STRATIFIED SAMPLING ON DATA NUGGETS: A STRATEGY FOR DATA REDUCTION

Ravi Kumar Gangadharan, Vanessa Petrarca, Maria Chiara Pagliarella, Giovanni Porzio

Department of Economics and Law, University of Cassino and Southern Lazio,
(e-mail: mc.pagliarella@unicas.it, porzio@unicas.it)

ABSTRACT: The increased volume and velocity of data production has been causing a growing cost in storing and analysing data. Thus, due to this continuously increasing phenomenon, the urgency of data reduction technique arises.

Data reduction aims at decreasing storage and computational costs for data analysis. In order to tackle with this very large and complex issue, many techniques have been developed and employed (such as clustering, principal points, support points, prototypes, etc.). Among the many, this work focuses on a recently introduced specific type of data reduction method which has been called Data Nuggets.

Data Nuggets reduces huge datasets and compresses the observations into few points, by saving essential information on the data structure. In parallel with standard classic procedures, Data Nuggets splits a dataset in several subsets (called Nuggets) which are defined by three main components: a Center, a Weight (representing the number of observations within each subset), and a Scale (representing the average Nugget within variance).

Particularly, our work aims at investigating to what extent Data Nuggets can be used as a tool to obtain stratified samples from large datasets so that some computational cost can be gained. A comparison in terms of efficiency with respect to statistical techniques applied to a simple random sample drawn from the same large dataset will be provided.

KEYWORDS: Large datasets, computational effort, data partition.

References

BEAVERS, T., CABRERA, J., & LUBOMIRSKI, M. 2020. datanugget: Create, Refine, and Cluster Data Nuggets. *R package version 1.0.0, https://CRAN.R-project.org/package=datanugget*.

CHERASIA K.E., CABRERA J., FERNHOLZ L.T., & FERNHOLZ R. 2023. Data Nuggets in Supervised Learning. In M. Yi, K. Nordhausen (eds.), *Robust and Multivariate Statistical Methods*, Springer, https://doi.org/10.1007/978-3-031-22687-8_20.