Published in final edited form as:

Stat. 2014 ; 3(1): 337–350. doi:10.1002/sta4.67.

A Model-Free Machine Learning Method for Risk Classification and Survival Probability Prediction

Yuan Geng^a, Wenbin Lu^{b,*}, and Hao Helen Zhang^c

^aBoehringer Ingelheim International Trading Co., Ltd., Shanghai, 200040, P. R. China

^bDepartment of Statistics, North Carolina State University, Raleigh, NC 27695, USA

^cDepartment of Mathematics, University of Arizona, Tucson, AZ 85721, USA

Abstract

Risk classification and survival probability prediction are two major goals in survival data analysis since they play an important role in patients' risk stratification, long-term diagnosis, and treatment selection. In this article, we propose a new model-free machine learning framework for risk classification and survival probability prediction based on weighted support vector machines. The new procedure does not require any specific parametric or semiparametric model assumption on data, and is therefore capable of capturing nonlinear covariate effects. We use numerous simulation examples to demonstrate finite sample performance of the proposed method under various settings. Applications to a glioma tumor data and a breast cancer gene expression survival data are shown to illustrate the new methodology in real data analysis.

Keywords

Model-free; Risk classification; Support vector machines; Survival probability prediction

1. Introduction

Survival probability prediction is one important problem encountered in medical studies when the primary endpoint of interest is time to an event. An accurate survival probability prediction can provide useful information for patients' risk prediction and stratification, which is critical in selecting prevention and treatment strategies. In the past few decades, extensive research has been dedicated to survival probability prediction, and various parametric and semiparametric models have been in use for analysis of survival data. Among many available tools, the proportional hazards (PH) model proposed by Cox (1972) has become the most popular model for survival probability prediction due to its flexibility and simplicity. A key assumption of the PH model is that the hazard ratio corresponding to different covariates is constant over time, which however may not hold in real applications. For example, in a randomized study for glioma tumor (Piantadosi, 1997), it was found that the estimated survival curves using the Kaplan-Meier method (Kaplan and Meier, 1958) for

the treatment and control groups cross around 50 weeks after receiving treatment, which suggests a possible violation of the PH model assumption. A variety of semiparametric models have been proposed for the situations where the PH assumption is not appropriate, including the additive hazards model (Aalen, 1980), proportioned odds (PO) model (Bennett, 1983), accelerated failure time (AFT) model (Cox and Oakes, 1984), linear transformation model (Clayton and Cuzick, 1985), and many others. These models provide useful alternatives to the PH model and give better survival probability prediction in some applications. For convenience, many commonly used semiparametric survival models assume the covariate effects to be linear. However, the true covariate effects can be much more complicated, and it is hard to pre-determine their parametric forms in practice. Based on this, model-based extensions to incorporate nonlinear/partially linear or time-varying covariate effects have been extensively studied.

Although semiparametric models are more flexible than parametric models for survival probability prediction, they still need to make certain assumptions on survival times. A number of testing procedures have been developed to check the goodness-of-fit of a particular model, for example, the test based on the cumulative martingale residual process proposed by Lin et al. (1993). However, it is generally difficult to select the optimal semiparametric model from a number of candidate models. When the model is misspecified, the parameter estimates could be biased, which may deteriorate prediction of survival probability. To further relax model assumptions, nonparametric methods for survival probability estimation have been studied in the literature, such as the stratified Kaplan-Meier estimator and the local Kaplan-Meier estimator (Dabrowska, 1987). These methods are useful only when there are a relatively small number of predictors due to the curse of dimensionality. Recently, Ishwaran et al. (2008) developed random survival forests for nonparametric survival probability prediction and Wang and Wang (2010) developed a Buckley-James boosting method for survival analysis with high-dimensional covariates.

Besides prediction of survival probability, risk classification is another important tool commonly used in treatment selection. In literature, the decision of risk classification is typically made based on the estimated survival probabilities of subjects. For example, subjects with t -year survival probability greater than a cut-off value are labeled as *low-risk*, and otherwise labeled as *high-risk*. A reliable risk classification procedure is critical, since doctors may rely on it to suggest personalized treatments that are suitable for high-risk and low-risk patients, respectively.

In this article, we develop a new machine learning framework for risk classification and survival probability prediction with censored survival data. The proposed estimation scheme is model-free, because it does not require any specific modeling assumption on survival times. The main idea is to develop an inverse probability of censoring weighted (IPCW) procedure based on large-margin classifiers called weighted Supported Vector Machines (wSVM) (Wang et al., 2008). We construct a nonparametric estimator for t -year survival probability by assembling a series of the wSVM decision rules associated with different weights. Compared to existing methods, a main advantage of the new approach, named the IPCW-wSVM, is that it makes less assumptions on data distribution and is hence more robust against possible model misspecification. An effective computation and tuning

algorithm is presented. We show that the IPCW-wSVM gives promising performance in estimating nonlinear classification boundaries and is capable of handling a large number of covariates.

The rest of the paper is organized as follows. Section 2 introduces the IPCW-wSVM methods for risk classification and survival probability prediction. Section 3 describes the computational algorithm and the model tuning procedure. Simulation studies are presented in Section 4. Two real data applications are shown in Section 5. Section 6 concludes the paper with discussions.

2. New Methodology

This section describes a new class of large-margin methods for risk classification and survival probability prediction in censored data analysis. These two problems have different goals but are closely related to each other. In the machine learning context, they are respectively treated as “hard” vs “soft” classification problems and useful for different purposes (Wahba, 2002).

2.1. Risk Classification

Consider a survival study of n subjects. Let T_i and C_i denote the failure time and the censoring time of the i th subject, respectively, and let $Z_i \sim Z$ be the associated p -dimensional vector of covariates. The observed data are denoted by (T_i, δ_i, Z_i) , where $T_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$. Throughout the paper we assume that T_i and C_i are independent given Z_i . In many disease studies, it is critical to accurately predict whether a patient can survive t years based on his/her covariate information, for some value t . For subject i , define $Y_i = 2I(T_i > t) - 1$. In other words, $Y_i = +1$ represents a “low risk” class while $Y_i = -1$ represents a “high risk” class. Given a patient’s covariate value z , define $p_t(z) \equiv P(T > t | Z = z) = P(Y = 1 | Z = z)$, the conditional t -year survival probability given z .

The primary goal of risk classification is to learn a decision function $f: Z \rightarrow \{-1, +1\}$ using the training sample. From the standard learning theory, the Bayes rule assigns a subject with covariate z to $+1$ class (low risk) if $p_t(z) \geq \pi$, and to -1 class (high risk) otherwise, where $\pi \in (0, 1)$ is a value chosen to minimize the expected misclassification cost. Most standard classification tools are designed to estimate the Bayes rule by first obtaining the probability estimator $\hat{p}_t(z)$ and then constructing the decision rule as $\hat{f}(z) = \text{sign}\{\hat{p}_t(z) - \pi\}$. For example, $\hat{p}_t(z)$ can be an estimator given by logistic regression. Alternatively, some hard classifiers like SVMs can directly estimate the sign of the Bayes rule.

When there is censoring in data, the true label Y_i ’s are not observable for subjects who are censored before time t . This missing data caused by censoring makes it difficult to learn the classification boundary. Define the observed label $\tilde{Y}_i = 2I(T_i > t) - 1$, $i = 1, \dots, n$. Then $\tilde{Y}_i = 1$ implies $Y_i = 1$, and $\tilde{Y}_i = -1$ and $\delta_i = 1$ implies $Y_i = -1$. Let $G(t|z)$ denote the survival function of censoring times given the covariate z , i.e., $G(t|z) = P(C_i > t | Z_i = z)$. We then have

$$\begin{aligned} E \left\{ I \left(\tilde{T}_i > t \right) / G(t|Z_i) | Z_i \right\} &= E \left\{ I \left(T_i > t \right) | Z_i \right\}, \\ E \left\{ \delta_i I \left(\tilde{T}_i \leq t \right) / G \left(\tilde{T}_i^- | Z_i \right) | Z_i \right\} &= E \left\{ I \left(T_i \leq t \right) | Z_i \right\}, \end{aligned}$$

where $G(t^-|z)$ is the left limit of $G(\cdot|z)$ at t . For any given π , we propose to learn the risk classification rule $f_\pi \in F$ by minimizing the regularized IPCW-wSVM loss

$$\min_{f_\pi \in F} \frac{1}{n} \sum_{i=1}^n \left\{ (1-\pi) \frac{I(\tilde{T}_i > t)}{\hat{G}(t|Z_i)} + \pi \frac{\delta_i I(\tilde{T}_i \leq t)}{\hat{G}(\tilde{T}_i^- | Z_i)} \right\} \left\{ 1 - \tilde{Y}_i f_\pi(Z_i) \right\}_+ + \frac{\lambda}{2} J(f_\pi), \quad (1)$$

where $\hat{G}(t|z)$ is a consistent estimator of $G(t|z)$, F is some function space containing f_π , $J(\cdot)$ is a measure of function complexity, and λ is a tuning parameter which balances the fit to data and the penalty term. In (1), each data point $(z_i, \tilde{Y}_i)$ is assigned with a weight $1 - \pi$ or π , depending on whether $T_i > t$ or $T_i \leq t$. The subscript π in f_π is used to emphasize that the solution is associated with the weight parameter π . The IPCW technique has been widely studied in survival analysis (e.g. Koul et al., 1981; Fan and Gijbels, 1994; Cheng et al., 1995, 1997; Fine et al., 1998; Lu and Li, 2011). Typically, the survival function of censoring times is usually estimated by the Kaplan-Meier method based on the assumption that censoring times are independent of both failure times and covariates. This assumption is reasonable for some clinical studies where study subjects are well followed and the majority of censoring is due to administrative censoring. However, it may be restrictive for some applications. Alternatively, a semiparametric model, like the PH model, can be posited for censoring times and used to estimate $G(t|z)$ based on the fitted model. The estimator is consistent as long as the model is correctly specified, though the nonparametric flavor of IPCW-wSVM is somewhat compromised. In our numerical studies presented in Sections 4 and 5, we use the Kaplan-Meier estimator for the survival function of the censoring times.

The optimization problem (1) provides a general framework for learning risk classification rules, allowing one to fit both linear and nonlinear classifiers. For linear classification, we assume $f_\pi(z) = b_\pi + \beta'_\pi z$ and $J(f_\pi) = \|\beta_\pi\|^2$. To achieve nonlinear classification, we assume $f_\pi(z) = b_\pi + h_\pi(z)$, where $h_\pi(z)$ is a function belonging to some function space, say, a reproducing kernel Hilbert space (RKHS) H_K associated with kernel $K(\cdot, \cdot)$ and norm $\|\cdot\|_{H_K}$, and $J(f_\pi) = \|h_\pi\|_{H_K}^2$. By the representer theorem (Kimeldorf and Wahba, 1971), the solution to (1) has a finite representation form, i.e., $h_\pi(z) = \sum_{i=1}^n c_{\pi,i} K(z, Z_i)$, where $c_\pi = (c_{\pi,1}, \dots, c_{\pi,n})'$ and K is the Gram matrix whose ij th entry is $K(Z_i, Z_j)$. This leads to $J(f_\pi) = c'_\pi K c_\pi$ and (1) can be written as the following equivalent optimization problem,

$$\min_{b_\pi, c_\pi} \frac{1}{n} \sum_{i=1}^n \left\{ (1-\pi) \frac{I(\tilde{T}_i > t)}{\hat{G}(t|Z_i)} + \pi \frac{\delta_i I(\tilde{T}_i \leq t)}{\hat{G}(\tilde{T}_i^- | Z_i)} \right\} \left[1 - \tilde{Y}_i \left\{ b_\pi + \sum_{j=1}^n c_{\pi,j} K(Z_i, Z_j) \right\} \right]_+ + \frac{\lambda}{2} c'_\pi K c_\pi. \quad (2)$$

The estimated classification rule is $\hat{f}_\pi(z) = \hat{b}_\pi + \sum_{i=1}^n \hat{c}_{\pi,i} K(z, Z_i)$, where $\hat{b}_\pi$ and $\hat{c}_{\pi,j}$'s are the solution to (2). The regularization problem (2) can be further transformed into its dual

form and solved by quadratic programming (QP). Computation details are provided in Section 3. Next we provide theoretical justifications for (1) by Lemma 1. In particular, we show that the IPCW-wSVM is Fisher consistent for censored data risk classification, as it directly estimates the Bayes rule $\text{sign}\{p_t(z) - \pi\}$.

Lemma 1—Provided that $\hat{G}(t|z)$ is a consistent estimator of $G(t|z)$ for any $t \in [0, \tau]$ and $G(\tau|z) - \varepsilon > 0$ for any covariate value z , where τ is the maximum follow-up time and ε is a constant, the IPCW-wSVM loss converges to

$$E \left[\{(1 - \pi) I(Y=1) + \pi I(Y=-1)\} \{1 - Y f_{\pi}(Z)\}_+ \right] \quad (3)$$

in probability as $n \rightarrow \infty$.

Proof of Lemma 1 is straightforward and omitted here. Lemma 1 essentially extends the Fisher consistency result in Wang et al. (2008) from the non-censored data context to censoring situations. Following the similar arguments of Lin (2002), when λ is properly chosen and the RKHS is rich enough, we can show that the IPCW-wSVM classifier given by (2) is consistent to $\text{sign}\{p_t(z) - \pi\}$.

2.2. Survival Probability Prediction

In survival analysis, the survival probability $p_t(z) = P(Y = 1|Z = z)$ is another important measure to assess a patient's survival status. In this section, we develop a mode-free procedure to estimate $p_t(z)$ by aggregating multiple classification rules learned from (1). The procedure consists of two steps: we first train a series of classifiers $\hat{f}_{\pi}(z)$ by solving (2) with various weights π 's from 0 to 1, and then extract the probability information from those decision rules.

Specifically, consider a set of grid points $0 = \pi_1 < \pi_2 < \dots < \pi_m < \pi_{m+1} = 1$. For each $j = 1, \dots, m+1$, we train the risk classifier $\hat{f}_{\pi_j}(z)$ by solving (2). In theory, for any fixed z , the quantity $\text{sign}\{p_t(z) - \pi\}$ is monotonically decreasing in π . This implies that $p_t(z)$ must fall into one of the intervals $[\pi_j, \pi_{j+1}]$, which satisfies $\text{sign}\{p_t(z) - \pi_j\} = 1$ and $\text{sign}\{p_t(z) - \pi_{j+1}\} = -1$. In other words, there exists a unique j such that $\text{sign}\{\hat{f}_{\pi_i}(z)\} = 1$ for $i \leq j$ and $\text{sign}\{\hat{f}_{\pi_i}(z)\} = -1$ for $i > j$. Based on this fact, a reasonable estimator for $p_t(z)$ would be the average of π_j and π_{j+1} . With finite samples, due to numerical variability in solutions, there is no guarantee on the monotonicity of $\text{sign}\{\hat{f}_{\pi}(z)\}$ in π . So we estimate $p_t(z)$ by $\frac{1}{2}(\pi^* + \pi^*)$, where $\pi^* = \arg \max_{\pi_j} \{\text{sign}\{\hat{f}_{\pi_j}(z)\} = 1\}$ and $\pi^* = \arg \min_{\pi_j} \{\text{sign}\{\hat{f}_{\pi_j}(z)\} = -1\}$, following the suggestion of Wang et al. (2008). A detailed step-by-step description is given as follows.

Step 1. Set $\pi_j = (j-1)/m, j = 1, \dots, m+1$.

Step 2. Train a series of IPCW-wSVM classifiers $\hat{f}_{\pi_j}(z)$ by solving (2) for $j = 1, \dots, m+1$ with different π_j 's.

Step 3. Estimate the label of z for each π_j by $\text{sign}\{\hat{f}_{\pi_j}(z)\}$.

Step 4. Find $\pi^* = \arg \max_{\pi_j} \{\text{sign}\{\hat{f}_{\pi_j}(z)\} = 1\}$ and $\pi_* = \arg \min_{\pi_j} \{\text{sign}\{\hat{f}_{\pi_j}(z)\} = -1\}$. Construct the t -year survival probability estimator as $p_t(\hat{z}) = 1/2 (\pi_* + \pi^*)$.

The estimator $p_t(\hat{z})$ is a model-free probability estimation procedure, since it does not require any parametric or semiparametric assumption on the survival time distribution or the hazard function. It is easy to see that, as m increases, the interval that covers $p_t(z)$ becomes narrower. The larger m is, the better accuracy is achieved in estimating the survival probability; on the other hand, the computational cost is higher. In our simulation studies, m is set to be 100, which gives satisfactory results for the considered scenarios. When t is smaller than the smallest observed failure time, the estimated survival probability is equal to 1 for all subjects since no one is assigned to -1 class.

3. Computation and Tuning

In this section, we first present an effective computation algorithm to train the IPCW-wSVM classifier. Then the issue of parameter tuning issue will be discussed.

3.1. Computation Algorithm

Let $w_i = (1 - \pi) \frac{I(\tilde{T}_i > t)}{\hat{G}(t|Z_i)} + \pi \frac{\delta_i I(\tilde{T}_i \leq t)}{\hat{G}(\tilde{T}_i^- | Z_i)}$. For simplicity, we drop the subscript π in (2) and rewrite the problem as

$$\min_{b,c} \sum_{i=1}^n w_i \left[1 - \tilde{Y}_i \left\{ b + \sum_{j=1}^n c_j K(Z_i, Z_j) \right\} \right]_+ + \frac{n\lambda}{2} c' K c, \quad (4)$$

By introducing a set of non-negative slack variables ξ_i 's, (4) can be formulated into equivalent problem

$$\begin{aligned} \min_{\xi, b, c} \quad & \sum_{i=1}^n w_i \xi_i + \frac{n\lambda}{2} c' K c, \\ \text{subject to} \quad & \tilde{Y}_i \left\{ b + \sum_{j=1}^n c_j K(Z_i, Z_j) \right\} \geq 1 - \xi_i, \quad i=1, \dots, n, \\ & \xi_i \geq 0, \quad i=1, \dots, n. \end{aligned} \quad (5)$$

It is easy to show that (4) and (5) produce the same solution (b, c) for any given λ . Next, we solve (5) via Lagrange multipliers by considering the following primer function

$$L_p = \sum_{i=1}^n w_i \xi_i + \frac{n\lambda}{2} c' K c - \sum_{i=1}^n \alpha_i \left[\tilde{Y}_i \left\{ b + \sum_{j=1}^n c_j K(Z_i, Z_j) \right\} - 1 + \xi_i \right] - \sum_{i=1}^n \gamma_i \xi_i,$$

where $\alpha_i \geq 0$ and $\gamma_i \geq 0$, $i = 1, \dots, n$. By setting the derivatives of L_p equal to zero, we obtain the stationary conditions

$$c_i = \frac{1}{n\lambda} \alpha_i \tilde{Y}_i, \quad \sum_{i=1}^n \alpha_i \tilde{Y}_i = 0, \quad \alpha_i = w_i - \gamma_i, \quad i=1, \dots, n, \quad (6)$$

and the Karush-Kuhn-Tucker (KKT) conditions

$$\alpha_i \left[\tilde{Y}_i \left\{ b + \sum_{j=1}^n c_j K(Z_i, Z_j) \right\} - 1 + \xi_i \right] = 0, \quad \gamma_i \xi_i = 0, \quad i=1, \dots, n. \quad (7)$$

This leads to the dual problem which maximizes a quadratic function with respect to α_i 's

$$\max_{\alpha_1, \dots, \alpha_n} \sum_{i=1}^n \alpha_i - \frac{1}{2n\lambda} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \tilde{Y}_i \tilde{Y}_j K(Z_i, Z_j),$$

subject to the stationary and KKT conditions (6) and (7). The dual problem can be solved by quadratic programming (QP). After obtaining $\hat{\alpha}_i$'s, we compute $\hat{c}_i$'s and $\hat{b}$ as

$$\begin{aligned} \hat{c}_i &= \frac{1}{n\lambda} \hat{\alpha}_i \tilde{Y}_i, \quad i=1, \dots, n; \\ \hat{b} &= \frac{\sum_{i=1}^n \{ \tilde{Y}_i - \sum_{j=1}^n \hat{c}_j K(Z_i, Z_j) \} \hat{\alpha}_i (w_i - \hat{\alpha}_i)}{\sum_{i=1}^n \hat{\alpha}_i (w_i - \hat{\alpha}_i)} \end{aligned}$$

Then $\hat{f}_\pi(z) = \hat{b} + \sum_{i=1}^n \hat{c}_i K(z, Z_i)$. To learn the survival probability, we repeat the above procedure for all of the π_j 's.

3.2. Choice of Tuning Parameter

The tuning parameter λ plays an important role in determining the accuracy of survival probability estimation and future prediction. In order to choose a proper λ , we use the following criterion based on a tuning data set $(\tilde{T}_i^*, \delta_i^*, Z_i^*), i = 1, \dots, n'$, to obtain an optimal value of λ ,

$$\frac{1}{n'} \sum_{i=1}^{n'} \left\{ (1 - \pi) \frac{I(\tilde{T}_i^* > t)}{\hat{G}^*(t|Z_i^*)} + \pi \frac{\delta_i^* I(\tilde{T}_i^* \leq t)}{\hat{G}^*(\tilde{T}_i^{*-}|Z_i^*)} \right\} \left\{ 1 - \tilde{Y}_i^* \hat{f}_\pi(Z_i^*) \right\}_+. \quad (8)$$

Here $\tilde{Y}_i^* = 2I(\tilde{T}_i^* > t) - 1$ and $\hat{G}^*(\cdot|z)$ is an estimate of the censoring time survival function based on the tuning data. A grid of λ values are considered, and the one which minimizes (8) is chosen as the best tuning parameter. When there is not a tuning data set available, we suggest to compute (8) based on cross validation. In the simulation studies, we generate a tuning data set and select the optimal λ for any given π . For real data applications, five-fold cross validation is used. Our numerical results suggest that the tuning based on (8) works effectively.

4. Simulation Studies

4.1. Risk Classification—We first demonstrate performance of the proposed IPCW-wSVM method for risk classification. As commonly done in literature, we use $\pi = 0.95$ as a critical cut-point to distinguish high-risk and low-risk subjects. That is, subjects with t -year survival probability greater than 0.95 are considered as *low-risk* class, while subjects with t -year survival probability smaller than 0.95 are classified as *high-risk* class.

We compare three estimators: IPCW-wSVM, PH, and PO, which are evaluated under the following five scenarios

- Case 1: (standard) PH model
- Case 2: (standard) PO model
- Case 3: a generalized Yang-Prentice (YP) model (Yang and Prentice, 2005)
- Case 4: quadratic PH model
- Case 5: quadratic PO model,

and each case corresponds to a different true model. The generalized YP model is given by

$$\Lambda(t|Z) = \int_0^t \frac{e^{(\beta+\gamma)'Z}}{e^{\beta'Z}F_0(s) + e^{\gamma'Z}S_0(s)} d\Lambda_0(s). \quad (9)$$

where $\Lambda(t|Z)$ is the cumulative hazard function for subjects with covariates Z , F_0 is a cumulative distribution function, and S_0 and Λ_0 are the corresponding survival and cumulative hazard functions. Note that when $\beta = \gamma$, the generalized YP model becomes the PH model; when $\gamma = 0$, it becomes the PO model; when β and γ have opposite signs, it may have crossing hazard functions.

In the cases of PH and PO models, we use the general transformation model $H(T) = -\beta^T Z + \varepsilon$ to generate survival times, where $H(t) = -1 + \log[0.1\{\exp(0.5t) - 1\}]$, $\beta = (1, 1)'$, and ε is independently and identically distributed. For the PH model, ε follows the standard extreme value distribution, i.e. the hazard function of ε is given by $\lambda_\varepsilon(x) = \exp(x)$. For the PO model, ε follows the standard logistic distribution, i.e. $\lambda_\varepsilon(x) = \exp(x)/\{1 + \exp(x)\}$. For the generalized YP model, we set $\Lambda_0(t) = t$ in (9) and let $\beta = (1, 1)'$ and $\gamma = (-1, -1)'$. In Cases 1 to 3, we consider two covariates $Z = (Z_1, Z_2)'$, which are generated from the multivariate normal distribution with mean 0, variance 1 and correlation $\rho = 0.5$. In Cases 4 and 5, we generate the survival times from the PH and PO models with the covariate $Z = (Z_1, Z_2, Z_1 \times Z_2, Z_1^2, Z_2^2)'$, $\beta = (0.5, 0.5, -2, 1, -1)'$, and $\rho = 0.1$.

For each case, we generate the censoring times from a uniform distribution on $[0, \tau]$, where τ is chosen to yield the censoring rate of 15% or 40%. In our method, the Kaplan-Meier estimator is used for estimating the survival function of the censoring times. The sample size of the training data is taken as $n = 200$ or 400. For each scenario, an independent tuning data set of size n is used to choose the best tuning parameter. Another independent test data set of

size 1000 is generated to evaluate the prediction performance. We perform risk classification with $\pi = 0.95$ using the IPCW-wSVM method with the radial basis kernel at three time points, such that the low risk proportion is around 0.65, 0.3, and 0.15 accordingly for Cases 1-3, and around 0.8, 0.35 and 0.25 for Cases 4-5. For comparison, we also report the classification errors given by the standard PH and PO models, where the predicted classification label is determined by $2I(\hat{S}(t|Z) > 0.95) - 1$. Here $\hat{S}(t|Z)$ is the predicted survival probability fitted by PH or PO model.

Tables 1 and 2 present the average misclassification rates calculated over 100 replications for the methods under comparison. The numbers in parentheses are the standard errors. We have the following observations. In Case 1 where PH model is true, the PH estimator is the best (which is expected) and the PO estimator is overall the second best. The IPCW-wSVM estimator can be slightly better, close to, or worse than the PO estimator, depending on the sample size and the reference time points. This is not that surprising, since it is known that the PO model can well approximate the PH model if the PH model is the true model, especially when the censoring proportion is high. When $n = 400$, the IPCW-wSVM estimator is very close to the PO estimator. Actually, when the sample size n increases to 600 (not reported here), we observe that the results begin to favor the IPCW-wSVM estimator compared with the PO estimator. In Case 2 where the PO model is true, the PO estimator is the best. The IPCW-wSVM is the second best and it consistently beats the model-misspecified PH estimator. In Cases 3, 4, 5, where either the true model is YP with crossing hazards or the covariates have nonlinear effects, the IPCW-wSVM estimator is consistently the best among the three; its advantages over the standard PH and PO estimators are substantial. These results suggest the robustness and flexibility of the IPCW-wSVM classifier for complex data, largely due to its model-free nature. Note that the proposed framework does not require prior knowledge or strong assumption on the explicit functional form of survival time. When the sample size n increases, we observe that the misclassification rate of IPCW-wSVM decreases. We also note that performance of IPCW-wSVM stays stable when the censoring rate increases.

4.2. Survival Probability Prediction—We now demonstrate the performance of the IPCW-wSVM in survival probability prediction. For the PH model, we set $\beta = \gamma = (1, 1)'$ and $\Lambda_0(t) = t$ in (9); for the PO model, we set $\beta = (1, 1)'$, $\gamma = (0, 0)'$ and $\Lambda_0(t) = \log(t + 1)$ in (9); for the generalized YP model, we set $\Lambda_0(t) = t$ and choose $\beta = (1, 1)'$ and $\gamma = (-1, -1)'$, which introduces crossing hazard functions. For the PH or PO model with quadratic effects, we use the same settings as in Section 4.1 except that $H(t) = \log(t)$. We consider the sample size $n = 100$ and 200. Other settings are set the same as those in risk classification experiments.

We compute the estimates of the survival function, $\hat{S}(t|z)$, at a few reference time points using the proposed IPCW-wSVM estimator with a radial basis kernel. In our simulations, the set of reference time points T are taken as the $1/6, \dots, 5/6$ quantiles of the survival times in the training data. In order to assess the prediction performance of the estimates, we calculate the mean square root of the aggregated squared error (MRASE) as

$$\frac{1}{1000} \sum_{j=1}^{1000} \left[\sum_{t \in T} \left\{ \hat{S}(t|Z_j^*) - S_0(t|Z_j^*) \right\}^2 \right]^{1/2},$$

where $Z_1^*, \dots, Z_{1000}^*$ are 1000 independent copies of the covariates that are generated following the same distribution as the training data. Under each setting, we report the mean MRASE for each method over 100 simulations. For comparison, we also include the mean MRASE for the estimated survival functions based on the standard PH estimator with all covariates. The results are summarized in Table 3. The numbers in the parentheses are the standard errors of mean MRASE over 100 runs.

From Table 3, we can make several observations. First, when PH model is the true model, the estimates based on the PH model have much smaller mean MRASE than the IPCW-wSVM fits, which is expected. Second, when PO model is the true model, it is a little surprising to observe that the estimates based on the PH model, which is a misspecified model, give smaller mean MRASEs than the IPCW-wSVM. By studying the numerical results more carefully, we note that though the estimated regression coefficients are biased, the estimated survival functions given by the PH are close to the true curves. This implies a certain degree of robustness of the PH estimator in estimating survival probabilities as long as the deviation from the PH model assumption is not severe. Third, when the true model is the YP model with crossing hazards, the IPCW-wSVM shows much better prediction performance than the PH estimator by exhibiting much smaller mean MRASE. This suggests flexibility of the IPCW-wSVM for predicting survival probabilities under complex situations, say, the models with crossing hazards. This finding is also observed from the analysis of the glioma tumor data given in Section 5.1. Furthermore, when the true model is PH or PO with quadratic effects, the IPCW-wSVM outperforms the standard PH estimator which mis-specifies the functional form for covariates.

5. Data Applications

5.1. Survival Probability Prediction for Glioma Tumor Data

We first consider the glioma tumor data (Piantadosi, 1997) application. The data is from a randomized study of 222 patients with gliomas. One major goal is to test the safety and efficacy of the treatment of implanting the polymer with Carmustine (BCNU) at the tumor site after surgery in order to prolong the exposure of drug. Among 222 patients, 110 were assigned to the treatment group receiving BCNU polymers and 112 were assigned to the control group receiving polymers with placebo. Following Liu et al. (2008), we only focus on a subgroup of 148 patients with the tumor type glioblastoma. Among them, 76 patients were in the treatment group and 72 were in the control group. The endpoint of interest is the death time. Except for three patients, all subjects died before the end of the study. Liu et al. (2008) analyzed the data and found that the estimated Kaplan-Meier curves for the two treatment groups crossed at around the 50th week. Here we fit the data using the PH estimator and the proposed IPCW-wSVM estimator with the treatment indicator as a covariate, and use Kaplan-Meier estimates as the gold standard for comparison. The tuning

parameters in the IPCW-wSVM method are chosen using 5-fold cross validation based on (8).

The fitted survival curves given by the IPCW-wSVM (blue color) are shown in the left panel of Figure 1, while those by the PH estimator (red color) are in the right panel of Figure 1. For comparison, the Kaplan-Meier curves (black color) are given in both panels of Figure 1. The survival curves for the treatment group (group 1) are plotted by dashed lines while those for the control group (group 0) are plotted by solid lines. From the Kaplan-Meier curves, the BCNU treatment has beneficial effect for the short-term survival, but it eventually becomes worse than the control group. Also, the fitted survival curves using the IPCW-wSVM are very close to the Kaplan-Meier curves and they can capture the crossing hazard pattern at nearly the same time point as for the Kaplan-Meier curves. However, the fitted survival curves based on the PH estimator show obvious deviation from the Kaplan-Meier curves and are not crossing due to the underlying PH model assumption. In addition, we calculate the sum of squared error (SSE) of the fitted survival curves compared with the Kaplan-Meier curves for the treatment and control groups, which is defined as

$$\sum_{t \in T_0} \left\{ \hat{S}(t|Z=0) - \hat{S}^{KM}(t|Z=0) \right\}^2 + \sum_{t \in T_1} \left\{ \hat{S}(t|Z=1) - \hat{S}^{KM}(t|Z=1) \right\}^2,$$

where T_0 and T_1 are the sets of all the observed failure times for the control and treatment groups, respectively, and $\hat{S}^{KM}(t|Z=0)$ and $\hat{S}^{KM}(t|Z=1)$ are the Kaplan-Meier curves for the control and treatment groups, respectively. The SSE of the IPCW-wSVM method is 0.082, while the SSE of the PH model fit is 0.173. All these results suggest that the IPCW-wSVM method has a better prediction of survival probabilities than the PH fit when there are crossing hazards/survival curves.

5.2. Risk Classification for Breast Cancer Data

Here we consider the analysis of a breast cancer data set from Van Houwelingen et al. (2006). Survival times and gene-expression measurements for 4,919 genes were collected for 295 patients with breast cancer, among which 73.2% of the survival times were censored. We randomly divide 295 subjects into a training set of size 200 and a testing set of size 95. We pre-select the top 70 most significant genes based on p-values of the Wald test of the marginal PH model with a single gene. We compare three methods: the IPCW-wSVM estimator using radial basis kernel, the standard PH estimator, and the standard PO estimator. Five-fold cross validation is used to determine the tuning parameters.

To evaluate the accuracy of risk classification, we use the receiver operating characteristic (ROC) curve and the area under the curve (AUC) for censored survival data as studied in Heagerty et al. (2000). Here the predicted survival probability is considered as the risk marker. An ROC curve demonstrates the true positive (TP) rate (or sensitivity) and the false positive (FP) rate (or 1-specificity) of a test at all the critique values. For survival data, an ROC curve is determined by $\{(TP_t(c), FP_t(c)) : c \in [0, 1]\}$, where

$$TP_t(c) = P \left\{ \hat{S}(t|Z) < c | T \leq t \right\},$$

$$FP_t(c) = P \left\{ \hat{S}(t|Z) < c | T > t \right\}.$$

Here t is the reference time of interest, which is set to be 40% and 80% percentiles of all the death times for illustration, i.e. 3.2 years and 6.5 years for the breast cancer data. To construct the ROC curves, the $TP_t(c)$ and $FP_t(c)$ are estimated based on the testing data following the method of Heagerty et al. (2000). The 3.2-year and 6.5-year ROC curves for the considered three estimation methods are shown in Figure 2. The corresponding AUCs are also provided under the plots. Based on the ROC curves and the AUCs, the IPCW-wSVM shows much better risk classification performance than the standard PH and PO fits.

6. Discussions

In this paper, we proposed a machine learning framework for risk classification and survival probability prediction in censored data analysis. The proposed IPCW-wSVM method does not rely on any model assumption or the specification of the function form of the covariate effects, so it provides a flexible alternative to popular survival models. In this paper, the focus is on the point estimate of survival probability. The estimation of its standard error is a challenging topic and needs further investigation.

Acknowledgement

W. Lu's work was partially supported by NIH/NCI grants R01 CA140632 and P01 CA142538. H. H. Zhang's work was partially supported by NSF grants DMS-1418172, DMS-1309507 and DBI-1261830.

References

- Aalen OO. A model for nonparametric regression analysis of counting processes. *Lecture Notes in Statistics*. 1980; 2:1–25.
- Bair E, Paul D, Tibshirani R. Prediction by supervised principal components. *Journal of the American Statistical Association*. 2006; 101:119–137.
- Bennett S. Analysis of survival data by the proportional odds model. *Statistics in Medicine*. 1983; 2:273–277. [PubMed: 6648142]
- Beran, R. Technical report. University of California; Berkeley, CA: 1981. Nonparametric regression with randomly censored survival data.
- Cheng SC, Wei LJ, Ying Z. Analysis of transformation models with censored data. *Biometrika*. 1995; 82:835–845.
- Cheng SC, Wei LJ, Ying Z. Predicting survival probabilities with semiparametric transformation models. *Journal of the American Statistical Association*. 1997; 92:227–235.
- Clayton D, Cuzick J. Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society*. 1985; 148:82–117. Series A (General).
- Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society*. 1972; 34:187–220. Series B (Methodological).
- Cox, DR.; Oakes, D. *Analysis of Survival Data*. Chapman & Hall; London: 1984.
- Dabrowska D. Non-parametric regression with censored survival time data. *Scandinavian Journal of Statistics*. 1987; 14:181–197.
- Fan J, Gijbels I. Censored regression: Local linear approximations and their applications. *Journal of the American Statistical Association*. 1994; 89:560–570.

- Fine JP, Ying Z, Wei LG. On the linear transformation model for censored data. *Biometrika*. 1998; 85:980–986.
- Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*. 2000; 56:337–344. [PubMed: 10877287]
- Ishwaran H, Kogalur UB, Glackstone EH, Lauer MS. Random survival forests. *The Annals of Applied Statistics*. 2008; 2:841–860.
- Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*. 1958; 53:457–481.
- Kimeldorf G, Wahba G. Some results on Tchebychean spline functions. *Journal of Mathematical Analysis and Application*. 1971; 33:82–95.
- Kong L, Cai J, Sen PK. Weighted estimating equations for semiparametric transformation models with censored data from a case-cohort design. *Biometrika*. 2004; 91:305–319.
- Koul H, Susarla V, Ryzin JV. Regression analysis with randomly right-censored data. *The Annals of Statistics*. 1981; 9:1276–1288.
- Li H, Gui J. Partial cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics*. 2004; 20:208–215.
- Li H, Luan Y. Kernel Cox regression models for linking gene expression profiles to censored survival data. *Pac Symp Biocomput*. 2003; 8:65–76. [PubMed: 12603018]
- Lin DY, Wei LJ, Ying ZL. Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*. 1993; 80:557–572.
- Lin Y. Support vector machines and the Bayes rule in classification. *Data Mining and Knowledge Discovery*. 2002; 6:259–275.
- Liu M, Lu W, Shao Y. A Monte Carlo approach for change-point detection in the Cox proportional hazards model. *Statistics in Medicine*. 2008; 27:3894–909. [PubMed: 18254131]
- Lu W, Li L. Sufficient dimension reduction for censored regressions. *Biometrics*. 2011; 67:513–523. [PubMed: 20880013]
- Piantadosi, S. *Clinical Trials: A Methodologic Perspective*. Wiley; New York: 1997.
- Van Houwelingen HC, Bruinsma T, Hart AA, Van't Veer LJ, Wessels LF. Cross-validated Cox regression on microarray gene expression data. *Statistics in Medicine*. 2006; 25:3201–3216. [PubMed: 16143967]
- Wahba G. Soft and hard classification by reproducing kernel Hilbert space methods. *Proceedings of the National Academy of Sciences*. 2002; 99:16524–16530.
- Wang J, Shen X, Liu Y. Probability estimation for large-margin classifiers. *Biometrika*. 2008; 95:149–167.
- Wang Z, Wang CY. Buckley-James boosting for survival analysis with high-dimensional biomarker data. *Statistical Applications in Genetics and Molecular Biology*. 2010; 9 Article 24.
- Yang S, Prentice R. Semiparametric analysis of short-term and long-term hazard ratios with two-sample survival data. *Biometrika*. 2005; 92:1–17.

Figure 1.

The estimated survival curves for the two treatment groups using the IPCW-wSVM method (left panel) and the PH model (right panel). The SSE between the estimated survival curves and the Kaplan-Meier curves is also provided.

Figure 2.
The ROC curves and AUCs of the three methods for analysis of the breast cancer data.

Table 1Mean and standard errors (in parentheses) of misclassification rates with $n = 200$.

True Model	Censoring Rate	Reference Time	Low Risk Proportion	PH	PO	IPCW-wSVM
PH	0.15	t_1	0.683	0.043 (0.003)	0.067 (0.004)	0.111 (0.007)
		t_2	0.304	0.035 (0.002)	0.104 (0.004)	0.101 (0.006)
		t_3	0.115	0.029 (0.002)	0.095 (0.004)	0.073 (0.004)
	0.40	t_1	0.683	0.048 (0.003)	0.063 (0.004)	0.112 (0.008)
		t_2	0.304	0.042 (0.003)	0.092 (0.005)	0.105 (0.006)
		t_3	0.115	0.034 (0.003)	0.078 (0.005)	0.086 (0.004)
PO	0.15	t_1	0.688	0.136 (0.009)	0.059 (0.003)	0.102 (0.005)
		t_2	0.308	0.185 (0.006)	0.045 (0.003)	0.091 (0.007)
		t_3	0.119	0.095 (0.002)	0.039 (0.003)	0.073 (0.003)
	0.40	t_1	0.688	0.113 (0.007)	0.061 (0.004)	0.103 (0.005)
		t_2	0.308	0.136 (0.007)	0.048 (0.004)	0.094 (0.007)
		t_3	0.119	0.077 (0.003)	0.044 (0.003)	0.078 (0.003)
YP	0.15	t_1	0.647	0.529 (0.018)	0.329 (0.015)	0.183 (0.009)
		t_2	0.348	0.358 (0.004)	0.337 (0.005)	0.138 (0.011)
		t_3	0.140	0.140 (0.000)	0.140 (0.000)	0.081 (0.003)
	0.40	t_1	0.647	0.409 (0.015)	0.256 (0.012)	0.177 (0.008)
		t_2	0.348	0.348 (0.001)	0.307 (0.005)	0.123 (0.010)
		t_3	0.140	0.140 (0.000)	0.140 (0.000)	0.081 (0.003)
Quadratic PH	0.15	t_1	0.809	0.477 (0.025)	0.415 (0.016)	0.135 (0.008)
		t_2	0.358	0.358 (0.000)	0.332 (0.004)	0.106 (0.005)
		t_3	0.234	0.234 (0.000)	0.233 (0.000)	0.098 (0.006)
	0.40	t_1	0.809	0.459 (0.024)	0.412 (0.018)	0.136 (0.008)
		t_2	0.358	0.358 (0.001)	0.341 (0.003)	0.122 (0.008)
		t_3	0.234	0.234 (0.000)	0.234 (0.000)	0.110 (0.006)
Quadratic PO	0.15	t_1	0.820	0.387 (0.023)	0.364 (0.014)	0.104 (0.007)
		t_2	0.361	0.361 (0.000)	0.336 (0.004)	0.122 (0.007)
		t_3	0.236	0.236 (0.000)	0.235 (0.001)	0.088 (0.004)
	0.40	t_1	0.820	0.382 (0.022)	0.363 (0.016)	0.104 (0.007)
		t_2	0.361	0.361 (0.001)	0.348 (0.003)	0.133 (0.008)
		t_3	0.236	0.236 (0.000)	0.236 (0.000)	0.100 (0.005)

Table 2Mean and standard error (in parentheses) of misclassification rates with $n = 400$.

True Model	Censoring Rate	Reference Time	Low Risk Proportion	PH	PO	IPCW-wSVM
PH	0.15	t_1	0.673	0.038 (0.003)	0.075 (0.003)	0.080 (0.004)
		t_2	0.289	0.024 (0.002)	0.116 (0.004)	0.081 (0.004)
		t_3	0.126	0.018 (0.001)	0.074 (0.002)	0.073 (0.004)
	0.40	t_1	0.673	0.039 (0.003)	0.070 (0.003)	0.081 (0.004)
		t_2	0.289	0.028 (0.002)	0.098 (0.004)	0.083 (0.004)
		t_3	0.126	0.020 (0.002)	0.061 (0.002)	0.055 (0.003)
PO	0.15	t_1	0.681	0.143 (0.008)	0.040 (0.003)	0.070 (0.004)
		t_2	0.293	0.175 (0.004)	0.033 (0.002)	0.070 (0.004)
		t_3	0.128	0.103 (0.002)	0.024 (0.002)	0.056 (0.004)
	0.40	t_1	0.681	0.112 (0.007)	0.041 (0.003)	0.072 (0.004)
		t_2	0.293	0.134 (0.004)	0.037 (0.003)	0.074 (0.004)
		t_3	0.128	0.083 (0.002)	0.027 (0.002)	0.060 (0.004)
YP	0.15	t_1	0.641	0.576 (0.018)	0.282 (0.013)	0.115 (0.006)
		t_2	0.335	0.335 (0.000)	0.325 (0.002)	0.081 (0.004)
		t_3	0.149	0.149 (0.000)	0.149 (0.000)	0.064 (0.004)
	0.40	t_1	0.641	0.404 (0.013)	0.205 (0.011)	0.112 (0.006)
		t_2	0.335	0.335 (0.000)	0.294 (0.004)	0.081 (0.004)
		t_3	0.149	0.149 (0.000)	0.149 (0.000)	0.066 (0.004)
Quadratic PH	0.15	t_1	0.808	0.476 (0.022)	0.415 (0.012)	0.082 (0.004)
		t_2	0.346	0.346 (0.000)	0.326 (0.002)	0.071 (0.003)
		t_3	0.233	0.233 (0.000)	0.233 (0.000)	0.060 (0.002)
	0.40	t_1	0.808	0.462 (0.021)	0.416 (0.014)	0.079 (0.004)
		t_2	0.346	0.346 (0.000)	0.334 (0.002)	0.073 (0.003)
		t_3	0.233	0.233 (0.000)	0.233 (0.000)	0.062 (0.002)
Quadratic PO	0.15	t_1	0.810	0.417 (0.022)	0.380 (0.013)	0.085 (0.004)
		t_2	0.349	0.349 (0.000)	0.329 (0.002)	0.074 (0.003)
		t_3	0.237	0.237 (0.000)	0.237 (0.000)	0.069 (0.003)
	0.40	t_1	0.810	0.408 (0.021)	0.378 (0.015)	0.089 (0.005)
		t_2	0.349	0.349 (0.000)	0.337 (0.002)	0.076 (0.003)
		t_3	0.237	0.237 (0.000)	0.237 (0.000)	0.095 (0.006)

Table 3

Mean and standard error (in parentheses) of MRASE for the PH and IPCW-wSVM estimates.

True Model	n	Censoring Rate	PH	IPCW-wSVM
PH	100	15%	0.0917(0.0031)	0.2067(0.0032)
		40%	0.1024(0.0028)	0.2338(0.0035)
	200	15%	0.0655(0.0021)	0.1518(0.0015)
		40%	0.0727(0.0025)	0.1579(0.0019)
PO	100	15%	0.1424(0.0043)	0.2018(0.0031)
		40%	0.1387(0.0039)	0.2460(0.0032)
	200	15%	0.1146(0.0030)	0.1676(0.0018)
		40%	0.1098(0.0028)	0.1595(0.0021)
YP	100	15%	0.3948(0.0025)	0.2429(0.0023)
		40%	0.3601(0.0027)	0.2662(0.0024)
	200	15%	0.3957(0.0020)	0.1792(0.0018)
		40%	0.3576(0.0019)	0.1995(0.0021)
Quadratic PH	100	15%	0.6221(0.0017)	0.2513(0.0027)
		40%	0.6206(0.0018)	0.2793(0.0032)
	200	15%	0.6203(0.0012)	0.2008(0.0018)
		40%	0.6207(0.0014)	0.2225(0.0020)
Quadratic PO	100	15%	0.5307(0.0019)	0.2661(0.0029)
		40%	0.5396(0.0017)	0.2973(0.0031)
	200	15%	0.5305(0.0014)	0.1953(0.0018)
		40%	0.5355(0.0013)	0.2153(0.0023)