blood

901. HEALTH SERVICES AND OUTCOMES RESEARCH: POSTER I | NOVEMBER 15, 2013

Survival Prediction In High Dimensional Datasets – Comparative Evaluation Of Lasso Regularization and Random Survival Forests

Erel Joffe, MD MSc,[*,1] Kevin R. Coombes, PhD;[*,2] Yi Hua Qiu, MD;[*,3] Suk Young Yoo, PhD;[*,2] Nianxiang Zhang, PhD,[*,2] Elmer V Bernstam, MD MS;[*,4] Steven M. Kornblau, MD[3]

[1]Department of Hematology, Tel Aviv Medical Center, Tel Aviv, Israel,

[2]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA,

[3]Department of Leukemia, The University of Texas MD Anderson Cancer Center, Houston, TX, USA,

[4]School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, USA

Blood **(2013) 122 (21) : 1728.**

http://doi.org/10.1182/blood.V122.21.1728.1728

Abstract

Background

High-dimensional data obtained using modern molecular technologies (e.g., gene expression, proteomics) and large clinical datasets is increasingly common. Risk stratification based on such high-dimensional data remains challenging. Traditional statistical models have a limited capability of handling large numbers of variables, non-linear effects, correlations and missing data. More importantly, as more variables are analyzed, models tend to over-fit (i.e., the model provides good predictions on the studied data but performs poorly on other data). Recently two methods have been proposed for handling multivariate analysis of high-dimensional data. The Least Absolute Shrinkage and Selection Operator (LASSO) minimizes the number of Cox regression coefficients, favoring models that are parsimonious and less likely to over-fit. LASSO is computationally efficient and capable of handling correlated variables. Random Survival Forests (RSF) combines multiple decision trees built on randomly selected subsets of variables. This enables the evaluation of all variables, even in the presence of significant correlations or missing data, and reduces over-fitting. Few studies have evaluated these models on realistic datasets. This study compared the performance of LASSO and RSF to that of the traditional Cox Proportional Hazard (CoxPH) with respect to their ability to predict survival based on high dimensional reverse phase protein array (RPPA) data from Acute Myeloid Leukemia (AML) patients.

Methods

Our data were derived from previous work to define the role of the pathway activation in AML using a custom made RPPA onto which were printed leukemia enriched cells from 511 newly diagnosed AML patients collected between 1992 and 2007. The RPPA was probed with 231 strictly validated antibodies. Data were normalized and analysis was performed for the 415 subjects that underwent therapy at MD Anderson. The dataset also included 38 clinical variables. We removed cases with a documented survival of less than 4 weeks and imputed a small number of sporadically missing values by random sampling with replacement. We generated LASSO, RSF, and CoxPH models using R. We built RSF models using 1000 trees and 10 random split points. We then identified key features using the RSF max-subtree and the glmnet cross validation methods. We built an RSF and Cox models to these variables only. Models were trained on a bootstrap sample of the size of the dataset, randomly sampled with replacement, and tested on the un-sampled remaining cases. The process was repeated for 50 iterations and results were averaged. We report Harrell's concordance index (C-Index) and the Brier score for model performance. Harrell's C-Index is a pairwise comparison of all observations in the testset. It evaluates the probability of erroneously assigining longer survival time to one case over the other. Brier score calculates the difference between the predicted and observed probabilities. It evaluates how well the model fits the entire dataset.

Results

Cox regression with LASSO regularization had the best performance based on both Brier score and C-Index (Table 1

Table 1

Comparison of model performance

	C-Index	**Brier Score**
Cox	Doesn't converge	Doesn't converge
RSF	0.66	0.16
Refitted-RSF	0.68	0.14
Cox-RSF	0.64	NA*
LASSO	0.70	0.08
Cox-LASSO	0.68	0.13

Cox-RSF, Refitted-RFS, Cox-LASSO – models limited to features selected by the RSF and LASSO. * Some of the constructed models did not converge pertaining the calculation of averaged Brier score.

and figure 1). For the complete dataset Cox regression models did not converge even when using forward feature selection. Cox regression following feature selection based on RSF had inferior performance compared to other methods.

Conclusions

LASSO and RSF allow for mutlivariate analysis of high-dimensional data that would have not been possible otherwise. LASSO regularization outperforms other methods in terms of accuracy and to select features for traditional Cox models. Using the latter approach has great appeal as traditional Cox models allow easy interpertation of the hazard associated with individual variables. Differences in Brier scores are more pronounced than C-Indexes, possibly indicating a tendency of LASSO regularization to overfit the data compared to RSF.

Disclosures:

No relevant conflicts of interest to declare.

Author notes

* Asterisk with author names denotes non-ASH members.