Wnioskowanie statystyczne na podstawie danych przedziałowych - interpretacje, problemy, przykłady

Przemysław Grzegorzewski

Wydział Matematyki i Nauk Informacyjnych Politechniki Warszawskiej
Instytut Badań Systemowych Polskiej Akademii Nauk

Streszczenie

Rozwój metod wnioskowania statystycznego oraz postęp w analizie danych i uczeniu maszynowym powiązany jest z przekraczaniem kolejnych ograniczeń dotyczących struktury danych. Zainteresowania współczesnej statystyki nie ograniczają się już - jak w klasycznej statystyce - do danych ujętych w postaci liczb i wektorów, ale obejmują również analizę danych funkcjonalnych (Cuevas, 2014), symbolicznych (Billard i Diday, 2003), rozmytych itd.

W ostatnim czasie dużym zainteresowaniem cieszą się metody wnioskowania na podstawie danych przedziałowych (por., np. Blanco-Fernández i in., 2013; Grzegorzewski i Ramos-Guajardo, 2015). Za pomocą przedziałów można w sposób stosunkowo prosty, a zarazem efektywny, modelować brak precyzji, niepewność wynikającą z braku informacji, fluktuacje mierzonej wielkości itp. Zdarzają się także sytuacje, w których dokładne dane liczbowe zastępowane są przedziałami w celu zabezpieczenia prywatności danych.

Konstrukcja metod wnioskowania statystycznego na podstawie danych przedziałowych nie powinna sprowadzać się do mechanicznego uogólniania metod klasycznych, uzupełnionych o arytmetykę przedziałową. Budowę konkretnej procedury statystycznej należy zacząć od ustalenia sposobu interpretownia danych, bowiem samo przyjęcie przedziałowej struktury danych nie pociąga za sobą domyślnej interpretacji (w przypadku danych przedziałowych można mówić co najmniej o dwóch podstawowych sposobach interpretacji - epistemicznym i ontycznym). Tymczasem konsekwencją wyboru interpretacji jest możność zastosowania takiego czy innego aparatu analitycznego. Wybór interpretacji przekłada się również na własności kostruowanych procedur statystycznych, jakość i sensowność dokonywanych za ich pomocą prognoz itd. (por. np. Couso i Dubois, 2014; Hüllermeier, 2014).

W referacie zostaną zasygnalizowane pewne problemy pojawiające się w trakcie budowy narzędzi statystycznych do wnioskowania na podstawie danych przedziałowych. Zostaną omówione możliwe scenariusze działania, wynikające z przyjętego sposobu interpretacji danych. Rozważania zostaną zilustrowane przykładami praktycznych rozwiązań.

Literatura

Billard, L., Diday, E. (2003). From the statistics of data to the statistics of knowledge: symbolic data analysis. *Journal of the American Statistical Association 98*, 470-487.

Blanco-Fernández, A., Colubi, A., González-Rodríguez, G. (2013). Linera regression analysis for interval-valued data based on set-arithmetic: a review. W: *Towards Advanced Data Analysis*. Borgelt et. all (Eds.). Springer. pp. 19-31.

Couso, I., Dubois, D. (2014). Statistical reasoning with set-valued information: Ontic vs. epistemic views. *International Journal of Approximate Reasoning 55*, 1502–1518.

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference 147*, 1-23.

Grzegorzewski, P., Ramos-Guajardo, A.B. (2015). Similarity based one-sided tests for the expected value and interval data. W: *Proceedings of the 2015 Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology*. Alonso J.M., Bustince H., Reformat M. (Eds.). Atlantis Press. pp. 960-966.

Hüllermeier E. (2014). Learning from imprecise and fuzzy observations: data disambiguation through generalized loss minimization. *International Journal of Approximate Reasoning 55*, 1519–1534.