Contribution ID: **54** Type: **not specified**

Accounting for misspecification in design based subsampling approaches: a comparative analysis

Monday, 16 September 2024 12:00 (30 minutes)

Supervised learning under measurement constraints presents a challenge in the field of machine learning. In this scenario, while predictor observations are available, obtaining response observations is arduous or cost-prohibitive. Consequently, the optimal approach involves selecting a subset of predictor observations, acquiring the corresponding responses, and subsequently training a supervised learning model on this subset.
Among various subsampling techniques, the design-inspired subsampling methods have attracted great interest in recent years (see Yu et al. (2023) for a review). Most of these approaches have shown remarkable performance in coefficient estimation and model prediction, but their performance heavily relies on a specified model. When the model is misspecified, misleading results may be obtained.
In this work we provide a comparative analysis of methods that account for model misspecification, as for instance the LowCon approach introduced by Meng et al. (2021) for selecting a subsample using an orthogonal Latin hypercube design, or other subsampling criteria based on space-filling designs. Furthermore, the robustness of these methods to outliers is also assessed (see Deldossi et al. (2023)). Empirical comparisons are conducted, providing insights into the relative performance of these approaches.

References

Deldossi, L., Pesce, E., Tommasi, C. (2023) Accounting for outliers in optimal subsampling methods. Statistical Papers, 64(4), 1119-1135.

Meng, C., Xie, R., Mandal, A., Zhang, X., Zhong, W., Ma, P. (2021) Lowcon: a design-based subsampling approach in a misspecified linear model. Journal of Computational and Graphical Statistics, 30:694–708

Yu, J., Ai, M., Ye, Z. (2024) A review on design inspired subsampling for big data. Statistical Papers, 65, 467–510

Type of presentation

Talk

Classification

Mainly methodology

Keywords

Massive data - Orthogonal design - Space-filling design

Primary author: DELDOSSI, Laura (Università Cattolica del Sacro Cuore)

Co-author: Dr TOMMASI, Chiara (University of Milan)

Presenter: DELDOSSI, Laura (Università Cattolica del Sacro Cuore)

Session Classification: SIS Invited session: Stratification, subsampling, randomization,issues and proposals

Track Classification: Other/ Special/ Invited