© 2013 FIZ Karlsruhe

io-port 06105187

Joenssen, Dieter William; Bankhofer, Udo

Hot deck methods for imputing missing data. The effects of limiting donor usage.

Perner, Petra (ed.), Machine learning and data mining in pattern recognition. 8th international conference, MLDM 2012, Berlin, Germany, July 13–20, 2012. Proceedings. Berlin: Springer (ISBN 978-3-642-31536-7/pbk). Lecture Notes in Computer Science 7376. Lecture Notes in Artificial Intelligence, 63-75 (2012).

Summary: Missing data methods, within the data mining context, are limited in computational complexity due to large data amounts. Amongst the computationally simple yet effective imputation methods are the hot deck procedures. Hot deck methods impute missing values within a data matrix by using available values from the same matrix. The object, from which these available values are taken for imputation within another, is called the donor. The replication of values leads to the problem, that a single donor might be selected to accommodate multiple recipients. The inherent risk posed by this is that too many, or even all, missing values may be imputed with the values from a single donor. To mitigate this risk, some hot deck variants limit the amount of times any one donor may be selected for donating its values. This inevitably leads to the question under which conditions such a limitation is sensible. This study aims to answer this question though an extensive simulation. The results show rather clear differences between imputations by hot deck methods in which the donor limit was varied. In addition to these differences, influencing factors are identified that determine whether or not a donor limit is sensible.

Keywords: Hot Deck Imputation; Missing Data; Nonresponse; Imputation; Simulation
doi:10.1007/978-3-642-31537-4_6