiernego, dostarczającego par liczb z przedziału (0; 1). Następnie mnożono pierwszą liczbę przez 150 i tak otrzymaną wartość zaokrąglano do najbliższej liczby naturalnej. W ten sposób otrzymywano numer obserwacji (wiersza). Drugą liczbę mnożono przez 4, zaokrąglano do najbliższej liczby naturalnej i w ten sposób otrzymywano numer kolumny (zmiennej). Gdy wylosowany numer wiersza oraz kolumny powtarzał się, losowano następną parę wartości.

2. Mechanizm MAR realizowano następująco. Aby wygenerować braki danych w określonej zmiennej, dla każdej obserwacji obliczano długość euklidesową wektora składającego się z komponentów pozostałych trzech zmiennych (liczba zmiennych wynosiła 4 w każdym ze zbiorów testowych)[1]. Następnie dokonywano liniowego uporządkowania zbioru danych według wzrastających wartości obliczonych w opisany wyżej sposób długości. Kolejną czynnością było usunięcie danych ze zmiennej, której komponent nie wchodził w skład 3-elementowego wektora. Aby to zrobić, losowo generowano numer wiersza (podobnie jak w MCAR), jednak zbiór obserwacji ograniczano jedynie do tych, dla których długość opisanego wyżej wektora była mniejsza od średniej wartości długości dla wszystkich obserwacji.

3. W mechanizmie NMAR, aby wygenerować braki danych dla określonej zmiennej, najpierw dokonywano liniowego uporządkowania zbioru według wzrastających wartości tej zmiennej. Następnie losowo generowano numer wiersza (podobnie jak w MCAR), jednak zbiór obserwacji ograniczano do tych obserwacji, dla których wartości zmiennej, z której usuwano dane, były ujemne.

Opisane schematy generowania zbiorów zawierających dane niekompletne powtarzano dla 3, 10 i 20% braków. Braki danych były rozłożone równomiernie w każdej ze zmiennych – odpowiednio po 5, 10 i 20 w każdej ze zmiennych.

W kolejnym etapie eksperymentu dla każdego ze zbiorów danych zweryfikowano hipotezę o MCAR za pomocą testu statystycznego przedstawionego w punkcie 3. Następnie uzupełniono brakujące dane, korzystając z poszczególnych metod.

[1] W mechanizmie MAR powstanie braku danych jest zależne od pozostałych zmiennych, a nie jest zależne od samej zmiennej, w której wystąpił określony brak danej.

Ponieważ dane eksperymentalne miały charakter przedziałowy, za miarę błędu szacunku przyjęto błąd bezwzględny według wzoru:

$$b_b = |y_i - \hat{y}_i|, \tag{3}$$

gdzie: b_b – błąd bezwzględny, y_i – wartość prawdziwa, $\hat{y}_i$ – wartość oszacowana.

Następnie obliczono średnią wartość tego błędu dla każdego z testowanych zbiorów.

5. Wyniki badań i wnioski końcowe

Poniżej zamieszczono wyniki eksperymentu symulacyjnego. Tabela 1 przedstawia wartości *p-value* opisanego w punkcie 3 testu MCAR dla każdego ze zbiorów danych. Należało się spodziewać, że w każdym z przypadków, z wyjątkiem MCAR, test powinien odrzucić hipotezę o MCAR. Tak się jednak nie stało. Jedynie w danych typu NMAR na zbiorze dwumodalnym hipotezę tę odrzucono na poziomie istotności 0,05 (co zaznaczono drukiem wytłuszczonym). Powodem tego może być, z jednej strony, słaba moc testu (test ten jest jedynie asymptotycznie zbieżny do statystyki χ^2), a z drugiej – sposób realizacji mechanizmu generowania braków danych typu MAR na zbiorach multimodalnych i unimodalnych oraz NMAR na zbiorach unimodalnych w eksperymencie symulacyjnym. Niewątpliwie zagadnienie to wymaga dalszych badań.

Tabela 1. Wyniki testu MCAR dla poszczególnych zbiorów testowych

Rozkład	% braków	Mechanizm	*Chi*-kwadrat	DF	*p-value*
nmd24	3	mar	13,555	18	0,758
nmd24	10	mar	25,862	27	0,526
nmd24	20	mar	31,509	28	0,295
nu_4	3	mar	2,108	17	1,000
nu_4	10	mar	4,871	22	1,000
nu_4	20	mar	7,041	28	1,000
nmd24	3	mcar	10,289	14	0,741
nmd24	10	mcar	18,092	20	0,581
nmd24	20	mcar	33,611	27	0,178
nu_4	3	mcar	10,289	14	0,741
nu_4	10	mcar	18,092	20	0,581
nu_4	20	mcar	33,611	27	0,178
nmd24	**3**	**nmar**	**32,554**	**14**	**0,003**
nmd24	**10**	**nmar**	**80,199**	**25**	**0,000**
nmd24	**20**	**nmar**	**130,198**	**28**	**0,000**
nu_4	3	nmar	7,182	12	0,845
nu_4	10	nmar	24,633	23	0,369
nu_4	20	nmar	30,415	27	0,296

Źródło: opracowanie własne.

Tabela 2 przedstawia średnie wartości bezwzględnych błędów szacunku. Wyniki najlepsze w sensie średniego bezwzględnego błędu szacunku zaznaczono drukiem wytłuszczonym.

Tabela 2. Średnie wartości bezwzględnych błędów szacunku

Rozkład	% braków	Mechanizm	MI	SI-EM	SOM
nmd24	3	mar	1,92712	**1,46963**	1,53310
nmd24	10	mar	1,56577	**1,17915**	1,25635
nmd24	20	mar	1,41825	**1,15443**	1,17009
nu_4	3	mar	1,13964	**0,79885**	0,99311
nu_4	10	mar	1,19883	**0,81294**	0,98963
nu_4	20	mar	1,13235	**0,81392**	1,03393
nmd24	3	mcar	1,25791	**0,84826**	0,96087
nmd24	10	mcar	1,36310	0,90586	**0,90333**
nmd24	20	mcar	1,48713	**1,03335**	1,11998
nu_4	3	mcar	1,15768	**0,94277**	1,17345
nu_4	10	mcar	1,10612	**0,82661**	1,09066
nu_4	20	mcar	1,08843	**0,80299**	0,92347
nmd24	3	nmar	1,21225	0,87572	**0,79707**
nmd24	10	nmar	1,32881	**0,92063**	0,97309
nmd24	20	nmar	1,50083	1,05640	**0,99307**
nu_4	3	nmar	1,93008	**1,88016**	1,89575
nu_4	10	nmar	1,55128	**1,45817**	1,51228
nu_4	20	nmar	1,22405	**1,04227**	1,17691

Źródło: opracowanie własne.

Rezultaty badań skłaniają do wyciągnięcia następujących wniosków:

1. Metoda SI-EM okazała się najdokładniejszą w sensie średniego bezwzględnego błędu szacunku w większości przypadków.
2. W danych pochodzących z rozkładu normalnego i zastosowania mechanizmu MCAR oraz MAR różnica w dokładności była znaczna na korzyść metody SI-EM.
3. Jedynie w mechanizmie NMAR na zbiorze dwumodalnym w dwóch przypadkach metoda SI-EM okazała się gorsza od SOM.
4. W mechanizmie NMAR na zbiorze i stosunkowo niskim procencie braków danych wszystkie metody charakteryzowały się podobną dokładnością.

Przeprowadzony eksperyment symulacyjny potwierdza nieparametryczność metody SOM, tzn. jej zdolność do dopasowania się do rozkładu. Ma to znaczenie zwłaszcza wówczas, gdy rozkład w populacji, z której pochodzi próba, znacznie odbiega od rozkładu normalnego (np. ma strukturę grupową), a mechanizm powodujący powstawanie braków nie jest losowy (NMAR). Jeśli próba pochodzi z populacji o rozkładzie normalnym, należy stosować metody, których własności formalne są dobrze opisane (SI-EM, MI).

Literatura

Grabiński T., Wydymus S., Zeliaś A., *Z badań nad metodami predykacji brakujących informacji*, Zeszyty Naukowe AE nr 114 w Krakowie, Kraków 1979, s. 31-60.

Grabowski M., *Application of Self-Organizing Maps to Outlier Identification and Estimation of Missing Data*, [w:] *Advances in Data Science and Classification*, red. A. Rizzi, M. Vichi, H.H. Bock, Springer, Berlin 1998, s. 279-286.

Grabowski M., *Sieć Kohonena jako metoda szacowania brakujących danych*, Zeszyty Naukowe AE nr 522, Kraków 1999, s. 37-52.

Grabowski M., *Handling Missing Values in Marketing Research Using SOM*, [w:] *Innovations in Classification, Data Science, and Information Systems*, red. D. Baier, K.-D. Wernecke, Proc. 27th Annual GfKl Conference, University of Cottbus, 12-14 marca 2003, Springer-Verlag, Heidelberg-Berlin 2004, s. 322-329.

Hesterberg T., *A Graphical Representation of Little's Test for MCAR*, MathSoft, Research Report nr 94, 1999, http://www.statsci.com/Hesterberg/articles /tech94-mi-little.ps.

Kohonen T., *Self-Organizing Maps*, Springer-Verlag, Berlin 1995.

Kohonen T., Hynninen J., Kangas J., Laaksonen J., *SOM_PAK The Self-Organizing Map Program Package Version 3.1*, University of Technology, Helsinki 1995, http://www.cis.hut.fi/research/som_pak/.

Kordos J., *Jakość danych statystycznych*, PWE, Warszawa 1998.

Little R.J.A., *A Test of Missing Completely at Random for Multivariate Data with Missing Values*, „Journal of the American Statistical Association" 1988, 83(4), s. 1198-1202.

Little R.J.A., Rubin D.A., *Statistical Analysis with Missing Data*, John Wiley & Sons, New York 2002.

Pawełek B., *Metody szacowania brakujących informacji w szeregach przekrojowo-czasowych*, niepublikowana praca doktorska, AE, Kraków 1996.

Schafer J.L., *Analysis of Incomplete Multivariate Data*, Chapman & Hall, London 1997.

EFFICIENCY OF SELECTED METHODS FOR IMPUTATION OF INCOMPLETE DATA AND DIFFERENT MISSING DATA MECHANISMS

Summary

Incomplete data, issue addressed by various authors, constitute a serious problem for many classical data analysis methods. The concept of utilizing *SOM* neural network (Kohonen 1995) as an effective technique for handling data vectors containing missing components was presented in the previous papers of the author. The following article compares SOM with other advanced missing data handling methods across different missing data mechanisms (Little, Rubin 1987): MCAR, MAR and NMAR. Artificially created multidimensional random data sets with various predefined distributions will be the basis for this comparison.