Distributed Statistical Learning Framework for Small and Big Data

Subhadeep Mukhopadhyay
Temple University, Department of Statistics
Philadelphia, PA 19122, USA
Email: deep@temple.edu
Web: http://sites.temple.edu/deepstat/

Extended Abstract. We seek to develop a systematic and automatic statistical learning framework that will provide a generic platform to extend traditional and modern statistical modeling tools to large datasets using scalable distributed algorithms.

This talk proposes a generic statistical approach, called `MetaLP`, that addresses two main challenges of large datasets: (1) massive volume, and (2) variety or mixed data problem. We apply this general theory in the context of variable selection by developing a nonparametric distributed statistical inference framework that allows us to extend traditional and novel statistical methods to massive data that cannot be processed and analyzed all at once using standard statistical software. Our proposed algorithm leverages the power of distributed and parallel computing architecture, which makes it more scalable for large-scale data analysis. Furthermore, we show that how this broad statistical learning scheme can be successfully adapted for 'small' data like resolving the challenging problem of Simpson's paradox. We believe this is a great stepping stone towards developing 'United Statistical Algorithms' (Mukhopadhyay and Parzen, 2014) to bridge the increasing gap between the theory and practice of small and big data analysis.

This is a joint work with Professor Emanuel Parzen. Current research has been awarded Best Paper Award at the Fox Young Scholars Interdisciplinary Big Data Grant and by the Fox School PhD student research competition award. The paper is currently under review in the Journal of the American Statistical Association (JASA).

Keywords: Distributed algorithm; Small and big data modeling; LP transformation; Meta-analysis; Confidence distribution.

References

[1] Hedges, Larry V., and Ingram, Olkin. (1985). Statistical method for meta-analysis, London: Academic Press.

[2] Dean, Jeffrey, and Ghemawat, Sanjay (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM* **51**, 107–113.

[3] Mukhopadhyay, S. and Parzen, E. (2014). LP approach to statistical modeling. *arXiv:1405.2601*.