Regularized Weighted Linear Regression for High-dimensional Censored Data

Yan Li* Bhanukiran Vinzamuri† Chandan K. Reddy‡

Abstract

Survival analysis aims at modeling time to event data which occurs ubiquitously in many biomedical and healthcare applications. One of the critical challenges with modeling such survival data is the presence of censored outcomes which cannot be handled by standard regression models. In this paper, we propose a regularized linear regression model with weighted least-squares to handle the survival prediction in the presence of censored instances. We also employ the elastic net penalty term for inducing sparsity into the linear model for effectively handling high-dimensional data. As opposed to the existing censored linear models, the parameter estimation of our model does not need any prior estimation of survival times of censored instances. In addition, we propose a self-training framework which is able to improve the prediction performance of our proposed linear model. We demonstrate the performance of the proposed model using several real-world high-dimensional biomedical benchmark datasets and our experimental results indicate that our model outperforms other related competing methods and attains very competitive performance on different datasets.

Keywords: survival analysis; linear regression; censored data; self-training; sparse methods; least squares; high-dimensional data.

1 Introduction

Survival analysis aims at modeling data in longitudinal studies where the observations are monitored over period of time [1]. The monitoring continues until the occurrence of a certain *event of interest*. However, the event of interest may not always be observed during the study period which gives rise to censoring in the dataset. *Censoring* makes survival analysis more challenging compared to the standard regression setting, and for such instances the last observed time is known as *censored time*. The most common form of censoring that occurs in real-world scenarios is *right censoring* where the survival time is known to be longer than or equal to censored time, but its precise value is unknown. In the rest of the paper, we refer to right censored data as censored data, unless otherwise specified.

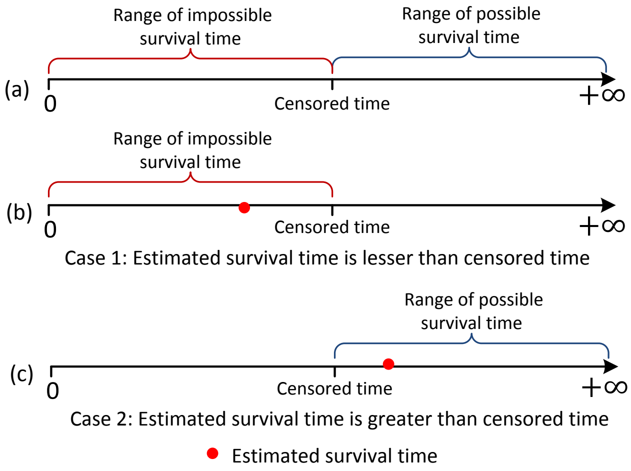

Range of impossible survival time
Range of possible survival time
(a)
0
Censored time
+∞
Range of impossible survival time
(b)
0
Censored time
+∞
Case 1: Estimated survival time is lesser than censored time
Range of possible survival time
(c)
0
Censored time
+∞
Case 2: Estimated survival time is greater than censored time
● Estimated survival time

Figure 1: Relationship between estimated survival time and censored time for a right censored observation.

Figure 1 shows three timelines to demonstrate the relationship between the estimated survival time and the censored time. Based on the definition of right censoring, the timeline can be separated into two parts, the range of impossible survival time and the range of possible survival time using the censored time (Figure 1(a)). Hence, there are two cases that can arise. Case 1: once the estimated survival time is lesser than the censored time, then it falls into the range of impossible survival time and the real difference between the estimated survival time and the actual survival time is definitely greater than the difference between the estimated survival time and observed censored time (Figure 1(b)). Thus, the model should give more emphasis to this case with assigning more weight for this case in the loss function to reduce such an occurrence. Case 2: If the estimated survival time is greater than the censored time, then it falls into the range of possible survival time (Figure 1(c)), and hence, in the loss function, the model should assign less weight for this case.

Motivated by this observation, in this paper, we propose a *Regularized Weighted Residual Sum-*

*Department of Computer Science, Wayne State University, Detroit, MI. E-mail: rock_liyan@wayne.edu

†Department of Computer Science, Wayne State University, Detroit, MI. E-mail: bhanukiranv@wayne.edu

‡Department of Computer Science, Wayne State University, Detroit, MI. E-mail: reddy@cs.wayne.edu

of-Squares (RWRSS) algorithm which imposes more penalty to the first case and less penalty to the second case. Thus, the proposed RWRSS is able to effectively handle the censored instances. Additionally, we also employ the elastic net as penalty to sparsify the learned coefficients, so that the RWRSS is able to avoid overfitting and will be able to deal with high-dimensional datasets. We propose a linear model because it is simple, effective and scalable. In survival analysis, linear regression for data analysis with censored observations is an alternative research direction that has attracted broad attention in the fields of data mining and biostatistics [2, 3, 4].

Some linear models such as Tobit regression [5] and Buckley-James (BJ) regression [2] were proposed to handle censored observations. These methods employ the Kaplan-Meier (K-M) estimator [6] (in the case of BJ) and the Gaussian distribution (in the case of Tobit regression) to approximate the survival times of censored instances for satisfying the least-squares principle. However, these approximation methods will induce bias into the final model since the actual survival times of censored instances cannot be observed. It induces bias because the KM estimation cannot accurately estimate the survival time of censored instances, and this estimated inaccurate survival time will be used to train the model. This makes the prediction problem more complex because the survival time of censored instances are calculated using the integral of the KM estimator.

In contrast to these existing methods, we do not approximate the survival times of censored instances. RWRSS aims at directly minimizing the difference between the estimated survival time and actual survival time of uncensored instances and ensures that the estimated survival time of censored instances is longer than the censored times. Thus, comparing to the existing linear censored regression models, our proposed model simplifies the prediction problem. In this paper, we demonstrate that such a simplification improves the prediction performance of the proposed model. The concordance index (C-index) [7] is the most commonly used performance metric in survival analysis which measures the concordance between the orderings of the survival times and predicted marker values. *Because the actual survival time of censored instances are unknown, it is infeasible to calculate the concordance between a pair of censored instances. Hence, an accurate estimation of the survival time of censored instances is not possible or needed.*

From the viewpoint of traditional data mining, survival analysis can also be viewed as a semi-supervised learning problem, where the uncensored instances can be viewed as labeled data and the censored instances can be viewed as unlabeled data. However, different from most of the existing semi-supervised algorithms which are focused on classification, survival analysis is a regression problem. Motivated by self-training [8], which uses the confidence estimated labels of unlabeled data in the next training round, we develop a framework which uses the proposed RWRSS model to infer the survival time of censored instances.

However, different from the traditional self-training approaches, in the first training round RWRSS is trained from both labeled (uncensored) and unlabeled (censored) instances. In the right censoring scenario, for certain censored instances, the censored time should be equal to or less than the survival time. Thus, once the estimated survival time is greater than the censored time, we will use the estimation to approximate the survival time of censored instances and train a new model based on the updated training instances in the next training round. Experimental results over high-dimensional biomedical datasets indicate that our model outperforms other related competing methods and attains very competitive C-index values on high-dimensional datasets.

The main contributions of this paper can be summarized as follows:

- Propose a novel weighted linear regression method for prediction problems with censored observations which avoids the use of approximate survival time of censored data during the training phase.
- Develop a self-training framework which involves both uncensored and censored instances in each training round and is able to improve the prediction performance of the proposed RWRSS model.
- Demonstrate the performance of the proposed censored regression method using real-world high-dimensional cancer gene expression survival datasets and compare it with several existing survival estimation methods.

The rest of the paper is organized as follows. In Section 2, the related data mining approaches for survival analysis are discussed. Our proposed approaches are explained in detail in Section 3. Section 4 demonstrates our experimental results on several real-world datasets while Section 5 concludes our discussion.

2 Related Work

In this section, we present the related work in the area of data mining methods for survival analysis and highlight the differences and relationships between our proposed model and other existing works.

Cox proportional hazards model [9] is one of the earliest and most widely used survival analysis method which has garnered significant interest from researchers in both statistics and data mining communities. To deal with high-dimensional data some regularization methods have been integrated with it. These methods include LASSO-COX [10] which introduces the L_1 norm penalty in the Cox log-likelihood loss function, Elastic-Net Cox (EN-COX) [11] which uses the elastic net penalty term and the kernel elastic net penalized Cox regression [12, 13].

The linear regression model, together with the least-squares estimator, is one of the fundamental models in data analysis. One of its main drawbacks is that it can not be directly used in survival analysis because the actual survival times of censored instances are missing for censored instances. The Tobit model [5] is the earliest attempt to extend the linear regression for data analysis with censored observations. Then, in the late 1970s and early 1980s, a number of works [14, 2, 15] have extended the least-squares principle to handle censored observations. The estimator of Miller [14] requires that the censoring time satisfy the same regression model as the survival time, while the estimator of Koul et al. [15] requires that the censoring time be independent of covariates. Miller and Halpern's study [16] have shown that the Buckley-James (BJ) estimator [2] is robust compared to the other methods. Wang et al. applied the elastic net penalty to the BJ regression (EN-BJ) [3] to handle the high-dimensional survival data. Accelerated failure time (AFT) model [17] can also be viewed as an extension of linear model which assumes that the relationship of the logarithm of survival time T and the covariates is linear in nature [1].

In this paper, we propose the RWRSS algorithm to handle the survival prediction with censored instances in high-dimensional data. Different from the Tobit regression, we solve the prediction problem by optimizing the desired objective function directly rather than doing a maximum likelihood estimation. The loss function that is optimized is regularized using the elastic net penalty which can induce the required sparsity and efficiently handle the high-dimensionality. Comparing with the BJ and EN-BJ methods, our model does not need to compute the K-M estimator to approximate the survival time of censored instances during the training process. In addition, we also propose a framework motivated by the idea of self-training which can improve the prediction performance of our proposed linear model.

3 Regularized Weighted Linear Regression for Survival Analysis

In this section, we will explain the details of the proposed regularized weighted linear regression model for predicting survival times. We will first discuss the proposed weighted loss function along with its main intuition. Later, the regularized weighted linear regression and the optimization procedure will be explained in details. Finally, a self-training framework for handling survival data with right censored instances will be discussed. This self-training framework is used with our regularized weighted linear regression as the base learning algorithm. In survival analysis, one can either observe the survival time (T_i) or the censored time (U_i) for i^{th} instance but not both of them. The dataset is considered to be right censored if only if $y_i = \min(T_i, U_i)$ can be observed during the study. An instance in the survival data is usually represented by a triplet (X_i, y_i, δ_i), where X_i is $1 \times p$ feature vector, δ_i is the censored indicator; $\delta_i = 1$ for a uncensored instance and $\delta_i = 0$ for a censored instance [18]. The *observed time* y_i is equal to the survival time T_i for uncensored instances and U_i otherwise.

$$y_i = \begin{cases} T_i & \text{if} \quad \delta_i = 1 \\ U_i & \text{if} \quad \delta_i = 0 \end{cases} \tag{3.1}$$

3.1 Objective Function For censored observations, the exact difference between the estimated outcome and the actual target value cannot be measured. The estimated survival time for a right censored instance should be either equal to or larger than its censored time. For the i^{th} instance, if $\delta_i = 1$, then the estimated survival time of the proposed model should be as close as possible to y_i, and hence the standard squared residual can be used as loss function for uncensored instances; however, if $\delta_i = 0$, the estimated survival time should be greater than y_i, and hence in the loss function, we should give more weight to the censored instances whose estimated survival time is lesser than censored time and less weight to the censored instances whose estimated survival time is greater than censored time. Thus, we propose the following weighted residual sum-of-squares (WRSS) as the objective function to minimize.

$$WRSS = \sum_{i=1}^{N} (y_i - X_i\beta)^2 w_i \tag{3.2}$$

where weight w_i is defined as follows:

$$w_i = \begin{cases} 1 & \text{if} \quad \delta_i = 1 \\ \tau & \text{if} \quad \delta_i = 0 \text{ and } y_i \geq X_i\beta \\ 0 & \text{if} \quad \delta_i = 0 \text{ and } y_i < X_i\beta \end{cases} \tag{3.3}$$

From Eqs.(3.2) and (3.3), we can see that the WRSS calculates the standard residual value $(y_i - X_i\beta)$ for uncensored observations ($\delta_i = 1$). However, for censored observations, it ignores the difference between the estimated output and the censored time when the estimated output is greater than the censored time. For the right censored observations, we know that the actual survival time is equal to or greater than the censored time; therefore, when the estimated output is lesser than the censored time, then the difference between the actual survival time and the estimated survival time is indeed greater than $(y_i - X_i\beta)$. Hence, τ is a constant which should be greater than 1 and is a parameter of the model that needs to be empirically determined because it is infeasible to measure the true difference between the estimated output and the actual survival time of the corresponding censored instance.

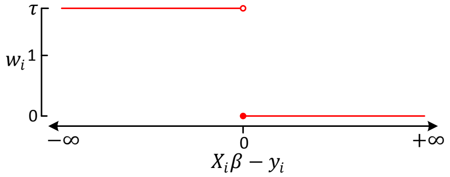


Figure 2: w_i is a step function for censored instances.

In our proposed model, the elastic net [19] is used as the penalty term. The corresponding optimization problem is formulated below:

$$arg \min_{\beta} \frac{1}{N} \sum_{i=1}^{N} (y_i - X_i\beta)^2 w_i + \lambda \left(\alpha \|\beta\|_1 + \frac{1-\alpha}{2} \|\beta\|_2^2 \right) \tag{3.4}$$

where $\lambda \geq 0$ is the regularization parameter, and $0 \leq \alpha \leq 1$ is used to adjust the weights of the L_1 and L_2 norm penalties.

3.2 Optimization From Eq. (3.3), we can see that w_i is a constant when $\delta_i = 1$ and it is a step function when $\delta_i = 0$ (see Figure 2). To handle this discontinuity in Eq. (3.4) due to the presence of the step function, we propose an iterative optimization method based on coordinate descent method. Coordinate descent minimizes a multi-variate function by minimizing it along only one direction at a time. Let us suppose, except β_k, all other $\tilde{\beta}_l \, (l = 1, 2, 3, \cdots, p$ and $l \neq k)$ have already been estimated and we would like to partially optimize with respect to β_k. The coordinate-wise updating [20] is done as follows:

$$\tilde{\beta}_k \leftarrow \frac{S(\frac{1}{N}\sum_{i=1}^{N} w_i x_{ik}(y_i - \tilde{y}_{i\bar{k}}), \ \lambda\alpha)}{\frac{1}{N}\sum_{i=1}^{N} w_i x_{ik}^2 + \lambda(1-\alpha)} \tag{3.5}$$

where $\tilde{y}_{i\bar{k}} = \sum_{l \neq k} x_{il}\tilde{\beta}_l$ is the fitted value excluding the contribution from x_{ik}, $S(Z, \ \gamma) = sign(Z) \cdot (|Z| - \gamma)_+$ is the soft-thresholding operation; in addition, $sign(\cdot)$ is the signum function, and $(|Z| - \gamma)_+$ refers to the positive part, which is $|Z| - \gamma$ if $(|Z| - \gamma) > 0$ and 0 otherwise. This update is simply the univariate regression coefficient of the partial residual sum of squares $(y_i - \tilde{y}_{i\bar{k}})$ on the k^{th} variable. In each iteration, all of the p coefficient variables are repeatedly updated until convergence.

Now, one of the problems that arise during this optimization is the determination of w_i for each observation in each iteration. From Eq. (3.3), we can see that the value of w_i is determined by δ_i, y_i, and $X_i\beta$, where for a particular i^{th} observation, δ_i, y_i, and X_i will not change during the coordinate decent process, but each element of β will be updated one by one based on the coordinate-wise method. In the optimization process, we use the latest updated coefficient vector to approximate β; thus, in the d^{th} iteration before updating the coefficient of k^{th} feature, the latest updated $\beta^{d,k-1}$ can be represented by

$$\beta^{(d,k-1)} = \{\beta_1^d, \ldots, \beta_{(k-1)}^d, \beta_k^{d-1}, \beta_{(k+1)}^{d-1}, \ldots, \beta_p^{d-1}\}$$

and we have

$$\begin{aligned} X_i\beta &\approx X_i\beta^{(d,k-1)} \\ &= X_i\beta^{(d,k-2)} + X_{i,k-1} \cdot \beta_{(k-1)}^d - X_{i,k-1} \cdot \beta_{(k-1)}^{d-1} \end{aligned} \tag{3.6}$$

Algorithm 1 outlines the basic learning methodology for the proposed RWRSS model. In line 1, we initialize the estimator of the parameter $\hat{\beta}$ to be the zero vector. In lines 4-7, we calculate the updated w_i for each training instance, and each element of the coefficient vector is updated using the coordinate-wise update in line 8. Eq.(3.6) can be updated in $O(1)$ based on the previous result; thus, for N instances, a complete cycle costs $O(Np)$ operations where p is the number of features. Hence, the overall time complexity of the proposed model is $O(Np)$.

Algorithm 1: Regularized weighted residual sum-of-squares (RWRSS)

```
   Input: Training data $(X, \delta, y)$, Regularization
          parameter $\lambda$, Adjustment Weight $\alpha$
   Output: $\hat{\beta}$
1  Initialize: $\hat{\beta} \leftarrow \mathbf{0}$;
2  repeat
3      for $k = 1$ to $p$ do
4          for $i = 1$ to $N$ do
5              Calculate $X_i\beta$ using Eq.(3.6);
6              Update $w_i$ using Eq.(3.3);
7          end
8          $\tilde{\beta}_k \leftarrow \frac{S(\frac{1}{N}\sum_{i=1}^{N} w_i x_{ik}(y_i - \tilde{y}_{i\bar{k}}), \ \lambda\alpha)}{\frac{1}{N}\sum_{i=1}^{N} w_i x_{ik}^2 + \lambda(1-\alpha)}$;
9      end
10     $\hat{\beta} \leftarrow \tilde{\beta}$;
11 until Convergence of $\beta$;
```

3.3 Theoretical Analysis We will now provide the convergence analysis of Algorithm 1. Since w_i is updated iteratively based on the approximation made in Eq.(3.6) (line 6 of Algorithm 1), the proposed objective function is an iteratively re-weighted least squares (IRLS) which is different from the standard weighted update of coordinate descent. Thus, the analysis of the descent property is needed to ensure that the proposed RWRSS algorithm indeed converges.

LEMMA 3.1. *The optimum value of Eq.(3.4) with iteratively updated w_i is upper bounded by the optimum value of Eq.(3.4) with constant initial w_i (denote by $w_i^{(0)}$).*

Proof. Since β is initialized by a zero vector, we have $X_i\beta = 0 \leq y_i$ for all i. Then based on Eq.(3.3), we have

$$w_i^{(0)} = \begin{cases} 1 & \text{if} \quad \delta_i = 1 \\ \tau & \text{if} \quad \delta_i = 0 \end{cases}$$

Let $w_i^{(f)}$ denote the final updated weight of i^{th} instance, then we have

$$w_i^{(0)} \geq w_i^{(f)} \text{ for all } i \tag{3.7}$$

Let $L(w_i^{(0)}, \hat{\beta}^{(0)})$ be the optimum value of Eq.(3.4) with constant initial w_i, where $\hat{\beta}^{(0)}$ is the corresponding learned coefficient. Similarly, let $L(w_i^{(f)}, \hat{\beta}^{(f)})$ be the optimum value of Eq.(3.4) with iteratively updated w_i, where $\hat{\beta}^{(f)}$ is the corresponding learned coefficient. Thus, we have

$$L(w_i^{(0)}, \hat{\beta}^{(0)}) \geq L(w_i^{(f)}, \hat{\beta}^{(0)}) \geq L(w_i^{(f)}, \hat{\beta}^{(f)})$$

where the first inequality is based on Eq.(3.7), and the second inequality is because $\hat{\beta}^{(f)}$ is the learned optimal coefficient with respect to $w_i^{(f)}$. Therefore, the optimum value of Eq.(3.4) with iteratively updated w_i is upper bounded.

THEOREM 3.1. *The objective function given in Eq. (3.4) converges during the learning process.*

Proof. In the d^{th} iteration of coordinate descent, the k^{th} coefficient is updated by

$$\beta_k^d \leftarrow \min_{\beta_k} L(w_i^{(d,k-1)}, \beta_1^d, \ldots, \beta_{(k-1)}^d, \beta_k, \beta_{(k+1)}^{d-1}, \ldots, \beta_p^{d-1}) \tag{3.8}$$

where $w_i^{(d,k-1)}$ is the latest updated weight of i^{th} instances based on $\beta^{(d,k-1)}$. Similarly, as discussed in Lemma (3.1) we have $w_i^{(0)} \geq w_i^{(d,k-1)}$, and

$$\begin{aligned} & L(w_i^{(0)}, \beta_1^d, \ldots, \beta_{(k-1)}^d, \beta_k^{d(0)}, \beta_{(k+1)}^{d-1}, \ldots, \beta_p^{d-1}) \\ & \geq L(w_i^{(d,k-1)}, \beta_1^d, \ldots, \beta_{(k-1)}^d, \beta_k^{d(0)}, \beta_{(k+1)}^{d-1}, \ldots, \beta_p^{d-1}) \\ & \geq L(w_i^{(d,k-1)}, \beta_1^d, \ldots, \beta_{(k-1)}^d, \beta_k^d, \beta_{(k+1)}^{d-1}, \ldots, \beta_p^{d-1}) \end{aligned}$$

where $\beta_k^{d(0)}$ is the optimal value of the k^{th} coefficient in d^{th} iteration if the weight of the instances is initialized as $w_i^{(0)}$. Thus, we can say that in each step of the learning process, the value of the objective function is upper bounded by the constant weighted ($w_i^{(0)}$) objective function. Based on the convergence of coordinate descent we know that the constant weighted objective function has converged during the learning process. Therefore, we can conclude that the objective function converges during the learning process.

3.4 Self-training Framework for Right Censored data As pointed in the previous sections, mining from dataset with censored observations is closely related to semi-supervised learning. The censored observations can be considered as unlabeled instances since the event of interest can take place in the future. Most of the existing semi-supervised learning methods focus on classification rather than regression, and unlabeled observations do not contain any labeling information at all. However, in survival analysis, for each censored instance we can observe a lower bound of the target value. In this section, we propose a self-training framework for right censored data (STC). The goal of this framework is to infer the correct event labels for the given censored instances. In our proposed framework, we employ the RWRSS as our base learning method and label the censored instances based on both the prediction and the corresponding censored time.

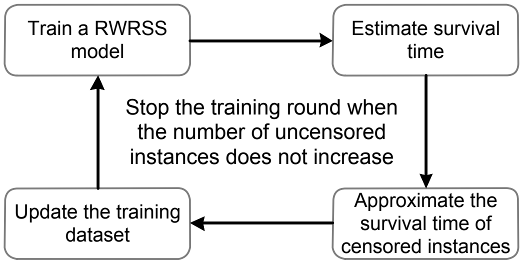

Train a RWRSS model
Estimate survival time
Stop the training round when the number of uncensored instances does not increase
Update the training dataset
Approximate the survival time of censored instances

Figure 3: A self-training framework for right censored data.

Figure 3 shows the diagram of the proposed STC framework. In survival data, it is evident that the actual (unobserved) survival time of any right censored instance should be no less than the observed censored time. Thus, in the t^{th} self-training learning round of the STC framework, if the estimated survival time of a censored instance is greater than or equal to the censored time, it will fall into the range of possible survival time and it will be viewed as a correct prediction. Then in the $t+1^{th}$ learning round, this instance will be viewed as an uncensored object whose survival time will be the estimated output in the t^{th} round and its corresponding censored indicator will be changed from 0 to 1. Hereby, in the $t+1^{th}$ learning round, the updated training data will contain more uncensored instances than the training data in the t^{th} round, and hence a robust model can be learned in the $t+1^{th}$ learning round.

Algorithm 2 describes the STC framework for right censored data. In line 3, we estimate the coefficients based on the proposed RWRSS algorithm. In lines 4-11, the objective values and the δ values for censored instances in the training dataset will be updated based on the predicted and the observed time. The training rounds will stop when the status (δ) and observed time (y) of the training data is not updated any further.

Algorithm 2: Self-Training framework for right Censored Data (STC)

```
   Input: Training data ($X$, $\delta$, $y$), Regularization
          parameter $\lambda$, Adjustment weight $\alpha$
   Output: $\beta$
1  Initialize: $c \leftarrow 0$, $\hat{\beta} \leftarrow \mathbf{0}$;
2  repeat
3      $\hat{\beta} = RWRSS(X, y, \delta, \lambda, \alpha)$;
4      for $i = 1$ to $N$ do
5          if $\delta_i == 0$ then
6              $\hat{y}_i = X_i\hat{\beta}$;
7              if $\hat{y}_i > y_i$ then
8                  $y_i = \hat{y}_i$; $\delta_i = 1$;
9              end
10         end
11     end
12 until $\delta$ and $y$ are not updated;
```

4 Experimental Results

In this section, we will first describe the datasets used in our evaluation and then provide the performance results along with the implementation details.

4.1 Dataset Description For our evaluation, we used several publicly available high-dimensional gene expression cancer survival benchmark datasets which can be downloaded from [1]. Here are the list of datasets that are used in our experiments.

- Norway/Stanford Breast Cancer Data (NSBCD).
- Van de Vijver's Microarray breast cancer (VDV).
- Lung adenocarcinoma (Lung).
- Mantle Cell Lymphoma (MCL) [2].
- The Dutch Breast Cancer Data (DBCD).
- Diffuse Large B-Cell Lymphoma (DLBCL).

All these datasets measure cancer survival using gene expression levels. Table 1 provides the details of the datasets that are being used. In this table, the column titled "# Censored" corresponds to the number of censored instances in each dataset. We used 5-fold cross validation when the number of instances is greater than 150 and 3-fold cross validation otherwise.

Table 1: Details of the datasets used in this paper.

Dataset	**# Instances**	**# Features**	**# Censored**
NSBCD	115	549	77
VDV	78	4705	44
Lung	86	7129	62
MCL	92	8810	28
DBCD	295	4919	216
DLBCL	240	7399	102

4.2 Evaluation Metrics Concordance index (C-index) or the *concordance probability*, is used to measure the performance of prediction models in survival analysis [7]. Let us consider a pair of bivariate observations (y_1, $\hat{y}_1$) and (y_2, $\hat{y}_2$), where y_i is the actual observation, and $\hat{y}_i$ is the predicted one. The concordance probability is defined as:

$$c = Pr(\hat{y}_1 > \hat{y}_2 | y_1 \geq y_2) \tag{4.9}$$

By definition, the C-index has the same scale as the area under the ROC (AUC) in binary classification, and if y_i is binary, then the C-index is same as the AUC. In the hazards ratio based regression models, the instances with a low hazard rate should survive longer, and the C-index will be calculated as follows:

$$c = \frac{1}{num} \sum_{i \in \{1 \cdots N\} \delta_i = 1} \sum_{y_j > y_i} I[X_i\hat{\beta} > X_j\hat{\beta}] \tag{4.10}$$

where *num* denotes the number of comparable pairs and $I[\cdot]$ is the indicator function. The C-index in other censored regression methods, which directly target the survival time, should be calculated as:

$$c = \frac{1}{num} \sum_{i \in \{1 \cdots N\} \delta_i = 1} \sum_{y_j > y_i} I[S(\hat{y}_j | X_j) > S(\hat{y}_i | X_i)] \tag{4.11}$$

where $S(\hat{y}_i | X_i)$ is the predicted target value for X_i.

4.3 Implementation Details All of the seven methods used for comparisons are implemented in R. The Cox and Tobit regression models are obtained from the *survival* package [21]. In the *survival* package, the *coxph* function is employed to train the cox model and

[1] http://user.it.uu.se/~liuya610/download.html

[2] http://llmpp.nih.gov/MCL/

Table 2: Performance comparison of the proposed method and seven other existing related methods using C-index values (along with their standard deviation).

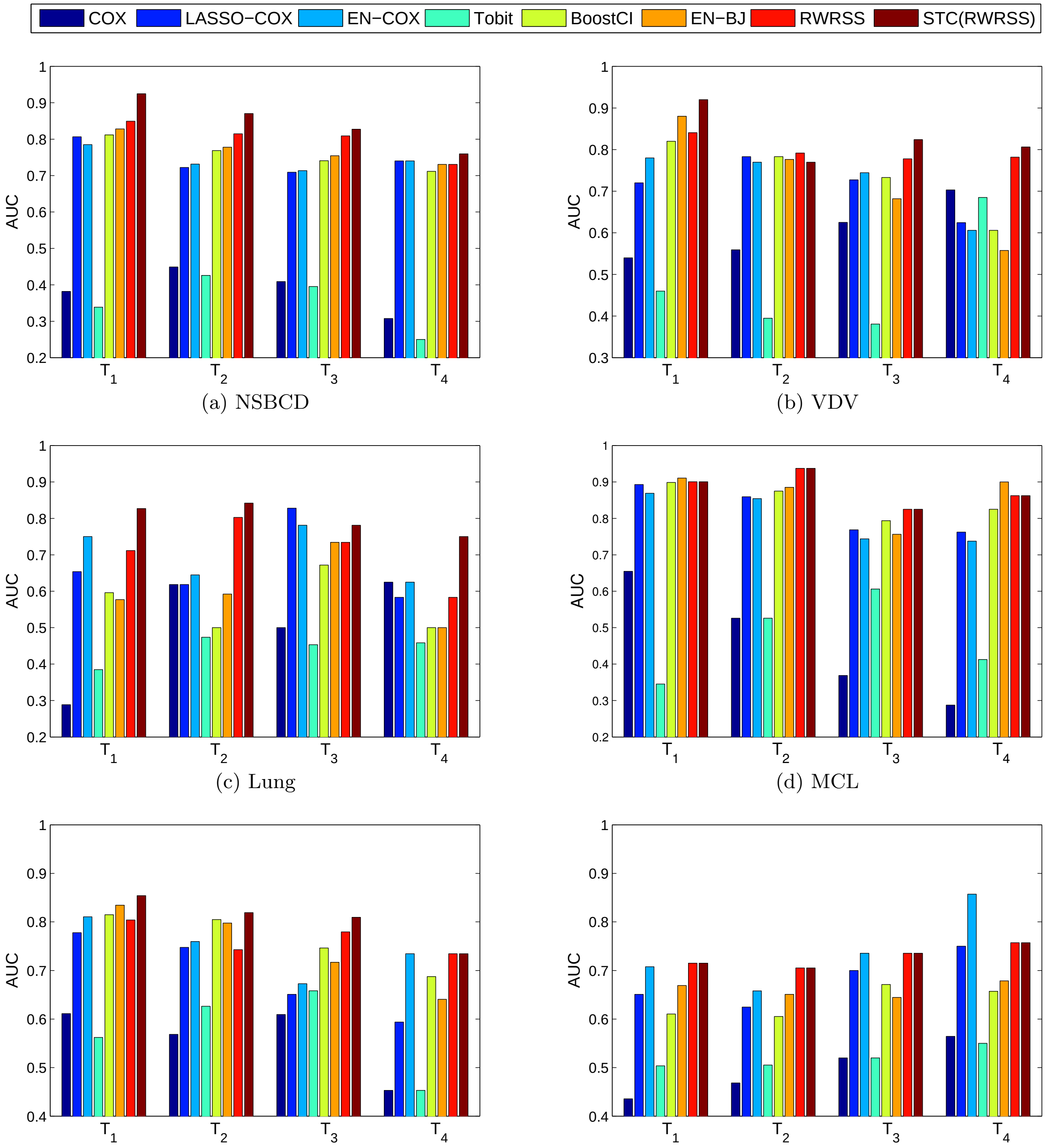
DataSet	**COX**	**LASSO-COX**	**EN-COX**	**BoostCI**	**OLS**	**Tobit**	**EN-BJ**	**RWRSS**	**STC(RWRSS)**
NSBCD	0.4411	0.5910	0.6046	0.6263	0.6333	0.3733	0.6215	0.6766	**0.7149**
	(0.0589)	(0.1086)	(0.1000)	(0.0831)	(0.1108)	(0.0214)	(0.0924)	(0.1277)	(**0.0836**)
VDV	0.5947	0.6428	0.6384	0.6641	0.5315	0.5112	0.6077	0.7207	**0.7445**
	(0.0997)	(0.0254)	(0.0603)	(0.0560)	(0.0086)	(0.1491)	(0.0648)	(0.0705)	(**0.0195**)
Lung	0.5139	0.6684	0.6639	0.5708	0.5716	0.4695	0.6634	0.6969	**0.7316**
	(0.1372)	(0.0867)	(0.0661)	(0.0883)	(0.0610)	(0.1321)	(0.1284)	(0.0430)	(**0.0313**)
MCL	0.5715	0.6742	0.6594	0.6895	0.4881	0.4914	0.7023	**0.7118**	**0.7118**
	(0.0446)	(0.0688)	(0.0655)	(0.0897)	(0.0414)	(0.0875)	(0.1038)	(**0.0737**)	(**0.0737**)
DBCD	0.5294	0.6850	0.7188	0.7045	0.5599	0.4869	0.7175	0.7216	**0.7404**
	(0.0634)	(0.0417)	(0.0304)	(0.0380)	(0.0717)	(0.0784)	(0.0396)	(0.0446)	(**0.0475**)
DLBCL	0.5097	0.6242	**0.6372**	0.5954	0.5052	0.4917	0.6228	0.6265	0.6265
	(0.0293)	(0.0416)	(**0.0359**)	(0.0170)	(0.0891)	(0.0524)	(0.0611)	(0.0657)	(0.0657)

the Efron's method [22] is used to handle the tied observations. The Tobit regression methods are trained using the *survreg* function with Gaussian distributions. Three sparse regression methods, namely, LASSO-COX, EN-COX, and EN-BJ, which are penalized versions using lasso and elastic net penalty terms are also used for our comparisons. LASSO-COX and EN-COX are built using the *cocktail* function in the *fastcox* package [23], while EN-BJ is implemented using the *bujar* package [4]. Boosting concordance index (BoostCI) [24] for survival data is an approach where the concordance index metric is modified to an equivalent smoothed criterion using the sigmoid function. In addition to the above mentioned six survival analysis methods, we also compared with the ordinary least squares (OLS) linear regression which has a similar form to the proposed methods. Note that, the OLS is only learned using the uncensored instances rather than the entire set of training instances since it cannot handle the censored instances, while the other methods are trained based on both uncensored and censored instances. The proposed model is implemented using C++ and the code will be made publicly available upon the acceptance of this paper.

In the experiments we use the C-index of the training dataset as a training error rate to monitor the training round of STC framework and terminate the learning round when the C-index decreases. In this scenario, the output of STC(RWRSS) is the model learned in the penultimate learning round.

4.4 Results and Discussion Table 2 provides the C-index values obtained with various censored regressions and OLS methods on the real-world high-dimensional cancer microarray datasets. The results show that our proposed model obtains higher C-index in most of the datasets and the STC framework is able to improve the prediction performance of RWRSS in some of the cases.

Figure 4 provides the histogram plots of the AUC values for each dataset at four different splits which is chosen corresponding to the time points when the 25%, 50%, 75%, and 100% of events occurred in each dataset. In other words, to show the time-dependent prediction capability of various survival analysis methods, the original regression problem has been reformulated into four classification problems which indicate whether a patient can survive at each time point or not and the prediction performance of each classifier is evaluated using AUC [25]; we exclude OLS in the plots since it is not a censored regression method. The AUC values for our proposed models are higher than or close to those of the other existing survival analysis methods indicating that the time-dependent prediction capability of our proposed models is higher than or as good as that of the other six survival prediction methods. It should be noted that the AUC values of the original RWRSS and the STC version of it are same for MCL and DLBCL datasets because the STC framework did not improve the discriminative power of RWRSS for these two datasets and hence the learning process is terminated after the second round. The AUC values of our proposed model on five datasets (NSBCD, VDV, Lung, MCL, and DBCD) are higher than or around 0.8, which indicates our proposed model is able to predict the binary problem (survival or death) at different time points effectively. This is efficient because there is no need to re-train new model for estimating whether a patient has survived or not at various time points.



(a) NSBCD

(b) VDV

(c) Lung

(d) MCL

(e) DBCD

(f) DLBCL

Figure 4: AUC values for different survival regression methods at different points of survival time. For each plot, T_1, T_2, T_3, and T_4 are the time points corresponding to the 25%, 50%, 75%, and 100% of events occurred, respectively.

Figure 5 presents the C-index values of the proposed RWRSS by varying the parameter τ from 1 to 3. We can see that the C-index values of all six datasets do not vary much when τ is greater than 1.6, and this phenomenon demonstrates that the RWRSS is not sensitive to the choice of τ that is chosen for parameter selection. Note that the RWRSS does not reduce to the standard OLS model when τ equals to 1 because the weight of some censored instances is 0 according to Eq.(3.3).

5 Conclusion

In this paper, we developed a novel regularized weighted linear regression method for high-dimensional (right)

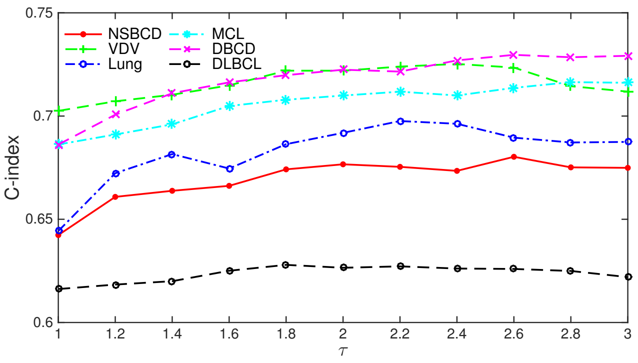


Figure 5: The effect of τ in the RWRSS alorithm on C-index.

censored data. Using the notion that the latent survival time of censored instances should be no earlier than censored time, we proposed a weighted scheme which induces more penalty for the incorrectly predicted censored instances. The elastic net penalty is used to induce sparseness into the resulting coefficients thus avoiding over-fitting the data especially in high-dimensional datasets. In addition, we also developed a self-training framework for censored regression based on this linear model. We compared the performance of the proposed methods with several state-of-the-art censored regression methods using various publicly available benchmark datasets which contain microarray gene expressions for the diseased patients. We plan to extend this work using other semi-supervised learning approaches such as label propagation in the context of survival analysis.

Acknowledgments This work was supported in part by NSF grants IIS-1527827 and IIS-1231742.

References

[1] E. T. Lee and J. Wang, *Statistical methods for survival data analysis*. Wiley. com, 2003, vol. 476.

[2] J. Buckley and I. James, "Linear regression with censored data," *Biometrika*, vol. 66, no. 3, pp. 429–436, 1979.

[3] S. Wang, B. Nan, J. Zhu, and D. G. Beer, "Doubly penalized buckley–james method for survival data with high-dimensional covariates," *Biometrics*, vol. 64, no. 1, pp. 132–140, 2008.

[4] Z. Wang and C. Wang, "Buckley-james boosting for survival analysis with high-dimensional biomarker data," *Statistical Applications in Genetics and Molecular Biology*, vol. 9, no. 1, 2010.

[5] J. Tobin, "Estimation of relationships for limited dependent variables," *Econometrica: journal of the Econometric Society*, pp. 24–36, 1958.

[6] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American statistical association*, vol. 53, no. 282, pp. 457–481, 1958.

[7] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, "Evaluating the yield of medical tests," *Jama*, vol. 247, no. 18, pp. 2543–2546, 1982.

[8] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: transfer learning from unlabeled data," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 759–766.

[9] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 187–220, 1972.

[10] R. Tibshirani *et al.*, "The lasso method for variable selection in the cox model," *Statistics in medicine*, vol. 16, no. 4, pp. 385–395, 1997.

[11] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for coxs proportional hazards model via coordinate descent," *Journal of statistical software*, vol. 39, no. 5, pp. 1–13, 2011.

[12] B. Vinzamuri and C. K. Reddy, "Cox regression with correlation based regularization for electronic health records," *IEEE 13th International Conference on Data Mining (ICDM)*, pp. 757–766, 2013.

[13] B. Vinzamuri, Y. Li, and C. K. Reddy, "Active learning based survival regression for censored data," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 2014, pp. 241–250.

[14] R. G. Miller, "Least squares regression with censored data," *Biometrika*, vol. 63, no. 3, pp. 449–464, 1976.

[15] H. Koul, V. Susarla, and J. Van Ryzin, "Regression analysis with randomly right-censored data," *The Annals of Statistics*, pp. 1276–1288, 1981.

[16] R. Miller and J. Halpern, "Regression with censored data," *Biometrika*, vol. 69, no. 3, pp. 521–531, 1982.

[17] L. Wei, "The accelerated failure time model: a useful alternative to the cox regression model in survival analysis," *Statistics in medicine*, vol. 11, no. 14-15, pp. 1871–1879, 1992.

[18] C. K. Reddy and Y. Li, "A review of clinical prediction models," in *Healthcare Data Analytics*, C. K. Reddy and C. C. Aggarwal, Eds. Chapman and Hall/CRC Press, 2015.

[19] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[20] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, vol. 33, no. 1, p. 1, 2010.

[21] T. Therneau, "A package for survival analysis in s. r package version 2.37-4," *URL http://CRAN. R-project. org/package= survival. Box*, vol. 980032, pp. 23 298–0032, 2013.

[22] B. Efron, "The efficiency of cox's likelihood function for censored data," *Journal of the American statistical Association*, vol. 72, no. 359, pp. 557–565, 1977.

[23] Y. Yang and H. Zou, "A cocktail algorithm for solving the elastic net penalized cox's regression in high dimensions," *Statistics and its Interface*, vol. 6, no. 2, pp. 167–173, 2012.

[24] A. Mayr and M. Schmid, "Boosting the concordance index for survival data–a unified framework to derive and evaluate biomarker combinations," *PloS one*, vol. 9, no. 1, p. e84483, 2014.

[25] L. E. Chambless and G. Diao, "Estimation of time-dependent area under the roc curve for long-term risk prediction," *Statistics in medicine*, vol. 25, no. 20, pp. 3474–3486, 2006.