Supplemental Materials: Machine Learning for Survival Analysis: A Survey

PING WANG, Virginia Tech
YAN LI, University of Michigan, Ann Arbor
CHANDAN K. REDDY, Virginia Tech

The purpose of this supplemental material is to provide more details on the following topics. (i) Section 1 introduces the detailed algorithms for both the Nelson-Aalen (NA) estimator and Life Table (LT) method in non-parametric methods. (ii) In Section 2, both the C-index and the Brier Score (BS) are adapted to evaluate prediction performance over a time interval. (iii) Section 3 provides more details about handling complex events in survival analysis.

1 NON-PARAMETRIC METHODS

In this section, we introduce two non-parametric methods in detail, the NA estimator and the LT method.

1.1 Nelson-Aalen Estimator

The NA estimator (Nelson 1972; Aalen 1978) is a non-parametric estimator of the Cumulative Hazard Function (CHF) for censored (or incomplete) data. It was first introduced by Nelson in the context of reliability (Nelson 1972) and later Aalen (Aalen 1978) rediscovered and derived the estimator using modern counting process techniques. The NA estimator of the CHF is defined as:

$$\hat{H}(t) = \sum_{t_i \leq t} \frac{d_i}{r_i}, \tag{1}$$

where d_i is the number of deaths at time t_i and r_i is the number of individuals at risk at t_i. This cumulative hazard rate function can be used to estimate the survival function as follows:

$$\hat{S}(t) = e^{-\hat{H}(t)}. \tag{2}$$

The NA and Kaplan-Meier (KM) estimators are asymptotically equivalent. The differences between them are minor, particularly at early survival time points (Lee and Wang 2003).

1.2 Life Table Method

The clinical LT method (Cutler and Ederer 1958) is effectively the application of the KM method to the interval grouped survival data. It is generated based on the conditional probability of events within an interval, with the total number of subjects N being partitioned into J intervals based on the observed times of the events occur. The j^{th} interval, normally denoted I_j, is defined as $I_j = [t_j, t_{j+1}), j = 0, 1, \ldots, J$, where $t_0 = 0$ and $t_{J+1} = \infty$, and the length of I_j is $l_j = t_{j+1} - t_j$.

For I_j, let r'_j be the number of survivors at the beginning of the jth interval, c_j be the number of censored cases during the jth interval and d_j be the number of deaths within the j^{th} interval. We then have $r'_1 = n$ and $r'_j = r'_{j-1} - d_{j-1} - c_{j-1}, j = 2, \ldots, J + 1$.

0360-0300/2019/02-ART110 $15.00
https://doi.org/10.1145/3214306

Table 1. Summary of Characteristics of Different Statistical Methods for Survival Analysis

Type	Methods	Assumptions	What it does	Optimization	Complexity
Non-parametric	KM	The distribution of time is not required.	Estimates the survival function.	Not required.	$O(N)$
	NA		Estimates the cumulative hazard function.		
	LT		Estimates the survival function for grouped data.		
Semi-parametric	Cox	PH assumption.	Baseline hazard function is not specified.	Maximum partial likelihood estimator (MPLE) and Newton-Raphson (NR) method.	$O(NP^2)$
	Ridge-Cox		Selects the correlated features.		
	Lasso-Cox		Selects features and estimates parameters simultaneously.	MPLE and Coordinate Gradient Descent (CGD).	$O(NP)$
	EN-Cox		Selects features and deals with correlations between features simultaneously.		
	Oscar-Cox		Selects features for highly correlated features.		
	CoxBoost		Considers mandatory features.	Gradient boosting.	$O(NP^2)$
	TD-Cox	PH assumption does not hold.	Considers time-dependent features.	MPLE and NR method.	
Parametric	Tobit	The survival times or the logarithm of the survival times of all instances are assumed to follow a particular distribution.	Adapts linear regression for survival analysis.	Maximum-likelihood estimation (MLE) and NR method.	$O(NP^2)$
	AFT		Features have multiplicative effects on the outcome.		
	BJ		Combines KM and AFT.		
	Weighted regression		Handles censored data by assigning weights to different instances.	MPLE and CGD.	$O(NP)$
	Structured regularization		Infers the underlying structure of survival data.		

Suppose we have the survival function $S(t)$ and define

$$P_j = S(t_j) = P(T \geq t_j)$$

$$p_j = P(T \geq t_j | T \geq t_{j-1}) = \frac{S(t_j)}{S(t_{j-1})} = \frac{P_j}{P_{j-1}}$$

$$q_j = P(t_{j-1} \leq T < t_j | T \geq t_{j-1}) = 1 - p_j$$

for $j = 2, \ldots, J+1$. Thus, we have

$$S(t_j) = p_j S(t_{j-1}) = \cdots = p_j p_{j-1} \cdots p_2 p_1 P_0 = \prod_{i=1}^{j} p_i \tag{3}$$

since $P_0 = P(T \geq 0) = 1$. To estimate $S(t_j)$, we need to estimate p_j or q_j. Note that if $c_j = 0$ ($\forall j = 1, \ldots, J$), then $\hat{q}_j = d_j / r'_j$. However, if $c_j \neq 0$, $\hat{q}_j = d_j / r'_j$ is expected to underestimate q_j, since it is possible that individuals censored in I_j might die before the end of I_j. The problem becomes more complicated due to the fact that we do not know exactly when an event occurs during each time interval. Thus, in the standard LT method, $r_j = r'_j - c_j/2$ is assumed to be the number of survivors on average half-way through the interval. This is appropriate if the censorings occur uniformly throughout the interval.

Table 2. Summary of Properties of Machine Learning Methods for Survival Analysis

Methods	Assumptions		Optimization	Complexity
Survival trees	Classification and regression trees are integrated for survival analysis.		Varies with different methods.	Varies with different methods.
Bayesian methods	Predicts the probability of an event.		Varies with different methods.	Varies with different methods.
Neural networks (NN)	Combines Cox model with a NN or utilizes the output of NN as the survival or hazard probability.		Varies with different methods.	Varies with different methods.
SVM	Adapt support vector regression for survival analysis.		Sequential minimal optimization (SMO)	$O(N^3)$
RSF	Adapts tree structured models for survival analysis.	Average predictions made by a single survival tree.	Varies with different survival analysis methods.	$O(NP^2)$
BST		Average predictions of all survival trees.		
Boosting	Extends the gradient boosting algorithm to solve survival analysis problems.		Gradient boosting	Varies with different methods.
Active learning	Integrates active learning method with the Cox model to take into account the opinions of an expert.		Coordinate majorization descent	$O(NPK)$
Transfer learning	Performs knowledge transfer from related tasks to solve insufficient data problems.		Fast iterative shrinkage thresholding algorithm (FISTA)	$O(NP)$
Multi-task learning	Captures dependencies between the outcomes at various time points using a shared representation across related tasks.		Alternating direction method of multipliers (ADMM) algorithm	$O(NPK)$

Then, as in the KM estimator, the conditional probability of surviving during the j^{th} interval is estimated as

$$\hat{p}_j = 1 - \frac{d_j}{r_j}. \tag{4}$$

The corresponding survival function is estimated by the product

$$\hat{S}(t_j) = \prod_{i:i<j} \left(1 - \frac{d_i}{r_i}\right). \tag{5}$$

The recursive formula and the variance of $\hat{S}(t_j)$ can be calculated in a manner similar to that used in the KM method. Based on the survival function obtained above, we can also estimate the death density function and the hazard function. Let t_{mj} denote the midpoint of $[t_{j-1}, t_j)$ and $b_j = t_j - t_{j-1}$. The death density function at t_{mj} is estimated as

$$\hat{f}(t_{mj}) = \frac{\hat{S}(t_{j-1})\hat{q}_j}{b_j}. \tag{6}$$

The hazard function at t_{mj} is estimated as

$$\hat{h}(t_{mj}) = \frac{2\hat{q}_j}{b_j(1+\hat{p}_j)}. \tag{7}$$

2 PERFORMANCE EVALUATION METRICS

In this section, we adapt both the C-index and the BS to evaluate the prediction performance of survival models during a time interval.

2.1 C-index

In order to evaluate the performance during a follow-up period, Heagerty and Zheng defined the C-index for a fixed follow-up time period $(0, t^*)$ as the weighted average of area under the curve (AUC) values at all possible observation time points (Heagerty and Zheng 2005). The time-dependent AUC for any specific survival time t can be calculated as

$$AUC(t) = P(\hat{y}_i < \hat{y}_j | y_i < t, y_j > t) = \frac{1}{num(t)} \sum_{i:y_i<t} \sum_{j:y_j>t} I(\hat{y}_i < \hat{y}_j), \tag{8}$$

where $t \in T_s$, the set of all possible survival times, and $num(t)$ represents the number of comparable pairs for the time point t. The C-index during the time period $(0, t^*)$, which is the weighted average of the time-dependent AUC obtained by Equation (8), can now be computed as

$$c_{t^*} = \frac{1}{num} \sum_{i:\delta_i=1} \sum_{j:y_i<y_j} I(\hat{y}_i < \hat{y}_j) = \sum_{t \in T_s} AUC(t) \cdot \frac{num(t)}{num}. \tag{9}$$

Thus c_{t^*} is the probability that the predictions are concordant with their outcomes for a given dataset during the time period $(0, t^*)$.

2.2 Brier Score

When predictions are assessed over a time period $(0, t^*)$, rather than for a particular time point t, the prediction error can be averaged over the time interval by using the Integrated BS (IBS) (Graf et al. 1999), as shown in Equation (10):

$$IBS = \frac{1}{N} \sum_{i=1}^{N} \int_0^{t^*} w_i(t)[\hat{y}_i(t) - y_i(t)]^2 dW(t), \tag{10}$$

where $W(t)$ is a weight function, for which the natural choices are $W(t) = t/t^*$ or $W(t) = (1 - \hat{S}(t))/(1 - \hat{S}(t^*))$, where $\hat{S}(t)$ denotes the estimated survival function. Based on the definition of the BS given above, it is evident that this measures the mean squared difference between the predictions made and the actual outcomes; therefore, the lower the BS, the better the prediction model.

3 COMPLEX EVENTS

3.1 Competing Risks

To address the competing risks problem, the standard approach is to analyze each of these events separately using a survival analysis and considering other competing events as censored (Kleinbaum and Klein 2006). However, there are two primary drawbacks with such an approach. The first is that this method assumes that the competing risks are independent of each other, which is not always the case. Second, it would be difficult to interpret the survival probability estimated for each event separately by performing a survival analysis for each event of interest when competing

risks are involved. To overcome these drawbacks, two methods have been proposed in the survival analysis literature: the Cumulative Incidence Curve (CIC) and the Lunn-McNeil (LM) Approach.

Cumulative Incidence Curve (CIC) Approach: To avoid the questionable interpretation problem, the CIC (Putter et al. 2007) is often utilized for competing risks to estimate the marginal probability of each event q. The CIC is defined as

$$CIC_q(t) = \sum_{j:t_j \leq t} \hat{S}(t_{j-1})\hat{h}_q(t_j) = \sum_{j:t_j \leq t} \hat{S}(t_{j-1})\frac{n_{qj}}{n_j}, \tag{11}$$

where $\hat{h}_q(t_j)$ represents the estimated hazard at time t_j for event q ($q = 1, \ldots, Q$), n_{qj} is the number of events for the event q at t_j, n_j denotes the number of instances that are at risk of experiencing events at t_j, and $S(t_{j-1})$ denotes the survival probability at the previous time point t_{j-1}.

Lunn-McNeil (LM) Approach (Lunn and McNeil 1995): This is an alternative approach that can be applied to analyze the competing risks in survival problems, providing the flexibility to conduct statistical inferences from the features in standard competing risk models. It fits a single Cox PH model that considers all the events in competing risks rather than constructing separate models for each event (Kleinbaum and Klein 2006). Note that the LM approach is implemented using augmented data, where a dummy variable is created for each event to distinguish between different competing risks.

3.2 Recurrent Events

The following two approaches can be used to tackle the recurrent events problem.

Counting Process: In the Counting Process (CP) method, the data processing procedure is as follows: (i) For each instance, identify the time interval for each recurrent event and add one record to the data. Note that an additional record for the event-free time interval should also be included for each instance. (ii) For each instance, each data record should be labeled with the start time and end time of the corresponding time interval. These data format properties distinguish the CP method from other methods as they are significantly different from the regular survival data format for nonrecurrent event problems, which provides only the end time and contains only one record for each instance in the dataset.

The key idea when analyzing the survival data containing recurrent events is to treat the different time intervals for each instance as independent records from different instances. The basic Cox model is used to perform the CP approach. Each instance will not be removed from the risk set until the last time interval during the observation period. This means that, in survival problems with recurrent events, the partial likelihood function formula is different from that used for nonrecurrent event survival problems (Kleinbaum and Klein 2006).

Stratified Cox: The Stratified CP (Prentice et al. 1981), Marginal (Wei et al. 1989), and Gap Time (Prentice et al. 1981) approaches all use the stratified Cox method to differentiate the event occurrence order. (i) In the Stratified CP approach, the data format is exactly the same as that used in the CP approach, and the risk set for future events depends on the time of the first event. (ii) The Marginal approach uses the same data format as for nonrecurrent event survival data. This method considers the length of the survival time from the starting time of the follow-up until the time of a specific event occurrence and assumes that each event is independent of other events. For the k^{th} event ($k = 1, 2, \ldots$), the risk set therefore contains those instances which are at risk of experiencing the corresponding event after their entry into the observation. (iii) In the Gap Time approach, the data format (start, stop) is used, but the start time for each data record is 0 and the end time is the length of the interval from the previous experienced event. In this method, the risk set for future events will not be affected by the time of the first event.

REFERENCES

Odd Aalen. 1978. Nonparametric inference for a family of counting processes. *The Annals of Statistics* 6, 4 (1978), 701–726.

Sidney J. Cutler and Fred Ederer. 1958. Maximum utilization of the life table method in analyzing survival. *Journal of Chronic Diseases* 8, 6 (1958), 699–712.

Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. 1999. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* 18, 17-18 (1999), 2529–2545.

Patrick J. Heagerty and Yingye Zheng. 2005. Survival model predictive accuracy and ROC curves. *Biometrics* 61, 1 (2005), 92–105.

David G. Kleinbaum and Mitchel Klein. 2006. *Survival Analysis: A Self-learning Text*. Springer Science & Business Media.

Elisa T. Lee and John Wang. 2003. *Statistical Methods for Survival Data Analysis*. Vol. 476. John Wiley & Sons.

Mary Lunn and Don McNeil. 1995. Applying Cox regression to competing risks. *Biometrics* (1995), 524–532.

Wayne Nelson. 1972. Theory and applications of hazard plotting for censored failure data. *Technometrics* 14, 4 (1972), 945–966.

Ross L. Prentice, Benjamin J. Williams, and Arthur V. Peterson. 1981. On the regression analysis of multivariate failure time data. *Biometrika* 68, 2 (1981), 373–379.

Hein Putter, M. Fiocco, and R. B. Geskus. 2007. Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in Medicine* 26, 11 (2007), 2389–2430.

Lee-Jen Wei, Danyu Y. Lin, and Lisa Weissfeld. 1989. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* 84, 408 (1989), 1065–1073.