OncoCast: an improved interface for survival analysis using genomic data

Axel Martin
Memorial Sloan Kettering Cancer Center

Introduction

- steady decrease of whole-genome sequencing has lead this procedure to become a standard of care for cancer patients.
- large amounts of genomic data. Cancer being fundamentally a genetically driven disease this additional data is critical in the improvement of personalized medicine and survival prediction.
- patients are regularly not sequenced at the time of diagnosis, either through referral or simply because they were diagnosed at a time when sequencing was not a common practice.

Model

OncoCast has a variety of penalized regression and gradient boosted models. Namely least absolute shrinkage and selection operator (LASSO), elastic-net (ENET) and Generalized Boosted Regression Models (GBM). The algorithm repeatedly splits the data between training and testing set, the former is used as input with the selected model. At each iteration we use the trained model to generate a predicted risk for patients in the test set, we furthermore record the selected features and they associated coefficients. After performing a large amount of cross-validations we average the predicted risk for each patients, that we further rescale between 0 and 10 for comprehensibility. We then use the functional distribution of the risk score along with clinical relevance in order to generate clinically relevant risk groups.

In order to assess the accuracy of our algorithm we performed simulations. We generated 50 datasets for each setting with a mix of binary and continuous variables. We selected multiple distributions to mimic the different genomic data forms found in cancer research. The underlying true model has 5 strong coefficients, 5 medium strength coefficients, 5 low strength coefficients, 5 very low strength coefficients and 480 noise features.

Finally we present the results found in a metastatic adenocarcinoma cohort at Memorial Sloan-Kettering cancer center[1]. [1]

Simulation Results

Based on the underlying true model described previously we generated 250 random univariate and 250 random binary variables with respective correlation 0, 0.3 and 0.6 within each feature type. For each scenario we generate 50 datasets and we run our algorithm with 100 cross-validation for LASSO, ENET and GBM. At each iteration we calculate the cross-validated survival concordance index, we show below the median concordance and standard deviation. Moreover we know the true risk each patient is facing, thus we stratified them in 5 groups and we assess how often our algorithm correctly assigns each patients to the correct risk group.

		CI			CI Risk		
Corr	SS	ENET	LASSO	GBM	ENET	LASSO	GBM
0							
	100	.81 (.04)	.82 (.04)	.725 (.05)	.56 (.07)	.58 (.08)	.41 (.07)
	200	.87 (.02)	.87 (.02)	.82 (.03)	.70 (.05)	.72 (.05)	.53 (.05)
	500	.89 (.01)	.89 (.01)	.87 (.01)	.84 (.02)	0.85 (.02)	.685 (.03)
0.3							
	100	.84 (.03)	.83 (.04)	.83 (.02)	.62 (.06)	.64 (.07)	.57 (.06)
	200	.88 (.01)	.89 (.01)	.87 (.01)	.78 (.03)	.81 (.03)	.66 (.03)
	500	.9 (.01)	.9 (.01)	.89 (.01)	.88 (.02)	.9 (.01)	.73 (.02)
0.6							
	100	.89 (.01)	.88 (.02)	.88 (.01)	.7 (.05)	.75 (.05)	.69 (.05)
	200	.91 (.01)	.91 (.01)	.9 (.01)	.83 (.03)	.86 (.03)	.74 (.03)
	500	.92 (0)	.92 (0)	.91 (.01)	.9 (.02)	.91 (.02)	.77 (.02)

We observe that the cross validated concordance index is good for all methods even in small sample sizes. Note that the accuracy for the risk groups is lower due to the higher number of groups we gave (5). The concordance increases with correlation because selecting null features still have predictive power. Finally the poorer performance of the GBM algorithm comes from a lack of tuning that is necessary for optimal performance. Similar we registered the selected features and their coefficients for each method under each scenario. We report in the histogram below the the selection frequency for each type of feature importance.

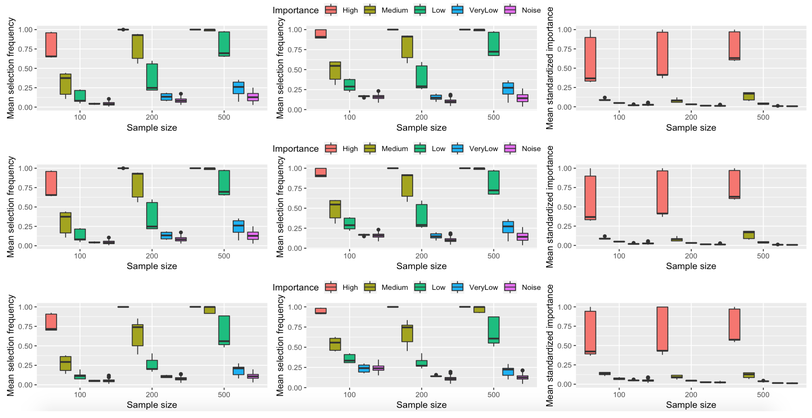

Fig. 1: Feature selection.

Application in Cancer

Using sequencing results from a cohort of 1,054 patients with advanced lung adenocarcinomas, we stratified this patient cohort into four risk groups based on tumor genomic profile.

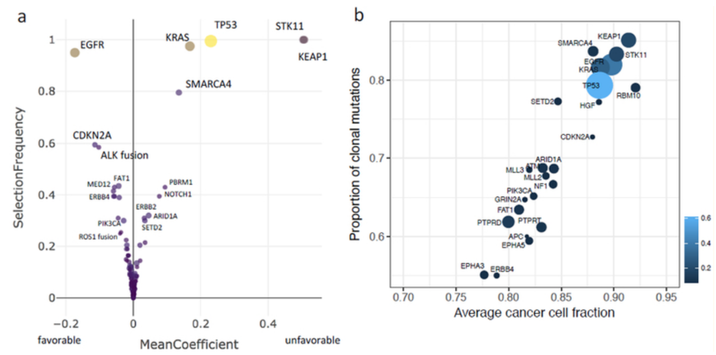

Fig. 2: Feature selection and clonal mutations.

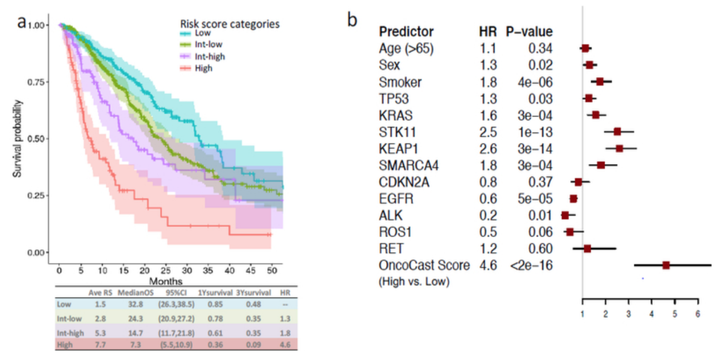

Fig. 3: Startification and OncoCast risk score prognostics.

Patients whose tumors harbored a high-risk profile had a median survival of 7.3 months (95% CI 5.5-10.9), compared to a low risk group with a median survival of 32.8 months (95% CI 26.3-38.5), with a hazard ratio of 4.6 (P<2e-16), far superior to any individual gene predictor or standard clinical characteristics.

Acknowledgements

I would like to thank Ronglai Shen, Gregory Riely and Katherine Pangenea for their help and support.

References

[1] A. Martin R. Shen and G. Riely. “Harnessing Clinical Sequencing Data for Survival Stratification of Patients with Metastatic Lung Adenocarcinomas”. In: *JCO Precision Oncology* (Apr. 2019).