Proceedings of the 5th International Conference on Statistics: Theory and Applications (ICSTA'23)
Brunel University, London, United Kingdom – August 03-05, 2023
Paper No. 114
DOI: 10.11159/icsta23.114

Big Data and Design of Experiments: A Combination Which Can Enhance the Performances of Machine Learning Models

Rosa Arboretti[1], Livio Corain[2], Alberto Molena[2]
[1]Department of Civil Environmental and Architectural Engineering, University of Padova,
Via Marzolo, 9, Padova, 35131, Italy
rosa.arboretti@unipd.it
[2]Department of Management and Engineering, University of Padova,
Stradella S. Nicola, 3, Vicenza, 36100, Italy
livio.corain@unipd.it; alberto.molena.1@phd.unipd.it

Extended Abstract

Big Data are a huge quantity of data that are automatically accrued and/or obtained by merging several sources of information. Their availability is a great challenge currently. However, the quality of the Big Data information might be poor because they usually arise from observational studies and not through controlled experimentation. Some authors have dealt with the idea of selecting a subsample of the Big Dataset for a better management of the inferential properties of Big Data and achieve different inferential goals. This topic has been already studied by Ma and Sun [1], Drovandi et al. [2], Wang et al. [3], Wang et al. [4] and Campbell and Broderick [5], among others.

In particular, Drovandi et al. [2], propose to improve the analysis of Big Data through retrospective designed sampling in order to answer particular questions of interest.

They suggest that, depending on the aim of the analysis, to adopt an optimal experimental design perspective whereby instead of (or as well as) analysing all of the data, a retrospective sample set is drawn in accordance with a sampling plan or experimental design, based on an identified statistical question and corresponding utility function. The analyses and inferences are then based on this designed sample. Thus, the novelty here is the ability to extract the required design points. The suggested approach can also be considered as a targeted way of undertaking sampling in divide-and-conquer algorithms or for "sequential learning" in which a given design is applied to incoming data or new data sets until the question of interest is answered with sufficient precision, or a pre-determined criterion is reached. It can also be used for evaluating the quality of the data, including potential biases and data gaps, when the required optimal or near-optimal design points cannot be extracted from the data.

Recently Deldossi and Tommasi [6] consider the theory of optimal design to extract subsamples, which allows to incorporate the inferential goal in the sampling strategy. They propose a purposive selection strategy (named Optimal Design Based, ODB method) consisting of two steps: at first, they identify the most informative values of the explanatory variables according to an optimality criterion, then, they select those observations from the full dataset that are closer to these theoretical optimal values. Hence, this optimal-sampling approach allows to select the most "informative" observations from the Big Dataset.

A paper by Arboretti et Al. [7], discusses the jointly application of Design of Experiments (DOE) and Machine Learning (ML) techniques to optimize data collection and analysis: here, the main goal is to understand which combination of experimental design and ML model help improving performances and accuracy in data analytics. Moreover, another paper by Arboretti et Al. [8], introduces an algorithm that can be used for sequential data collection in Physical Experiments when investigating three or more responses, called ALPERC. Such procedure selects the best experimental combination each time, taking into account both uncertainty and variable importance.

Reviewing above contributions and comparing them, allows us to summarize the research done so far in this direction and identify gaps and new directions to be explored also merging different ideas and techniques.

References

[1] P. Ma and X. Sun, "Leveraging for big data regression," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 7, no. 1, pp. 70–76, 2015.

[2] C. C. Drovandi, C. Holmes, J. M. McGree, K. Mengersen, S. Richardson, and E. G. Ryan, "Principles of experimental design for big data analysis," *Stat. Sci. Rev. J. Inst. Math. Stat.*, vol. 32, no. 3, p. 385, 2017.

[3] H. Wang, M. Yang, and J. Stufken, "Information-based optimal subdata selection for big data linear regression," *J. Am. Stat. Assoc.*, vol. 114, no. 525, pp. 393–405, 2019.

[4] H. Wang, R. Zhu, and P. Ma, "Optimal subsampling for large sample logistic regression," *J. Am. Stat. Assoc.*, vol. 113, no. 522, pp. 829–844, 2018.

[5] T. Campbell and T. Broderick, "Automated scalable Bayesian inference via Hilbert coresets," *J. Mach. Learn. Res.*, vol. 20, no. 1, pp. 551–588, 2019.

[6] L. Deldossi and C. Tommasi, "Optimal design subsampling from Big Datasets," *J. Qual. Technol.*, vol. 54, no. 1, pp. 93–101, 2021.

[7] R. Arboretti, R. Ceccato, L. Pegoraro, and L. Salmaso, "Design choice and machine learning model performances," *Qual. Reliab. Eng. Int.*, 2022.

[8] R. Arboretti, R. Ceccato, L. Pegoraro, and L. Salmaso, "Active learning for noisy physical experiments with more than two responses," *Chemom. Intell. Lab. Syst.*, p. 104595, 2022.