Time dependent ROC curves for the estimation of true prognostic capacity of microarray data

R. Danger and Y. Foucher *

Department of Biostatistics - EA 4275, INSERM U1064, ITUN, University of Nantes, 1 rue G. Veil, 44035 Nantes, France

* Yohann.Foucher@univ-nantes.fr

Introduction

The use of microarray technology has revolutionized the identification of molecular **signatures for the prediction of patient outcome**. The samples (blood, urine, etc.) are collected under the same experimental conditions (event-free) and the individuals are followed up in order to identify features predictive of the event. The time of the event may be observed or not (right-censored).

The **Cox model with lasso penalty** appeared to be a reference a method in this context. Schumacher (2007) demonstrated that the 0.632+ estimator of the prediction error is an interesting indicator for evaluating and comparing different models. It takes into account **overfitting** without splitting the available data into training and validation sets. Nevertheless, 3 limitations should be considered.

- ✓ The tuning parameter is defined using the full data (5-fold cross-validation), while the selection of the model complexity has to be included in each bootstrap iteration in order to avoid overoptimistic results.
- ✓ The prediction error can be used for model comparisons, but it is not a meaningful indicator for biologists or clinicians.
- ✓ The prediction error is based on the regression residuals and therefore depends on the incidence of the event, but the sample may not necessarily represent the targeted population.

Objective

We proposed a 0.632+ estimator of the area under the time-dependent Receiver Operating Characteristic (ROC) curve (Heagerty, 2000). The method is designed for the analysis of censored and/or truncated survival data.

The Cox model with lasso penalty

Let X be the vector of the P feature expressions and β the associated regression coefficients. Let T_j be the time of the event occurrence for the individual j $(j = 1, \dots, N)$. The hazard function is:

$$h(t_j|x_j) = h_0\,(t_j)\exp(\underbrace{\beta x_j}_{\eta_j})$$

The lasso (Tibshirani, 1996) shrinks the regression coefficients towards zero by penalizing the partial likelihood by the sum of their absolute value multiplied the tuning parameter λ:

$$\hat{\beta} = \arg max \left\{ l(\beta) - \lambda \sum_{p=1}^{P} |\beta_p| \right\} \quad (1)$$

We proposed the use of the cross-validation algorithm recently reported by Goeman (2011) that efficiently estimates λ.

The ability of the score η to predict the event up to time τ

If $\eta_j > c$ then the event is predicted before τ.The false negative and positive rate are respectively:

$$FNR_\tau(c) = \{G(c) - S(-\infty,\tau) + S(c,\tau)\}/\{1 - S(-\infty,\tau)\} \quad (2)$$

$$FPR_\tau(c) = S(c,\tau)/S(-\infty,\tau) \quad (3)$$

where $G(c)$ is the empirical distribution function and $S(c,\tau)$ is the joint bivariate survival distribution of (η, T) estimated using the Akritas estimator involved in the Heagerty framework (2000).

The principle of the iterative bootstrap procedure is to estimate the regression parameters (1) and the associated value of λ for each bootstrap sample. The corresponding estimations of (2) and (3) allow to obtain the **apparent error rates** by performing the average over all the bootstrap sample. These estimations (2) and (3) can also be performed for each sample based on the patients not included in the bootstrap sample. The average of these values gives **the bootstrap cross-validation (BCV) error rates**.

	False Negative Rate	False Positive Rate
Apparent	$FNR_{a,\tau}(c)$	$FPR_{a,\tau}(c)$
BCV	$FNR_{b,\tau}(c)$	$FPR_{b,\tau}(c)$

The 0.632+ estimator

The no-information rates respectively associated with the FNR and FPR may be estimated using all the data and considering the independence between η and T: $\gamma_{N,\tau}(c) = 1 - \gamma_{P,\tau}(c) = G(c)$. These no-information probabilities are used to define the overfitting rates. For $K = (N, P)$:

$$r_{K,\tau}(c) = \{FKR_{b,\tau}(c) - FKR_{a,\tau}(c)\}/\{\gamma_{K,\tau}(c) - FKR_{a,\tau}(c)\} \quad (4)$$

We assigned these rates to 0 for negative values and to 1 for values higher than 1. The 0.632+ estimations of the FKR, $K = (N, P)$, are thus defined by:

$$FKR_{.632,\tau}(c) = \{1 - \psi(r_{K,\tau}(c))\}FKR_{a,\tau}(c) + \psi(r_{K,\tau}(c))\, FKR_{a,\tau}(c) \quad (5)$$

where $\psi(x) = 0.632/(1 - 0.368x)$. The corresponding 0.632+ ROC curve for a prognostic up to time τ is $\{FPR_{.632,\tau}, 1 - FNR_{.632,\tau},\ c \in \mathcal{R}\}$.

Proposition of a R package

This method has been implemented in an **R package called ROC632** available at **www.divat.fr/en/softwares** or upon request from authors.

Validation by simulations

Different values of N were used: 60, 125 and 250. Only the results for $N = 250$ are presented in the following table. The times-to-event were simulated using a Weibull PH model. The feature expressions were obtained by assuming independent standard normal distributions. The censoring times were simulated independently respecting Exponential distributions to obtain three different censoring rates at 6 months: 30, 50 and 70%. We distinguished the 3 following scenarios:

- ✓ **Total overfitting.** Among 750 features, no feature is associated with the time-to-event, the true value of the AUC is 0.5.
- ✓ **No overfitting.** Among 3 features, 2 features are associated with the time-to-event, the true value of the AUC is obtained by using the apparent estimator.
- ✓ **High overfitting.** Among 750 features, 2 features are associated with the time-to-event. We added 747 features independently associated with the time-to-event to the previous scenario. The true value of the AUC is thus similar to the apparent estimation in the second scenario.

For each possible combination of the overfitting levels, sample sizes and censoring rates, 250 samples were simulated. 100 bootstrap iterations were used for each simulated sample.

Adequate correction of the overfitting

Overfitting level	Censoring rate	Apparent	0.632+
Total (0/750)	0.3	0.799 (0.035)	0.521 (0.046)
	0.5	0.777 (0.062)	0.516 (0.054)
	0.7	0.693 (0.121)	0.519 (0.067)
No overfitting (2/3)	0.3	0.838 (0.032)	0.830 (0.033)
	0.5	0.836 (0.033)	0.824 (0.035)
	0.7	0.828 (0.047)	0.806 (0.051)
High overfitting (2/750)	0.3	0.811 (0.027)	0.843 (0.025)
	0.5	0.789 (0.037)	0.825 (0.034)
	0.7	0.729 (0.060)	0.766 (0.059)

Table: Mean and standard deviation (between brackets) of the time dependent area under the ROC curves at to 6 months

No correction when no overfitting

Adequate correction of the overfitting

Warning: These adequate estimations were also observed for N =125 with 30% of censoring. For the worst situations (lower sample sizes and/or higher censoring rates), the 0.632+ estimator underestimated the prognostic capacity.

The application to the DLBCL study

Rosenwald (2002) evaluated tumor samples from 240 DLBCL (diffuse large-B-cell lymphoma) patients treated with anthracycline-based therapy. The full dataset was split into training (N=160) and test (N=80) sets, which is associated to an increase of the type II error.

The overfitting was high with an apparent AUC around 0.95. The AUCs obtained by using the 0.632+ estimator were between 0.70 and 0.65 (depending on the prognostic time between 2 and 14 years). This illustrates the utility of this signature to predict mortality up to 14 years, but it also illustrates that this signature alone is not sufficient for medical decision-making. Indeed, a patient who will die before 10 years has a 32% chance of having a score lower than a patient who will be alive at this time.

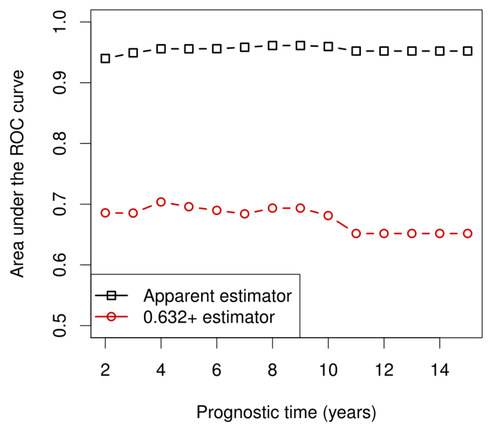

Figure: AUC according to the prognostic times (DLBCL data)

Conclusions

We propose a 0.632+ estimator of the time-dependent ROC curve to exceed the latter limitations. First, ROC-based methodologies are well-accepted in the community of biologists and physicians. The AUC represents the ability of a prognostic factor to correctly distinguish patients who will develop events in the future from those who will not.

Second, the time-dependent FNR and FPR are conditional distributions of the prognostic index assuming the distribution of the time-to-event). The AUC is an invariant prognostic indicator among populations with different incidences.

Third, the tuning parameter is re-estimated at each bootstrap iteration according to the recent efficient algorithm of Goeman (2011).

We validated the methodology by simulating three overfitting levels (total, high, low) according to different sample sizes and censoring rates. If the available information is sufficient (N = 250 with less than 50% of censoring or N = 125 with less than 30% of censoring), the 0.632+ estimators appeared to be similar to the true expected areas under the curve.

The approach can be affected by the model misspecification, as all the regression analysis. Given the high dimension data, it is not realistic to test every assumption for every feature. However, the algorithm has been formulated in a general way and can be subsequently applied to other models.

References

M Schumacher, H Binder, and T Gerds. Assessment of survival prediction models based on microarray data. *Bioinformatics*, 23:1768–1774, 2007.

PJ Heagerty, T Lumley, and SP Pepe. Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker. *Biometrics*, 56:337–44,2000.

R Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B,* 58:267–288, 1996.

J Goeman. L1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal*, 52(1):70–84, 2010.

A Rosenwald et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New England Journal of Medicine*, 346(25):1937–47, 2002.

Divat Données Informatisées et VAlidées en Transplantation www.divat.fr & EA-4275 Biostatistique, Pharmacoépidémiologie et Mesures Subjectives en Santé www.erbiomes.anaqol.org

Statistical Applications in Genetics and Molecular Biology. 2012. In press.