SVM-Based Approaches for Predictive Modeling of Survival Data

Han-Tai Shiao and Vladimir Cherkassky
Department of Electrical and Computer Engineering
University of Minnesota, Twin Cities
Minneapolis, Minnesota 55455, U.S.A.
Email: {shiao003, cherk001}@umn.edu

***Abstract*—Survival data is common in medical applications. The challenge in applying predictive data-analytic methods to survival data is in the treatment of censored observations. The survival times for these observations are unknown. This paper presents formalization of the analysis of survival data as a binary classification problem. For this binary classification setting, we propose two different strategies for encoding censored data, leading to two advanced SVM-based formulations: SVM+ and SVM with uncertain class labels. Further, we present empirical comparison of the advanced SVM methods and the classical Cox modeling approach for predictive modeling of survival data. These comparisons suggest that the proposed SVM-based models consistently yield better predictive performance (than classical statistical modeling) for real-life survival data sets.**

***Index Terms*—classification, survival analysis, Support Vector Machine (SVM), SVM+, Learning Using Privileged Information (LUPI), SVM with uncertain labels, Cox model.**

I. Introduction

A significant proportion of medical data is a collection of time-to-event observations. Methods for survival analysis developed in classical statistics have been used to model such data. Survival analysis focuses on the time elapsed from an initiating event to an event, or endpoint, of interest [1]. Classical examples are the time from birth to death, from disease onset to death, and from entry to a study to relapse, *etc*. All these times are generally known as the *survival time*, even when the endpoint is something different from death. This statistical methodology can also be used in many different settings, such as the reliability engineering, and financial insurance. Even though the purpose of a statistical analysis may vary from one situation to another, the ambitious aim of most statistical analyses is to build a model that relates explanatory variables and the occurrences of the event.

The field of machine learning is also targeting the same or similar goals. Learning is the process of estimating an unknown dependency between system's inputs and its output, based on a limited number of observations [2]. However, the machine learning techniques have not been widely used for survival analysis for two major reasons.

First, the survival time is not necessarily observed in all samples. For example, patients might not experience the occurrence of event (death or relapse) during the study, or they were lost to follow-up. Hence, the survival time is incomplete and only known "up-to-a-point," which is quite different from the traditional notion of 'missing data.'

The second reason is methodological. Machine learning techniques are usually developed and applied under predictive setting, where the main goal is the prediction accuracy for future (or test) samples. In contrast, classical statistical methods aim at estimating the true probabilistic model of available data. So the prediction accuracy is just one of several performance indices. The methodological assumption is that if an estimated model is 'correct,' then it should yield good predictions. So the classical statistical methodology often does not clearly differentiate between training (model estimation) and prediction (or test) stages. This paper assumes a predictive setting, which is appropriate for many applications. Under this predictive setting, the survival time is known for training data, but it is not available during the prediction (or testing) stage. Thus, modifications are required for applying existing machine learning approaches to survival data analysis.

Previously, several studies applied Support Vector Machines (SVM) to survival data [3]–[5]. Most of these efforts formalize the problem under the regression setting. Specifically, the SVM regression was used to estimate a model that predicts the survival time. However, formalization using regression setting is intrinsically more difficult than classification. Further, practitioners generally use the modeling outputs as a reference and they are usually concerned with the status of a patient at a given time, such as six-month after surgery or two-year post transplant.

In this paper, we propose to use a special classification formulation that addresses the issues of incomplete information in the survival time. Instead of predicting the survival time, we try to estimate a model that predicts a subject's status at a time point of interest. This paper is organized as follows. The characteristics of the survival data are summarized in Section II. The predictive problem setting for survival analysis is introduced in Section III. The proposed SVM-based formulations are introduced in Section IV. Empirical comparisons for several synthetic and real-life data sets are presented in Section V and VI. Finally, the discussion and conclusion are given in Section VII.

II. Survival Data Analysis

This section provides general background description of survival data analysis and its terminology.

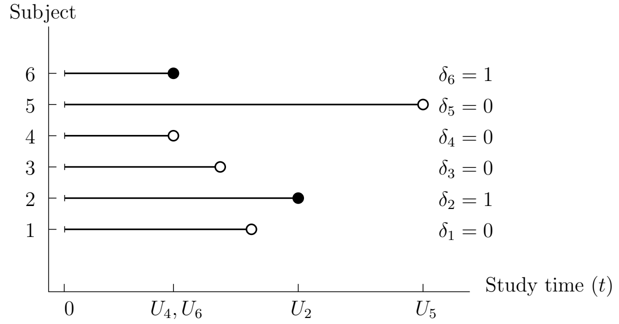

Fig. 1. Example of survival data in a study-time scale. The exact observations are indicated by solid dots, and the censored observations by hollow dots.

The survival data (or failure time data) are obtained by observing individuals from a certain initial time to either the occurrence of a predefined event or the end of the study. The predefined event is often the failure of a subject or the relapse of a disease. The major difference between survival data and other types of numerical data is the time to the event occurring is not necessarily observed in all individuals.

A common feature of these data sets is they contain censored observations. Censored data arise when an individual's life length is known to occur only in a certain period of time. Possible censoring schemes are *right censoring*, where all that is known is that the individual is still alive at a given time, *left censoring* when all that is known is that the individual has experienced the event of interest prior to the start of the study, or *interval censoring*, where the only information is that the event occurs within some interval. In this paper, we only consider the right censoring scheme.

The graphical representation of the survival data for a hypothetical study with six subjects is shown in Figure 1. In this study, subject 2 and 6 experienced the event of interest prior to the end of the study and they are the exact observations. Subject 1, 3, and 5, who experienced the event after the end of the study, are only known to be alive at the end of the study. Subject 4 was included in the study for some time but further observation cannot be obtained. The data for subject 1, 3, 4, and 5 are called censored (right-censored) observations. Thus, for the censored observations, it is known that the survival time is greater than a certain value, but it is not known by how much.

Suppose T denotes the event time, such as death or lifetime; C denotes the censoring time, *e.g.*, the end of study or the time an individual withdraws from the study. The T's are assumed to be independent and identically distributed with probability density function $\varphi(t)$ and survival function $S(t)$. For right censoring scheme, we only know $T_i > C_i$ with observed C_i. Then the survival data can be represented by pairs of random variables (U_i, δ_i), $i = 1, \ldots, n$. The δ_i indicates whether the observed survival time U_i corresponds to an event ($\delta_i = 1$) or is censored ($\delta_i = 0$). The U_i is equal to T_i if the lifetime or event is observed, and to C_i if it is censored. Mathematically, U_i and δ_i are defined as

$$U_i = \min(T_i, C_i), \tag{1}$$

$$\delta_i = I(T_i \le C_i) = \begin{cases} 0 & \text{censored observation,} \\ 1 & \text{event occurred.} \end{cases} \tag{2}$$

In Figure 1, subject 4 and 6 have the same observed survival time ($U_4 = U_6$), but their censoring indicators are different ($\delta_4 = 0, \delta_6 = 1$). Therefore, in the survival analysis, we are given a set of data, $(\mathbf{x}_i, U_i, \delta_i)$, $i = 1, \ldots, n$, where $\mathbf{x}_i \in \mathbf{R}^d$, $U_i \in \mathbf{R}_+$ and $\delta_i \in \{0, 1\}$. In contrast, under supervised learning setting, we are given a set of training data, $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, where $\mathbf{x}_i \in \mathbf{R}^d$ and $y_i \in \mathbf{R}$. The target values y_i's can be real-valued such as in standard regression, or binary class labels in classification.

Classical statistical approach to modeling survival data aims at estimating the survival function $S(t)$, which is the probability that the time of death is greater than certain time t. More generally, the goal is to estimate $S(t|\mathbf{x})$, or survival function conditioned on patient's characteristics, denoted as feature vector $\mathbf{x}$. Assuming that the probabilistic model $S(t|\mathbf{x})$ is known, or can be accurately estimated from available data, this model provides complete statistical characterization of the data. In particular, it can be used for prediction and for explanation (*i.e.*, identifying input features that are strongly associated with an outcome, such as death).

III. Predictive Modeling of Survival Data

In many applications, the goal is to estimate (predict) survival at a pre-specified time point τ, *e.g.*, survival of cancer patients two years after initial diagnosis, or the survival status of patients one year after bone marrow transplant procedure. Generally τ can be about half of the maximum observed survival time. Next we describe possible formalization of this problem under predictive setting, leading to a binary classification formulation.

Classification problem setting: Given the training survival data, $(\mathbf{x}_i, U_i, \delta_i, y_i)$, $i = 1, \ldots, n$, where $\mathbf{x}_i \in \mathbf{R}^d$, $U_i \in \mathbf{R}_+$, $\delta_i \in \{0, 1\}$, and $y_i \in \{-1, +1\}$, estimate a classification model $f(\mathbf{x})$ that predicts a subject's status at a pre-specified time τ based on the input (or covariates) $\mathbf{x}$.

The status of subject i at time τ is a binary class label through the following encoding

$$y_i = \begin{cases} +1, & \text{if } U_i < \tau, \\ -1, & \text{if } U_i \ge \tau. \end{cases} \tag{3}$$

Note that U_i and δ_i are only available for training, not for prediction (or testing stage). So the challenge of predictive modeling is to develop novel classification formulations that incorporate uncertain nature of censored data.

In a hypothetical study as shown in Figure 2, suppose a subject's status is given by (3), then there is no ambiguity in the statuses of subject 2 and 6. Likewise, the survival status of subject 5 is known, even though the observation is censored. However, the survival statuses for subjects 1, 3, and 4 are unknown since the observed survival times are shorter than τ.

There are two simplistic ways to incorporate censored data into standard classification formulation:

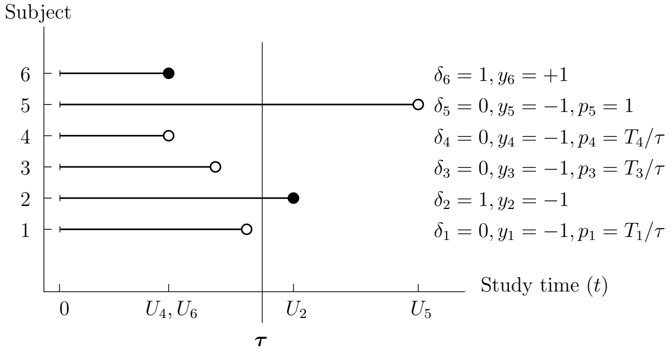
Subject
6
5
4
3
2
1
$\delta_6 = 1, y_6 = +1$
$\delta_5 = 0, y_5 = -1, p_5 = 1$
$\delta_4 = 0, y_4 = -1, p_4 = T_4/\tau$
$\delta_3 = 0, y_3 = -1, p_3 = T_3/\tau$
$\delta_2 = 1, y_2 = -1$
$\delta_1 = 0, y_1 = -1, p_1 = T_1/\tau$
Study time (t)
0
U_4, U_6
U_2
U_5
τ

Fig. 2. Example of survival data under the predictive problem setting. The goal is to find a model that predicts the subjects' statuses at time τ.

- Treat the censoring time as the actual event time, *i.e.*, replace T_i with C_i. This approach underestimates the actual event time because $T_i > C_i$.
- Simply ignore the censored data and estimate a binary classifier using only exact observations. This approach yields suboptimal models, as we ignore the information available in the censored data.

This paper investigates two different strategies for incorporating censored data in SVM-based classifiers:

1) Note that censoring information is available/known for training data, but not known during prediction, the censored data can be regarded as the privileged information under the so-called Learning Using Privileged Information (LUPI) paradigm [6], [7].
2) We can assign probabilities to reflect the uncertain status of censored data samples. One simple rule is to set the probability of a subject being alive at time τ proportional to the (known) survival time, as indicated in Figure 2. That is, $\Pr(y_i = -1|\mathbf{x}_i) = U_i/\tau$ or $\Pr(y_i = +1|\mathbf{x}_i) = 1 - U_i/\tau$. The idea is that if U_i is small, it is more likely subject i will not survive at time τ. On the other hand, if U_i is very close to τ, subject i will be alive at time τ with high probability. Therefore, the survival data $(\mathbf{x}_i, U_i, \delta_i)$, $i = 1, \ldots, n$, can be translated into $(\mathbf{x}_i, U_i, l_i)$, $i = 1, \ldots, n$. For exact observations, $l_i = y_i \in \{-1, +1\}$, $i = 1, \ldots, m$. For censored observations, $l_i = p_i \in [0, 1]$, $i = m+1, \ldots, n$, where
$$p_i = \Pr(y_i = -1|\, \mathbf{x}_i) = U_i/\tau \tag{4}$$
considers the uncertainty about the class membership of $\mathbf{x}_i$. The concept of assigning probability to the uncertain status can be extended to the exact observations. For a exact observation, we have its status y_i with probability $p_i = 1$. Then the survival data are represented as $(\mathbf{x}_i, U_i, p_i, y_i)$, $i = 1, \ldots, n$. This formalization of censored data leads to the so-called SVM with uncertain labels modeling approach [8].

Both modeling approaches are presented later in Section IV.

Finally, we describe application of classical survival analysis under predictive setting (introduced earlier in this section). Classical survival analysis models describe the occurrence of the event by means of survival curves and hazard rates and analyze the dependence (of this event) on covariates by means of regression models [1]. One of the most popular survival-curve estimation is the Cox modeling approach based on the proportional hazards model. Once a survival function $S(t|\mathbf{x})$ is known or estimated (from training data) it can be used for prediction. Specifically, for new (test) input $\mathbf{x}$ the prediction is obtained by a simple thresholding rule
$$y_i = \begin{cases} +1, & \text{if } S(t|\mathbf{x}_i) < r, \\ -1, & \text{if } S(t|\mathbf{x}_i) \geq r, \end{cases} \tag{5}$$
where the threshold value r should reflect the misclassification costs given *a priori*. In this paper, we assume equal misclassification costs. Hence, the threshold level is set to $r = 0.5$. This approach will be used to estimate the prediction accuracy (test error) of the Cox model in empirical comparisons presented in Sections V and VI.

IV. SVM-Based Formulations for Survival Analysis

This section presents two recent advanced SVM-based formulations appropriate for predictive modeling of survival data. Presentation starts with a general description of these SVM-based formulations, followed by specific description of incorporating censored data into these formulations.

A. SVM+

One strategy to handle the survival data is the setting known as Learning Using Privileged Information (LUPI) developed by Vapnik [6], [7]. In a data-rich world, there often exists additional information about training samples, which is not reflected in the training data. This additional information can be easily ignored by standard inductive methods such as SVM. Effective use of this additional information during training often results in improved generalization [7].

Under the LUPI setting, we are given a set of triplets $(\mathbf{x}_i, \mathbf{x}_i^*, y_i)$, $i = 1, \ldots, n$, where $\mathbf{x}_i \in \mathbf{R}^d$, $\mathbf{x}_i^* \in \mathbf{R}^k$, and $y_i \in \{-1, +1\}$. The $(\mathbf{x}, y)$ is the 'usual' labeled training data and $(\mathbf{x}^*)$ denotes the additional *privileged* information available only for training data. Note that the privileged information is defined in a different feature space. This SVM+ approach maps inputs, $\mathbf{x}_i$ and $\mathbf{x}_i^*$, into two different spaces:

- *decision* space $\mathcal{Z}$ via the mapping $\Phi(\mathbf{x}) : \mathbf{x} \mapsto \mathbf{z}$, which is the same feature space used in standard SVM;
- *correcting* space $\mathcal{Z}^*$ via the mapping $\Phi^*(\mathbf{x}) : \mathbf{x} \mapsto \mathbf{z}^*$, which reflects the privileged information about the training data.

The goal of the SVM+ is to estimate a decision function $(\mathbf{w} \cdot \mathbf{z}) + b$ by using the correcting function $\xi(\mathbf{z}^*) = (\mathbf{w}^* \cdot \mathbf{z}^*) + d \geq 0$ as the additional constraints on the training errors (or slack variables) in the decision space. The SVM+ classifier is estimated from the training data by solving the following optimization problem:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{\gamma}{2}\|\mathbf{w}^*\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{subject to} & \boldsymbol{\xi} \succeq 0 \\ & y_i((\mathbf{w}\cdot\mathbf{z}_i)+b) \geq 1-\xi_i, \quad i=1,\ldots,n \\ & \xi_i = (\mathbf{w}^*\cdot\mathbf{z}_i^*)+d, \quad i=1,\ldots,n \end{array} \tag{6}$$

with $\mathbf{w} \in \mathbf{R}^d$, $b \in \mathbf{R}$, $\mathbf{w}^* \in \mathbf{R}^k$, $d \in \mathbf{R}$, and $\boldsymbol{\xi} \in \mathbf{R}_+^n$as the variables. The symbol $\succeq$ denotes componentwise inequality and $\mathbf{R}_+$ denotes non-negative real numbers.

Predictive modeling of survival data can be formalized under SVM+/LUPI formulation (6) as explained next. Available survival data $(\mathbf{x}_i, U_i, p_i, y_i)$ can be represented as $(\mathbf{x}_i, \mathbf{x}_i^*, y_i)$, where $\mathbf{x}_i^* = (U_i, p_i)$ is the privileged information. Then the problem of survival analysis can be formalized and modeled using the SVM+/LUPI paradigm.

B. SVM with Uncertain Labels

This section describes novel SVM-based formulation [8] that introduces the notion of uncertain class labels. That is, some instances (training samples) are not associated with definite class labels. For such uncertain labels, only the confidence levels (or probabilities) regarding the class memberships are provided. In the context of survival analysis, exact observations have known class labels, and censored observations have uncertain class labels.

For the non-separable survival data, we have the following optimization problem,

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i + \tilde{C}\sum_{i=m+1}^{n}(\xi_i^- + \xi_i^+) \\ \text{subject to} & \boldsymbol{\xi} \succeq 0 \\ & y_i((\mathbf{w}\cdot\mathbf{x}_i)+b) \geq 1-\xi_i, \quad i=1,\ldots,m \\ & \boldsymbol{\xi}^- \succeq 0 \\ & \boldsymbol{\xi}^+ \succeq 0 \\ & q_i^- - \xi_i^- \leq (\mathbf{w}\cdot\mathbf{x}_i)+b \leq q_i^+ + \xi_i^+, \\ & i=m+1,\ldots,n. \end{array} \tag{7}$$

with $\mathbf{w} \in \mathbf{R}^d$, $b \in \mathbf{R}$, $\boldsymbol{\xi} \in \mathbf{R}_+^m$, $\boldsymbol{\xi}^- \in \mathbf{R}_+^{n-m}$, and $\boldsymbol{\xi}^+ \in \mathbf{R}_+^{n-m}$ as the variables. The first part of the constraints is for the exact observations. As for the censored observations, their decision values, $(\mathbf{w}\cdot\mathbf{x}_i)+b$, are bounded by q_i^- and q_i^+. The boundaries are functions of p_i, a, and η, *i.e.*,

$$q_i^- = -\frac{1}{a}\log\left(\frac{1}{p_i-\eta}-1\right), \quad q_i^+ = -\frac{1}{a}\log\left(\frac{1}{p_i+\eta}-1\right),$$

where $a = \log(1/\eta - 1)$ is a constant and η is the max deviation of the probability estimate from p_i [8], [9].

The p_i values defined in (4) encode the information about survival time for both censored and exact observations, available in the training data. This formulation can be extended to nonlinear (kernel) parameterization using standard SVM methodology. This method is known (and will be referred to) as pSVM in this paper.

V. Empirical Comparisons for Synthetic Data

This section describes the empirical comparisons between the pSVM, SVM+/LUPI method and the Cox modeling approach [1]. Practical application of these methods to finite data, involves additional simplifications, as discussed next:

- For SVM+, the non-linearity is modeled only in the correcting space [10]. That is, in all experiments the decision space uses linear parameterization, and the correcting space is implemented via non-linear (RBF) kernels.
- pSVM uses either linear or non-linear mapping in the experiments.

Consequently, pSVM with RBF kernel has three tuning parameters, C, $\tilde{C}$, and σ (RBF width parameter), whereas SVM+ with RBF kernel has three tuning parameters, C, γ, and σ. Furthermore, pSVM with linear kernel has two tuning parameters (C and $\tilde{C}$). In contrast, there is no tunable parameter in the Cox modeling approach.

Empirical comparisons are designed to understand relative advantages and limitations of SVM-based methods for modeling the survival data sets with various statistical characteristics, such as the number of training samples, the noise in the observed survival times, and the proportion of censoring. The synthetic data set is generated as follows [11]:

- Set the number of input features d to 30.
- Generate $\mathbf{x} \in \mathbf{R}^d$ with each element x_i being a random number uniformly distributed within $[-1, 1]$.
- Define the coefficient vector as

$$\begin{aligned}\boldsymbol{\beta} = [&1,1,2,3,3,1,1,1,1,0,2,0,2,2,0,\\ &2,0,0,0,0,0,0,0,0,0,0,0,0,0,0].\end{aligned}$$

- Generate the event time T following $\mathrm{Exp}((\boldsymbol{\beta}\cdot\mathbf{x})+2)$ distribution. The Gaussian noise $\nu \sim \mathcal{N}(0, 0.2)$ is also added to the event time T. Generate the censoring time C following $\mathrm{Exp}(\lambda)$ distribution.
- The survival time and event indicator are obtained according to (1) and (2). The rate of the exponential distribution, λ, is used to control the proportion of censoring in the training set.
- Assign class label to each data vector by the rule in (3). The time of interest, τ, is set to the median value among the survival times. In this way, the prior probability for each class is about the same.
- Generate 400 samples for training, 400 for validation, and 2000 for testing.

This data set conforms to probabilistic assumptions (*i.e.*, exponential distribution) underlying the classical modeling approach. So the Cox modeling approach is expected to be very competitive for the synthetic data set.

The following experimental procedure was used in all experiments:

- Estimate the classifier using the training data.
- Find optimal tuning parameters for each method using the validation data. For the Cox modeling approach, the validation data are not used.

TABLE I
The test errors (%) for the synthetic data with 400 training samples.

Trial	1	2	3	4	5	6	7	8	9	10
Cox	26.6	27.1	26.3	29.6	27.4	27.1	28.3	28.7	27.4	26.9
pSVM linear	25.7	**22.6**	**25.0**	**27.5**	**24.2**	26.5	26.1	**26.0**	25.6	**26.1**
pSVM rbf	**24.6**	25.7	25.8	27.9	25.7	**25.4**	25.7	26.9	26.2	26.8
LUPI	25.2	25.5	25.6	29.6	25.7	25.5	**25.6**	27.2	**25.0**	26.5

TABLE II
The test errors (%) for the synthetic data with 250 training samples.

Trial	1	2	3	4	5	6	7	8	9	10
Cox	30.1	29.6	28.0	**27.6**	30.1	30.3	28.9	30.1	29.3	28.3
pSVM linear	**28.6**	**25.8**	**27.6**	28.1	29.8	**26.8**	**28.0**	28.1	27.3	29.0
pSVM rbf	28.9	26.9	30.4	**27.6**	30.5	28.1	27.5	**26.8**	27.7	28.1
LUPI	30.0	28.0	29.3	29.8	29.9	27.6	30.6	30.0	**25.0**	26.3

TABLE III
The test errors (%) for the synthetic data with 100 training samples.

Trial	1	2	3	4	5	6	7	8	9	10
Cox	35.6	**31.3**	34.0	32.3	27.7	**30.6**	**30.6**	33.5	31.4	**28.4**
pSVM linear	**32.5**	33.0	33.5	30.0	**25.1**	33.5	36.9	**30.4**	31.4	30.8
pSVM rbf	**32.5**	32.0	33.8	**29.3**	32.2	32.2	34.2	31.4	33.1	29.9
LUPI	33.6	37.1	**32.0**	32.0	26.0	41.0	33.6	37.0	**30.9**	29.3

TABLE IV
The test errors (%) for the synthetic data with 50 training samples.

Trial	1	2	3	4	5	6	7	8	9	10
Cox	35.0	**31.6**	37.5	39.3	33.7	46.5	40.2	41.2	**33.9**	42.1
pSVM linear	**34.3**	35.1	37.6	34.3	34.8	40.3	41.8	40.9	35.7	**38.1**
pSVM rbf	35.8	**31.6**	37.5	33.1	**34.1**	**38.0**	**38.1**	**35.5**	35.8	39.1
LUPI	37.8	35.4	**35.5**	32.0	39.4	41.3	41.5	39.3	38.4	42.0

- Estimate the test error of the final model using the test data.

The SVM+/LUPI has three tunable parameters, C, γ, and σ. These parameters are estimated using the validation data, and we consider C in the range of $[10^{-1}, 10^2]$, γ in $[10^{-3}, 10^1]$, and σ in $[2^{-2}, 2^2]$ for model selection. For pSVM with RBF kernel, we consider C and $\tilde{C}$ in the range of $[10^{-1}, 10^2]$, and σ in $[2^{-2}, 2^2]$.

Further, the experiment is performed ten times with different random realizations of the training, validation, and test data. In this experiment, the average proportion of the censored observation is 16.1% (or about 64 observations in the training set are censored). The test errors for ten trials are shown in Table I. The average test errors in percentage (along with standard deviations) for the Cox model, pSVM with linear kernel, pSVM with RBF kernel, and LUPI are 27.5±1.0, 25.6±1.4, 26.1±0.9, and 26.2±1.4, respectively.

The pSVM with linear kernel achieves the lowest test error among the methods in most trials. Comparing the pSVM method with different kernels, it is not surprising to find that pSVM with linear kernel performs better than that with RBF kernel. Because our synthetic data is generated from a nearly linear model and there is intrinsic linearity in the data. Methods with linear kernel are expected to perform better than those with RBF kernel.

The Cox model has the highest test error in most trails. The results illustrate potential advantage of using the SVM-based methods. Note that SVM-based methods yield similar or superior performance *vs.* classical Cox models, even thought the training and test data is generated using exponential distributions (for which the Cox method is known to be statistically optimal).

A. Number of Training Samples

To investigate the effect of training sample size on the test errors, the training sample size is reduced to 250, 100 and 50. The validation sample sizes are changed accordingly. The results are reported in Table II, III and IV.

For 250 training samples, the average test errors for the Cox model, pSVM with linear kernel, pSVM with RBF kernel, and LUPI are 29.2±1.0, 27.9±1.1, 28.3±1.3, and 28.7±1.9, respectively. The pSVM with linear kernel has the best performance in five trials. The relative performances between the pSVM with RBF kernel and LUPI are roughly the same. However, the performance gap between the Cox model and the pSVM with linear kernel is closing when the size of the training data is reduced. This observation is more evident when the sample size is reduced to 100. For 100 training samples, the Cox model has the lowest test error in four trials, whereas the pSVM with linear kernel has the best performance in three trials only.

When the training sample size is further reduced to 50, both the Cox model and the pSVM with linear kernel are outperformed by the pSVM with RBF kernel. This can be attributed to the high dimensionality of the input (feature) vectors. With high dimensional input vectors, methods with linear kernel fail to capture the linearity of the data when only 50 samples are available for training. It is also expected

TABLE V

TEST ERRORS AS A FUNCTION OF TRAINING SAMPLE SIZE.

Training size	50	100	250	400
Censoring	16.6%	15.9%	16.4%	16.1%
Cox	38.1 ± 4.6	**31.5 ± 2.4**	29.2 ± 1.0	27.5 ± 1.0
pSVM linear	37.3 ± 2.9	31.7 ± 3.1	**27.9 ± 1.1**	**25.6 ± 1.4**
pSVM rbf	**35.8 ± 2.4**	32.0 ± 1.5	28.3 ± 1.3	26.1 ± 0.9
LUPI	38.3 ± 3.2	33.2 ± 4.3	28.7 ± 1.9	26.2 ± 1.4

TABLE VI

TEST ERRORS AS A FUNCTION OF NOISE LEVEL.

Noise level	0	0.1	0.2	0.5
Censoring	15.9%	16.0%	17.2%	17.7%
Cox	**11.1 ± 0.4**	22.5 ± 1.7	28.7 ± 1.8	36.3 ± 1.3
pSVM linear	14.2 ± 1.0	**21.1 ± 1.8**	**27.1 ± 2.0**	34.8 ± 1.1
pSVM rbf	15.1 ± 1.5	22.5 ± 0.9	27.2 ± 2.1	36.0 ± 1.4
LUPI	14.3 ± 0.7	22.8 ± 1.7	27.5 ± 2.1	**34.7 ± 2.0**

that the estimated Cox model is not accurate due to the small sample size.

Table V shows the relative performance of the five methods, as a function of sample size. The pSVM with linear kernel outperforms all other methods when the training sample size is larger than 250. This is not surprising, because the linear space matches the synthetic data model. As expected, with increasing number of training samples, the relative advantage of the SVM-based methods is more noticeable. Nonetheless, the Cox model is more competitive for moderate training sample size (100).

B. *Noise Level in the Survival Time*

To examine the effect of noise level in the survival time on the test errors, noise with different variances are added to the survival time. The noise variance ranges from 0 to 0.5 and the training and validation sample sizes are kept at 250. The test errors are summarized in Table VI.

It is evident that the test errors are reduced in all methods when the noise variance is decreased. When there is no noise in the survival time, the data are generated from a distribution that follows the Cox modeling assumption. It is expected that the Cox model achieves the lowest test error under low-noise scenario. However, the increasing of noise level has much larger negative effect in the Cox modeling approach. The test error is increased from 11% to 36% when the noise level is raised from 0 to 0.5. Meanwhile, for the same changes in the noise levels, the test errors of the SVM-based approaches are raised from 14% to 35%.

Apart from the zero-noise scenario, the pSVM with linear kernel achieves the lowest average test error when the noise variance is less than 0.2. The LUPI, however, has the best performance when the noise level is higher than 0.2. It can be concluded that the SVM-based methods show more robustness to noisy data.

C. *Proportion of Censoring*

We also adjust the proportion of censoring in the training data to investigate the effect of censoring on the test errors. The percentage of censoring observations in the training data varies from 6% to 46% in our experiment. The noise variance is set to 0.2 and the training and validation sample sizes are kept at 250. The experiment results are summarized in Table VII.

TABLE VII

TEST ERRORS AS A FUNCTION OF CENSORING RATE.

Censoring	6.1%	30.6%	38.6%	46.0%
Cox	27.4 ± 2.0	33.8 ± 1.6	38.6 ± 2.2	42.0 ± 1.0
pSVM linear	**26.1 ± 1.6**	**31.5 ± 1.8**	36.8 ± 1.9	41.8 ± 2.4
pSVM rbf	26.9 ± 1.7	32.4 ± 2.5	**36.7 ± 1.3**	**39.9 ± 1.4**
LUPI	28.0 ± 2.7	32.5 ± 2.2	37.1 ± 2.1	41.3 ± 1.5

When less than 30% of the training data are censored, the pSVM linear gives the lowest test error. On the contrary, if a large portion of the observations are censored (about 40% or more), the pSVM with RBF kernel outperforms all other methods. With more censored observations in the training set, more observed survival times are obtained by the non-linear operator in (1). Hence, the linearity within the data is no longer maintained, and methods with non-linear parameterization (kernel) are expected to achieve better performances.

VI. REAL-LIFE DATA SETS

This section describes empirical comparisons using four real-life data sets from the *Survival* package in R [12]. For all comparisons, the common decision space for SVM+ uses the linear kernel while the unique correction space uses the RBF kernel. For the pSVM method, both linear and the RBF kernels are investigated. In all experiments, the time of interest τ was set to the median of the observed survival times. Our experiments for the four medical data sets follow the following procedure [2], [10]:

- Use five-fold cross-validation to estimate the test errors.
- Within each training fold, the parameter tuning (model selection) is performed through a five-fold resampling.

Our experimental set-up uses double resampling procedure [2]. One level of resampling is used for estimating the test error of a learning method, and the second level is for tuning the model parameters (or model selection). During the model selection stage, the possible choices of tuning parameters are C and $\tilde{C}$ in the range of $[10^{-1}, 10^{2}]$, γ in $[10^{-3}, 10^{1}]$, and σ in $[2^{-2}, 2^{2}]$. Since there is no definite class label for the censored observation with $U_i < \tau$, the test errors are reported based on samples with definite labels, *i.e.*, exact observations and censored observations with $U_i \geq \tau$. Further, model parameters are selected based on the performance with those samples with well-defined labels.

1) Veteran Data Set: The *veteran* data set is from the Veterans' Administration Lung Cancer Study which is a randomised trial of two treatment regimens for lung cancer. In the *veteran* data set, there are 137 instances (observations) and each instance has 10 attributes. Less than 7% of the instances are censored. Among the nine censored instances, one has the observed survival time less than the time of interest. In other words, only one instance is associated with the uncertain class label in the *veteran* data set.

2) Lung Data Set: The *lung* data set studied the survival and usual daily activities in patients with advanced lung cancer by the North Central Cancer Treatment Group (NCCTG). There are 167 instances in this data set, and each instance has 8 attributes. About 28% of the instances are censored, and 21 censored instances are linked to uncertain class labels.

TABLE VIII

SUMMARY OF THE *Survival* DATA SETS AND THE EXPERIMENT RESULTS.

Data set	Veteran	Lung	PBC	Stanford2
Size	137	167	258	157
Attributes	10	8	22	2
$\delta = 0$	9	47	147	55
Censored %	6.57	28.14	56.98	35.03
Uncertain cls	1	21	54	8
Cox	**23.4 ± 4.6**	43.3 ± 5.6	34.3 ± 7.1	51.9 ± 4.7
pSVM linear	27.2 ± 7.8	**38.3 ± 6.2**	26.2 ± 2.5	53.9 ± 7.4
pSVM rbf	32.0 ± 5.9	42.5 ± 8.0	**23.5 ± 5.2**	**34.3 ± 6.2**
LUPI	30.4 ± 4.5	38.3 ± 9.9	25.3 ± 10.6	42.4 ± 17.7

3) PBC Data Set: The *pbc* data set is from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. The *pbc* data set contains 258 instances and each instance has 22 attributes. More than half of the instances are censored, and 54 censored instances do not have the definite class labels.

4) Stanford2 Data Set: The fourth data set is the *stanford2* data set from the Stanford Heart Transplant data, which contains 157 instances, each with 2 attributes. More than 35% instances are censored and 8 of them are associated with the uncertain labels.

The descriptions of the data sets are summarized in Table VIII. The fourth row indicates the proportions of censored observations in the data sets. The fifth row shows the number of censored observation with $U_i < \tau$ when τ is set to the median of the observed survival times. Table VIII also shows the test errors from different methods applied to the four data sets. Note that the SVM-based approaches achieve the lowest test error in three of the four data sets. On the other hand, the Cox model gives the best performance in the *veteran* data set. In these experiments, the number of training samples is fixed, so we cannot make any conclusions regarding the effect of sample size on methods' performance. However, we can make inferences about inherent non-linearity in some of the data sets. For example, for the *stanford2* data set, non-linear pSVM performs much better than other methods using linear parameterization. So we can infer this data set requires non-linear modeling.

These results illustrate the effect of censoring on generalization performance. For small proportion of censoring (such as 6%) in the data, the Cox model gives the lowest test error. However, the SVM-based methods show their advantages when the proportion of censoring is increased. Further, relative advantage of SVM-based approaches becomes quite evident for higher-dimensional survival data.

These results also show large variability of estimated test errors, due to partitioning of available data into five (training, test) folds. This variability is reflected in large standard deviations of test error rates. Direct comparisons suggest that SVM-based methods yield smaller or similar test error in each (training, test) fold. Another reason for variability of the SVM-based model estimates is due to model selection via resampling. Notably, standard deviations of error rates for all SVM-based methods shown in Table VIII are consistently higher than standard deviations for the Cox model (which has no tunable parameters). This underscores the importance of robust model selection strategies for SVM-based methods, which would be the focus of our future work.

VII. DISCUSSION AND CONCLUSIONS

This paper proposes predictive modeling of high-dimensional survival data as a binary classification problem. We apply the LUPI formulation and SVM with uncertain class labels to solve the problem. Both methods incorporate the information about survival time to estimate an SVM classifier. We have illustrated the advantages and limitations of these modeling approaches using synthetic and real-life data sets.

Advanced SVM-based methods appear very effective when the proportion of censoring in training data is large, or the observed survival time does not follow the classical probabilistic assumptions, *e.g.*, the exponential distribution [1], [11]. On the other hand, with fewer censored observations the Cox modeling approach may perform better. Further, the relative performance of LUPI and pSVM depends on the intrinsic linearity/non-linearity of the data itself. In particular, superior performance of the pSVM with RBF kernel for the *stanford2* data indicates an intrinsic non-linearity of this data set.

The equal misclassification cost is assumed throughout this paper; however, realistic medical applications use unequal costs. We will incorporate different misclassification costs into the proposed SVM-based formulations. Further, our methodology for predictive modeling of survival data can be readily extended to other (non-medical) applications, such as predicting business failure (aka bankruptcy) or predicting marriage failure (aka divorce).

REFERENCES

[1] O. Aalen, Ø. Borgan, H. Gjessing, and S. Gjessing, *Survival and Event History Analysis: A Process Point of View*, ser. Statistics for Biology and Health. Springer-Verlag New York, 2008.

[2] V. Cherkassky and F. Mulier, *Learning from data: concepts, theory, and methods*. Wiley, 2007.

[3] F. Khan and V. Zubek, "Support Vector Regression for censored data (SVRc): A novel tool for survival analysis," in *Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on*, Dec. 2008, pp. 863–868.

[4] J. Shim and C. Hwang, "Support vector censored quantile regression under random censoring," *Comput. Stat. Data Anal.*, vol. 53, no. 4, pp. 912–919, Feb. 2009.

[5] P. K. Shivaswamy, W. Chu, and M. Jansche, "A support vector approach to censored targets," in *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ser. ICDM '07. Washington, DC, USA: IEEE Computer Society, 2007, pp. 655–660.

[6] V. N. Vapnik, *Estimation of dependences based on empirical data, Empirical inference science: afterword of 2006*. Springer, 2006.

[7] V. Vapnik and A. Vashist, "2009 special issue: A new learning paradigm: Learning using privileged information," *Neural Networks*, vol. 22, no. 5-6, pp. 544–557, July 2009.

[8] E. Niaf, R. Flamary, C. Lartizien, and S. Canu, "Handling uncertainties in SVM classification," in *Statistical Signal Processing Workshop (SSP), 2011 IEEE*, June 2011, pp. 757–760.

[9] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*. MIT Press, 1999, pp. 61–74.

[10] L. Liang, F. Cai, and V. Cherkassky, "Predictive learning with structured (grouped) data," *Neural Networks*, vol. 22, no. 5-6, pp. 766–773, 2009.

[11] M. Zhou, "Use software R to do survival analysis and simulation. a tutorial," http://www.ms.uky.edu/ mai/Rsurv.pdf.

[12] T. M. Therneau, *A Package for Survival Analysis in R*, 2013, r package version 2.37-4. [Online]. Available: http://CRAN.R-project.org/package=survival