

Sparseness determination for boosting estimation of high-dimensional survival models

Harald Binder^{1,2} and Martin Schumacher²

¹ Freiburg Center for Data Analysis and Modeling, University of Freiburg, Germany

² Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Germany

There are several techniques for fitting sparse survival models to high-dimensional data, arising e.g. from microarrays. However, sparseness of the estimates, obtained e.g. with Lasso-like techniques, often comes at the cost of underestimating effects for the few covariates which have a large influence. This is especially problematic when clinical covariates are included in addition to a large number of microarray features. We address this issue in a new boosting approach, allowing for mandatory covariates, whose coefficients are estimated unpenalized [1]. However, when it is not known which covariates may potentially have a large effect, a different strategy is needed. The boosting procedure is therefore extended to automatically adapt the penalties used for estimation after every boosting step. Decreasing the penalty for a selected covariate will favor sparse models and avoid underestimation of effects. A scheme which increases the penalty will favor models with groups of important covariates. As it is difficult for the user to decide on a specific penalty updating scheme, we investigate whether the integrated Brier score, obtained via bootstrap .632+ prediction error curve estimates [3], could be used for automatic sparseness determination. The performance of the resulting procedure is compared to an alternative boosting strategy that employs two rounds of boosting, a first one where the important covariates are identified and a second one where these are then preferentially selected [2]. The advantages of our new boosting procedure will be illustrated using microarray survival data from patients with diffuse large B-cell lymphoma.

References

- [1] Binder H, Schumacher M (2008) Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics* 9: 14
- [2] Bühlmann P (2007) Twin Boosting: Improved feature selection and prediction. Technical Report, Seminar für Statistik, ETH Zürich
- [3] Gerds TA, Schumacher (2007) Efron-type measures of prediction error for survival analysis. *Biometrics* 63: 1283-1287