About the selection of optimal subdata for big data regression

Dimitris Karlis

Athens University of Economics and Business, Greece

Abstract

The introduction of faster computer processing in conjunction with enhanced storage capabilities created an exponential growth in the size of datasets. However, in big data era, researchers face some serious problems, since statistical analyses and modeling processes of huge volumes of data may be infeasible due to limitations in the computational resources. Even linear regression can be either really difficult or problematic in case of a huge dataset, since for example the data cannot fit to the memory or manipulation with huge matrices are needed. A basic approach is based on selecting representative subdata to run the regression. Existing approaches select the subdata using information criteria or some basic properties
of the orthogonal arrays. We present a new approach that is based on the D-optimality approach, aiming at improving approaches that already exist in the current literature. The idea is that based on minimum additional time cost we can improve substantially the selected subdata. The proposed approach is evaluated through simulation experiments, in order to clarify the trade-offs between execution time and information gain. Real data applications are also provided.