Biostatistics and Biometrics
Open Access Journal
ISSN: 2573-2633

Juniper Publishers
key to the Researchers

Mini Review
Volume 6 Issue 2 - April 2018
DOI: 10.19080/BBOAJ.2018.06.555684

Biostat Biometrics Open Acc J

Model Averaging for High-Dimensional Linear Models

Juming Pan*

Department of Mathematics and Statistics, University of Minnesota Duluth, USA

Submission: January 30, 2018; **Published:** April 18, 2018

***Corresponding author:** Juming Pan, Department of Mathematics and Statistics, University of Minnesota Duluth, USA, Tel: 218-726-7315; Email: jpan@d.umn.edu

Abstract

Model averaging has attracted increasing attention in recent years for the analysis of high-dimensional data. By weighting several competing scientific models suitably, model averaging attempts to achieve stable and improved prediction in the case where the number of predictors greatly exceeds the sample size.

Keywords : Model averaging; High-dimensional regression models; Stable prediction.

Introduction

With the advent of high-throughput technologies, high-dimensional data have been frequently generated for the understanding of biological processes such as disease occurrence and cancer study. Motivated by these important applications, there has been a dramatic development in the statistical analysis of high-dimensional data, with the main goal of improving estimation and prediction. See [1] & [2], and examples therein. To obtain the parsimonious and compact representations of the data, a lot of work has been done under the context of mode selection using data oriented penalty functions, and a comprehensive overview can be found in [3]. Instead of relying on only one best model, model averaging focuses on prediction accuracy by assigning different weights to the candidate models. If the weights can be properly determined, then prediction performance could be enhanced. Developing for the data in which the number of predictors is more than the number of observations, model averaging for high dimensional regression is recently studied in [4], in which a two-stage procedure is proposed. First, candidate models are constructed by marginal correlation; Second, the optimal weights are found by delete-one-out cross-validation. One major contribution of [4] is that it relaxes the model weights to vary freely between 0 and 1 without the standard constraint of summing up to 1. Furthermore, it is shown that the proposed method could asymptotically achieve the lowest possible prediction loss, yet it is computationally efficient even when there are thousands of variables.

.However, [4] might cause an increasing variance due to the reuse of the same data for generating candidate models and then estimating weights for these candidate models in the two steps. Another approach introduced in [5] avoids reusing the same data by randomly splitting dataset multiple times and generating a second level dataset. Using two different datasets for these two steps could sufficiently avoid over fitting the dataset, and as a result, provides improved prediction accuracy. In addition to the proposed procedure, a weighted importance index can be obtained to rank the predictors from relevant predictors to irrelevant ones, which is often useful in gene expression study.

Discussion

In order to ensure excellent prediction accuracy, model averaging techniques are introduced under the setting of high-dimensional linear models. However, future work still needs to be done. For example, reducing the cost of computation is a concern for model averaging especially with the complexity of high-dimension regression. Moreover, missing values are quite common in high-dimensional data, which leaves space for future research in model averaging. Finally, it seems interesting to explore the possibility of further relaxing the weights to allow for negative values. These and many other unsettled issues deserves further investigation.

References

1. Bühlmann P, van de Geer S (2011) Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer.

2. Li X, Xu R (2009) High-Dimensional Data Analysis in Cancer Research. Springer.

3. Fan J, Lv J (2010) A selective overview of variable selection in high dimensional feature space. Statistica Sinica 20(1): 101-148.

4. Ando T, Li KC (2014) A model-averaging approach for high-dimensional regression. J Am Stat Assoc 109: 254-265.

5. Lin B, Wang Q, Zhang J, Pang Z (2017) Stable prediction in high-dimensional linear models. Statistics and Computing 27(5): 1401-1412.

DOI: 10.19080/BBOAJ.2018.06.555684

How to cite this article: Juming P. Model Averaging for High-Dimensional Linear Models. Biostat Biometrics Open Acc J. 2018; 4(3): 555639. DOI: 10.19080/BBOAJ.2018.06.555684.