

Predicting Survival from High Dimensional Data

Susmita Datta

*Department of Bioinformatics and Biostatistics,
University of Louisville, Louisville, KY 40202, USA
E-mail: susmita.datta@louisville.edu*

Key words: regression models, microarray, survival, high-dimensional data, partial least squares, LASSO

We consider the problem of predicting survival times of patients from their genomic profiles via linear regression modeling of log-transformed failure times. The ‘partial least squares’ (PLS) and ‘least absolute shrinkage and selection operator’ (LASSO) methodologies are used for this purpose where we first modify the data to account for censoring. A major objective of this work is to investigate the performances of PLS and LASSO in the context of genomic data where the number of covariates is very large and there are extremely few samples.

We demonstrate that LASSO outperforms PLS in terms of prediction error when the list of covariates includes a moderate to large percentage of useless or noise variables; otherwise, PLS may outperform LASSO. For a moderate sample size (one hundred with ten thousand covariates), LASSO performed better than a no covariate model (or noise based prediction).

We also consider a regularized version of PLS that seems to have better performance than the regular PLS.

Three approaches of handling right censored data - reweighting, mean imputation and multiple imputation are considered. The mean imputation method appears to best track the performance of the full data PLS or LASSO.