A Multivariate Missing Data Inputation Method Based on Clustering. Application to World Health Organization data.

K. Gibert [1,2], L. Salvador-Carulla, J. Morris, S. Saxena, S.

(1) Knowledge Engineering and Machine Learning Group (KEMLG), Universitat Politècnica de Catalunya, Barcelona, 08034 (Spain)

(2) Department of Statistics and Operations Research Universitat Politècnica de Catalunya, Barcelona, 08034 (Spain).

(3) Professor of Psychiatry. Department of Neurosciences. University of Cadiz. Plaza Falla 9 11003 Cadiz (Spain)

(4) Working Group on Clinical Management of the Spanish Society of Psychiatry (Grupo de Trabajo de Gestión Clínica de la Sociedad Española de Psiquiatría) (GClin-SEP)

(5) Department of Mental Health and Substance Abuse, World Health Organization, 20 Ave. Appia, Geneva 1211, Switzerland

In real applications, it is quite usual to find important rates of missing data that has to be preprocessed before the analysis. The literature for missing imputation is abundant. However, the most precise imputation methods require quite a long time, and sometimes specific software, and this implies a significant delay to get final results. In this work, we propose a missing imputation methodology based on clustering, that provides a good trade-off between precision and required time to prepare data for the analysis. The proposal is applied in the context of better understanding the Mental Health Systems in Low and Middle Income Countries, project leaded by the Mental Health Department of the World Health Organization.

The Multivariate Missing Inputation based on Clustering (MuMIC) method presented in this work is a non parametric method that uses the conditional mean for imputation according to the underlying structure of the dataset itself. A first subset of quasi-full relevant variables is selected together with the experts. The small set of missing data from this initial set of variables is inputed based on the background knowledge of the experts. The inputed data matrix is clustered and the class identifier is used to find conditional means for all remaining variables in the dataset. A method combining the prior expert knowledge with multivariate analysis to find input values that take into account the joint distribution of all variables and can be completed in a relatively short time, without requiring assumptions on the probabilistic models of the variables (normality, exponentiality, etc). Real applications shown a good performance in both quality of results and required time.

Ackowledgement

Juan Carlos Martín Sánchez, for helping with data processing.