SURVIVAL MODELS BUILT FROM GENE EXPRESSION DATA USING GENE GROUPS AS COVARIATES

Kai Kammers, Jörg Rahnenführer

Fakultät Statistik, Technische Universität Dortmund,
44221 Dortmund, Germany

Email: kammers@statistik.uni-dortmund.de

Abstract. We present prediction models for survival times built from high dimensional gene expression data. The challenge is to construct models that are complex enough to have high prediction accuracy but that at the same time are simple enough to allow biological interpretation.

Typical univariate approaches use single genes as covariates in survival time models, multivariate models perform dimension reduction through gene selection. Analysis of time-dependent ROC curves and the area under the curves (AUC) can be used to assess the predictive performance (Gui and Li, 2005).

We present models with higher interpretability by combining genes to gene groups (biological processes or molecular functions) and then using these groups as covariates in the survival models. The hierarchically ordered ”GO groups” (Gene Ontology) are particularly suitable. Cox models are used for detecting covariates that are significantly correlated with survival times. Based on these models statistical shrinkage procedures like Lasso-Regression are applied for variable selection. We make use of the R package penalized (Goeman, 2008) that provides algorithms for penalized estimation in generalized linear models, including linear regression, logistic regression and the Cox proportional hazards model.

Our aim is the combination of methods for survival prediction with biological a priori knowledge. First, we compare the prediction performance of models using single genes as covariates with models using gene groups as covariates on several real gene expression datasets. First results indicate that models built with gene groups alone have decreased prediction accuracy since many genes are not yet annotated to their corresponding functions. However, adding gene groups as covariates to models built from single genes improves interpretability while prediction performance remains stable.

In a next step, we integrate GO graph structure in the models (Alexa, Rahnenführer and Lengauer, 2006) in order to cope with the high correlations between neighboring GO groups.

References

Gui, J. and Li, H. (2005): Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics 21(13), 3001–3008.

Goeman, J.J. (2008): An efficient algorithm for L1 penalized estimation, submitted.

GO Consortium (2004): The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research 32:D258-D261. Oxford University Press.

Alexa, A., Rahnenführer, J. and Lengauer, T. (2006): Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 22(13), 1600–1607.

Keywords

Microarray, Survival Analysis, Gene Ontology