Method of divide-and-combine in regularized regression for Big Data

Lu Tang[1], **Ling Zhou**[1] **and Peter X.K. Song**[1]

[1] *University of Michigan, USA*

Abstract

When a data set is too big to be analyzed entirely once by one computer, the strategy of divide-and-combine (MODAC) has been the method of choice to overcome the computational hurdle. Although random data partition has been widely adopted, there is lack of clear theoretical justification and practical guidelines to combine results obtained from separately analyzed sub-datasets, especially when a regularization method such as LASSO [2] is utilized for variable selection in the generalized linear model regression. We develop a new strategy to combine separately regularized estimates of regression parameters by means of the confidence distributions [3] of biased corrected estimators. We first establish the theory for the construction of the confidence distribution and then show that the resulting MODAC estimator enjoys the Fishers efficiency, the efficiency of the maximum likelihood estimator obtained from the analysis of entire data once. Furthermore, using the MODAC estimator we propose a variable selection procedure, which is compared analytically and numerically via extensive simulations with the existing majority-voting method [1] and the gold standard of one-time entire data analysis.

Keywords

Confidence distribution, Generalized linear model, LASSO, Meta analysis.

References

[1] Chen, X. and M.-G. Xie (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica 24*, 16551684.

[2] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Methodological) 58*, 267288.

[3] Xie, M.-G. and K. Singh (2013). Confidence distribution, the frequentist distribution estimator of a parameter: a review. *International Statistical Review 81*, 339.