Imputacja brakujących danych binarnych w modelu autologistycznym

Marta Zalewska

Zakład Profilaktyki Zagrożeń Środowiskowych i Alergologii
Wydział Nauki o Zdrowiu, Warszawski Uniwersytet Medyczny
Żwirki i Wigury 61, 02-091 Warszawa
e-mail: zalewska.marta@gmail.com

Wojciech Niemiro

Wydział Matematyki i Informatyki
Uniwersytet Mikołaja Kopernika, Chopina 12/18, 87-100 Toruń
oraz Instytut Matematyki Stosowanej i Mechaniki, Uniwersytet Warszawski
Banacha 2, 02-097 Warszawa
e-mail: wniemiro@gmail.com

Bolesław Samoliński

Zakład Profilaktyki Zagrożeń Środowiskowych i Alergologii
Wydział Nauki o Zdrowiu, Warszawski Uniwersytet Medyczny
Żwirki i Wigury 61, 02-091 Warszawa
e-mail: bsamol@amwaw.edu.pl

Brakujące dane są poważnym problemem w badaniach statystycznych, szczególne w medycynie. Wiele metod analizy statystycznej wymaga aby zbiór danych był kompletną, prostokątną macierzą bez pustych miejsc. Imputacja jest techniką wypełniania brakujących danych. Zaprezentujemy algorytm, który reprezentuje *modelowe* podejście imputacji i wykorzystuje metody Monte Carlo (MCMC). Zakładamy bayesowski model statystyczny i brakujące dane losujemy z rozkładu predykcyjnego wyznaczonego przez dane obserwowane, przynajmniej w przybliżeniu.

Koncentrujemy się na przypadku zmiennych binarnych i na modelu autologistycznym, w którym wektor $x \in \{0,1\}^d$ ma rozkład prawdopodobieństwa $p(x|\beta) \propto \exp\left(\sum_{i,j=1}^{d} \beta_{ij} x_i x_j\right)$.

Nasz algorytm jest pewną wersją próbnika Gibbsa. Wykorzystujemy pewne aproksymacje, uzasadnione heurystycznie. Podajemy wyniki badań symulacyjnych, które potwierdzają użyteczność algorytmu. Wykorzystujemy rzeczywiste dane medyczne pochodzące z badania ECAP (epidemiologia alergii w Polsce). W tych danych generujemy sztucznie „braki" a następnie próbujemy je wypełniać używając różnych algorytmów i sprawdzamy zgodność z prawdziwymi, „zasłoniętymi" danymi.