2016 IEEE 16th International Conference on Data Mining

Transfer Learning for Survival Analysis via Efficient L2,1-norm Regularized Cox Regression

Yan Li*, Lu Wang†, Jie Wang*, Jieping Ye*‡ and Chandan K. Reddy§
*Dept. of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI -48109.
Email: {yanliwl, jwangumi, jpye}@umich.edu
†Dept. of Computer Science, Wayne State University, Detroit, MI - 48202. Email: lu.wang3@wayne.edu
‡Dept. of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI -48109.
§Dept. of Computer Science, Virginia Tech, Arlington, VA - 22203. Email: reddy@cs.vt.edu

***Abstract*—In survival analysis, the primary goal is to monitor several entities and model the occurrence of a particular event of interest. In such applications, it is quite often the case that the event of interest may not always be observed during the study period and this gives rise to the problem of censoring which cannot be easily handled in the standard regression approaches. In addition, obtaining sufficient labeled training instances for learning a robust prediction model is a very time consuming process and can be extremely difficult in practice. In this paper, we propose a transfer learning based Cox method, called *Transfer-Cox*, which uses auxiliary data to augment learning when there are insufficient amount of training examples. The proposed method aims to extract "useful" knowledge from the source domain and transfer it to the target domain, thus potentially improving the prediction performance in such time-to-event data. The proposed method uses the $l_{2,1}$-norm penalty to encourage multiple predictors to share similar sparsity patterns, thus learns a shared representation across source and target domains, potentially improving the model performance on the target task. To speedup the computation, we apply the screening approach and extend the strong rule to sparse survival analysis models in multiple high-dimensional censored datasets. We demonstrate the performance of the proposed transfer learning method using several synthetic and high-dimensional microarray gene expression benchmark datasets and compare with other related competing state-of-the-art methods. Our results show that the proposed screening approach significantly improves the computational efficiency of the proposed algorithm without compromising the prediction performance. We also demonstrate the scalability of the proposed approach and show that the time taken to obtain the results is linear with respect to both the number of instances and features.**

***Keywords*—Transfer learning; survival analysis; regularization; regression; high-dimensional data.**

I. Introduction

Due to the emergence of a wide range of data acquisition technologies, it has become a common practice in many domains to monitor subjects over a period of time in order to tell if there are any interesting events (such as device failure, disease occurrence, project success [1], etc.) that occur. Such monitoring typically starts from a particular time point and lasts until a certain event of interest occurs [2]. Due to time limitations or loss of data traces, however, the event of interest may not always be observed during the study period. This phenomenon is known as censoring and makes this problem more challenging for standard regression methods. For the instances where the event of interest is observed, the time to the event of interest is known as the failure time (or event time); while for the remaining (censored) instances, the last observed time is known as the censored time.

Survival analysis is an important branch of statistics which aims at predicting the time to the event of interest, and it can simultaneously model event data and censored data. Collecting labeling information of such problems is very time consuming, i.e., one has to wait for the occurrence of the event of interest from sufficient number of training instances to build robust models. Moreover, in many practical applications, appropriate feature collection can also be extremely expensive and tedious. A naive solution for this insufficient data problem is to merely integrate the data from related tasks into a consolidated form and build prediction models on such integrated data. However, such approaches often show poor performance since the target task (where the prediction needs to be done) will be overwhelmed by auxiliary data with different distributions. In such scenarios, knowledge transfer between related tasks will usually produce much better results compared to a mere integration scheme. Transfer learning methods have been extensively studied to solve classification and standard regression problems. However, transfer learning for survival analysis has not been studied in the literature so far, in spite of the clear practical need for this problem. In this paper, we employ the Cox proportional hazards model, one of the most popular survival analysis methods, for modeling time-to-event data.

The main objective of this paper is to improve the prediction performance of the Cox model in the target domain through knowledge transfer from the source domain in the context of survival models built on multiple high-dimensional datasets. The key component of our transfer learning method called *Transfer-Cox* is to identify the "useful" knowledge that can potentially improve the performance on the target data and transfer knowledge into the model to be learned on the target domain. Specifically, we propose to employ the $l_{2,1}$-norm to penalize the sum of the loss functions (Cox proportional hazards model) for both source and target domains [3], [4]. The $l_{2,1}$-norm penalty encourages multiple predictors to share similar sparsity patterns; thus, it will not only select important features but also learn a shared representation across source and target domains to improve the model performance on the target task. The proposed transfer learning formulation is solved via the fast iterative shrinkage thresholding algorithm (FISTA) [5]. In addition, with the help of a risk set updating method [6], the proposed *Transfer-Cox* algorithm achieves a linear time complexity with respect to both training sample size and feature dimensionality.

We demonstrate the prediction performance of our *Transfer-Cox* model using real-world high-dimensional microarray gene expression datasets which include patients from

2374-8486/16 $31.00 © 2016 IEEE
DOI 10.1109/ICDM.2016.129

TABLE I: Relationship between the proposed model and traditional multi-task learning related inductive transfer learning methods

Tasks	Source Domain Labels	Target Domain Labels	Related Literatures
Classification	Categorical / Fully informative	Categorical / Fully informative	[7],[8],[9]
Regression	Numeric / Fully informative	Numeric / Fully informative	[3],[10],[11]
Survival analysis	Numeric / Partially informative	Numeric / Partially informative	This paper

various cancer types. Our results demonstrate the power of the proposed *Transfer-Cox* model in transferring knowledge from the related cancer types to improve the survival prediction for one particular cancer type. Although the proposed algorithm is efficient, it is still time consuming due to the high dimensionality of the dataset ($19,171$ features). To this end, we adapt the idea of *screening* so that it is applicable to censored data by utilizing the strong rules [12]. Screening is a state-of-the-art technology which is able to efficiently identify the number of features whose corresponding coefficients are guaranteed to be zero. Removal of these features will dramatically reduce the dimensionality of the feature space. In this paper, we extend the strong rule to sparse survival analysis models with multiple datasets and significantly increase the efficiency of the algorithm without compromising the prediction performance.

The main contributions of our work are summarized as follows:

- Propose a novel transfer learning method *Transfer-Cox* for survival analysis which can select a subset of joint features to transfer the knowledge from the source domain to the target domain in the presence of censored data.
- Develop screening mechanism for censored data by extending the strong rule to sparse survival models with multiple datasets and use it to improve the efficiency of the algorithm without compromising the prediction performance.
- Demonstrate the performance of the proposed transfer learning method using several synthetic and high-dimensional microarray gene expression benchmark datasets, and compare it with state-of-the-art survival analysis methods.

The rest of this paper is organized as follows: Section II provides some relevant background regarding various transfer learning methods and regularized Cox regression models. Our novel transfer learning method for survival analysis, *Transfer-Cox*, is explained in detail in Section III. In Section IV, the effectiveness of the proposed *Transfer-Cox* method is demonstrated using several synthetic and high-dimensional microarray gene expression benchmark datasets. Finally, Section V concludes our discussion and gives some future research directions for the proposed work.

II. Related Work

In this section, we present the related works in the area of transfer learning and survival analysis and highlight the primary distinctions of the proposed work compared to the existing methods that are available in the literature.

A. Transfer learning

Transfer learning methods have been successfully applied in many real-world applications such as text mining [9], collaborative filtering [13] and biomedical data analysis [14] [15]. In transfer learning, the primary goal is to adapt a model built on source domain D_S (or distribution) for performing prediction on the target domain D_T. Pan *et al.* [16] categorized transfer learning methods into three different types, namely, *inductive*, *transductive* and *unsupervised* transfer learning, based on different settings for transfer. The model we propose in this paper belongs to the inductive transfer learning approach, more specifically, similar to *multi-task learning* [3]. In multi-task learning different tasks are learned simultaneously and equally weighted, while in the case of transfer learning, one set of data is selected as the target domain and the remaining is used as the source domains. Furthermore, it is very convenient to change the multi-task learning algorithm to transfer learning algorithm by merely enhancing the importance (weight) of the target task [16].

In all the methods described above and other related works (refer to [16]), the source and target tasks are either classification or regression problems. However, in this paper, the source and target tasks are the corresponding regression-based loss functions which include censored information. We propose a proportional hazards [17] based transfer learning model to transfer the useful and relevant knowledge from the source to the target domain. Our approach can effectively handle censored information based on the partial likelihood function which makes it unique compared to all the existing works. Table I summarizes the relationship between traditional multi-task learning related inductive transfer learning methods and the proposed model. It should be noted that in survival analysis, the label information is available but is partially imformative for censored instances. Hence, techniques like self-taught learning [18] and transductive learning [19] which handle scenarios with missing label information are not suitable for handling such partially informative label information.

Based on the underlying learning mechanism, transfer learning methods can be grouped into four categories [16]: *instance-based*, *feature-based*, *parameter-based*, and *relational knowledge-based*. Our proposed model is a feature-based transfer learning paradigm, which employs the $l_{2,1}$-norm [3], [4] to penalize the sum of the loss functions (Cox proportional hazards model) of both source and target tasks. The $l_{2,1}$-norm encourages multiple predictors to share similar sparsity patterns; thus, it can not only select important features and alleviate over-fitting in high-dimensional datasets but also learn a shared representation across source domain and target domain to improve the model performance on the target task.

B. Survival Analysis

Survival analysis is the field of statistics which produces optimal models for handling censored data [2]. Due to its flexibility in modeling and superiority in performance, the Cox proportional hazards model has been the most widely used model in survival analysis in the past several decades since its inception in the early 1970's [17]. It has garnered significant interest from researchers in both statistics and data mining

communities. Unlike parametric methods [20], this model does not require knowledge of the underlying distribution, but the attributes are assumed based on an exponential influence on the hazard ratio. The baseline hazard function in this model can be an arbitrary nonnegative function, but the baseline hazard functions of different individuals are assumed to be the same. The estimation and hypothesis testing of parameters in the model can be calculated by minimizing the negative log-partial likelihood function rather than the ordinary likelihood function.

Several variants of the basic Cox regression model have been proposed in the literature [21], [22]. The most interesting extension was done in terms of handling high-dimensional data, where the number of features are significantly larger than the number of samples, which typically creates the over-fitting issues with the basic Cox model. To tackle this problem, various sparsity-inducing regularization methods are widely used to penalize the negative log-partial likelihood function of Cox model. These methods include LASSO-COX [23] which employs the L_1 norm penalty, Elastic-Net Cox (EN-COX) [6] which uses the elastic net penalty term, and the kernel elastic net penalized Cox regression [21]. In the presence of limited amount of data, most of these methods produce inferior results and do not take advantage of the large amount of auxiliary data that is available. Recently, a multi-task learning formulation [24] has been proposed which reformulates the standard survival analysis as a series of related binary classification tasks. However, in this paper, the source and target tasks are both survival analysis. To the best of our knowledge, there is no work in the literature which provides transfer learning or multi-task learning for high-dimensional survival analysis.

III. Proposed Model

In this section, we will first introduce some basic concepts of survival analysis and Cox proportional hazards regression. Then we will propose the transfer learning methods based on $l_{2,1}$-norm regularized Cox model and the optimization approach.

A. Preliminaries

In survival analysis, for each data instance, we observe either a failure time (O_i) or a censored time (C_i), but not both. The dataset is said to be right-censored if and only if $y_i = \min(O_i, C_i)$ can be observed during the study. An instance in the survival data is usually represented by a triplet (X_i, T_i, δ_i), where X_i is a $1 \times p$ feature vector; δ_i is the censoring indicator, i.e. $\delta_i = 1$ for an uncensored instance, and $\delta_i = 0$ for a censored instance; and T_i denotes the *observed time* and is equal to the failure time O_i for uncensored instances and C_i otherwise, i.e.

$$T_i = \begin{cases} O_i & \text{if} \quad \delta_i = 1 \\ C_i & \text{if} \quad \delta_i = 0 \end{cases} \tag{1}$$

For censored instances, O_i is a latent value, and the goal of survival analysis is to model the relationship between X_i and O_i by using the triplets (X_i, T_i, δ_i) for censored and uncensored instances.

In survival analysis, one of the most important concepts in modeling such censored data is the *hazards function* $h_i(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq O_i < t + \Delta t | O_i \geq t)}{\Delta t}$, which is the event rate at time t conditional on survival until time t or later. In the Cox model, the proportional hazards assumption is

$$h(t, X_i) = h_0(t) \exp(X_i \beta) \tag{2}$$

for $i = 1, 2, ..., N$, where the $h_0(t)$ is the *baseline hazard function*, which can be an arbitrary non-negative function of time, and β is a $p \times 1$ regression coefficient vector of the Cox proportional hazards model. The Cox model is a semi-parametric model since all the instances share a same baseline hazard function and the coefficient estimation is independent from the form of $h_0(t)$. Let $O_1 < O_2 < \cdots < O_K$ be the increasing list of unique failure times of all N instances; given the fact that an event occurs at O_i, the conditional probability of the individual's corresponding covariate is X_i can be formulated as

$$Pr(X_i | O_i) = \frac{h(O_i, X_i)\Delta t}{\sum_{j \in R_i} h(O_i, X_j)\Delta t} = \frac{\exp(X_i \beta)}{\sum_{j \in R_i} \exp(X_j \beta)} \tag{3}$$

where R_i is the risk set at O_i which consists of all instances whose failure times are equal to or greater than O_i. Thus, the β can be learned via maximizing the partial likelihood:

$$L(\beta) = \prod_{i=1}^{K} \frac{exp(X_i \beta)}{\sum_{j \in R_i} \exp(X_j \beta)} \tag{4}$$

B. L2,1-norm regularized Cox model

In this paper, we propose a feature-based transfer learning method which aims at finding "good" features to transfer knowledge from source domian to target domain and miminize the prediction error of target task. In standard transfer learning the source and target tasks are either classification or standard regression, but in survival analysis, the source and target tasks are censored regression. Cox model is the most widely used survival analysis method, and we employ its loss function for both source and target tasks. However, Eq.(4) fails to handle the tied failures, i.e., two or more failure events that occur at same time. In this paper, the Breslow approximation [25] is used to deal with the tied failures. The partial likelihood is reformulated as follows

$$L(\beta) = \prod_{i=1}^{K} \frac{\exp(\sum_{j \in D_i} X_j \beta)}{[\sum_{j \in R_i} \exp(X_j \beta)]^{d_i}} \tag{5}$$

where D_i contains all instances whose failure time is O_i and $d_i = |D_i|$ is the size of D_i. Therefore, the coefficient vector can be learned via minimizing the negative log-partial likelihood.

$$l(\beta) = -\sum_{i=1}^{K} \left\{ \sum_{j \in D_i} X_j \beta - d_i \log[\sum_{j \in R_i} \exp(X_j \beta)] \right\} \tag{6}$$

To find "good" features for knowledge transfer, we propose a model which is able to learn a shared representation across source and target tasks. The $l_{2,1}$-norm is chosen to be one penalty term for our model because it encourages multiple coefficient vectors to share similar sparsity patterns. Therefore, the regularized model learns a shared representation across source and target tasks. In addition, a sparsity inducing penalty also helps the model deal with high-dimensional datasets and

alleviate model over-fitting. The proposed transfer learning model "*Transfer-Cox*" can be learned via solving the following minimization problem.

$$\min_{B} \sum_{t\in\{S,T\}} -\frac{w_t}{N_t} l(\beta_t) + \frac{\mu}{2} \parallel B \parallel_F^2 + \lambda \parallel B \parallel_{2,1} \quad (7)$$

where S and T denote the tasks in the source domain and target domain, respectively. $B = (\beta_S, \beta_T), B \in \mathbb{R}^{p\times 2}$, N_S and N_T are the number of training instances in the source domain and target domain, respectively. w_S and w_T are two empirically determined weight parameters, and usually $w_S < w_T$ which induces the model focusing more on the target task. The l_2 regularization on the coefficient matrix B is introduced to further reduce the variance of B and alleviate model over-fitting.

C. Optimization

The optimization problem proposed in Eq.(7) follows the standard $l_{1,2}$-norm regularization problem:

$$\min_{B\in\mathbb{R}^{p\times 2}} g(B) + \lambda \parallel B \parallel_{2,1} \quad (8)$$

where $\lambda > 0$ is the regularization parameter, and

$$g(B) = \sum_{t\in\{S,T\}} -\frac{w_t}{N_t} l(\beta_t) + \frac{\mu}{2} \parallel B \parallel_F^2$$

is a smooth convex loss function, and its first order derivative can be calculated as:

$$g^{'}(B) = \left[\frac{w_S}{N_S} l^{'}(\beta_S) + \mu\beta_S,\ \frac{w_T}{N_T} l^{'}(\beta_T) + \mu\beta_T\right] \quad (9)$$

where $l^{'}(\beta_S)$ and $l^{'}(\beta_T)$ are the gradient of the negative log-partial likelihood as shown in Eq.(6), and these two terms share the same formulation

$$l^{'}(\beta) = -\sum_{i=1}^{K}\left\{\sum_{j\in D_i} X_j - d_i \frac{\sum_{j\in R_i} X_j \exp(X_j\beta)}{\sum_{j\in R_i}\exp(X_j\beta)}\right\} \quad (10)$$

coresponding to the source and target datasets, respectively.

The optimization problem in Eq.(8) can be solved efficiently via the FISTA based algorithm (refer to the Appendix for more details) with the general updating step,

$$B^{(i+1)} = \pi_P(S^{(i)} - \frac{1}{\gamma_i} g^{'}(S^{(i)})) \quad (11)$$

where $S^{(i)} = B^{(i)} + \alpha_i(B^{(i)} - B^{(i-1)}) = [S_S^{(i)},\ S_T^{(i)}]$ are two search points of the source task and target task, respectively. α_i is the combination scaler, $g^{'}(S^{(i)})$ is the gradient of $g(\cdot)$ at point $S^{(i)}$, $\frac{1}{\gamma_i}$ is the possible biggest stepsize which is chosen by line search, and $\pi_P(\cdot)$ is the $l_{1,2}$-regularized Euclidean projection:

$$\pi_P(G(S^{(i)})) = \min \frac{1}{2} \parallel B - G(S^{(i)})) \parallel_F^2 + \lambda \parallel B \parallel_{2,1} \quad (12)$$

where $G(S^{(i)}) = S^{(i)} - \frac{1}{\gamma_i} g^{'}(S^{(i)})$. An efficient solution (Theorem 1) of Eq.(12) has been proposed in [4].

Theorem 1: Given λ, the primal optimal point $\hat{B}$ of Eq.(12) can be calculated as:

$$\hat{B}_j = \begin{cases} \left(1 - \frac{\lambda}{\|G(S^{(i)})_j\|_2}\right) G(S^{(i)})_j & \text{if } \lambda > 0, \parallel G(S^{(i)})_j \parallel_2 > \lambda \\ 0 & \text{if } \lambda > 0, \parallel G(S^{(i)})_j \parallel_2 \leq \lambda \\ G(S^{(i)})_j & \text{if } \lambda = 0 \end{cases} \quad (13)$$

where $G(S^{(i)})_j$ is the j^{th} row of $G(S^{(i)})$, and $\hat{B}_j$ is the j^{th} row of $\hat{B}$.

1) Complexity Analysis: The main cost per iteration of our optimization scheme is the computation of $g(\cdot)$ and $g^{'}(\cdot)$, more specifically, the computation of the negative log-partial likelihood and its gradient. From Eq.(6) and Eq.(10), we can see that, at each failure time point O_i, one needs to calculate $\sum_{j\in R_i} e^{X_j\beta}$ and $\sum_{j\in R_i} X_j e^{X_j\beta}$; thus, for all failure times, it needs $O(N^2 p)$ calculations, because R_i has $O(N)$ elements. To speedup the training process, we employ the risk set updating method proposed in [6] which is given as follows.

$$\sum_{j\in R_{i+1}} e^{X_j\beta} = \sum_{j\in R_i} e^{X_j\beta} - \sum_{j\in(R_i - R_{i+1})} e^{X_j\beta} \quad (14)$$
$$\sum_{j\in R_{i+1}} X_j e^{X_j\beta} = \sum_{j\in R_i} X_j e^{X_j\beta} - \sum_{j\in(R_i - R_{i+1})} X_j e^{X_j\beta}$$

Here, we only need to calculate $\sum_{j\in R_1} e^{X_j\beta}$ and $\sum_{j\in R_1} X_j e^{X_j\beta}$. Then for the subsequent failure time point O_i, we subtract the contribution from instances which are failed or censored between O_{i-1} and O_i. Therefore, the calculations of $l(\beta)$ and $l^{'}(\beta)$ are both reduced to $O(Np)$, and the computation cost of $g(\cdot)$ and $g^{'}(\cdot)$ are both $O((N_S + N_T)p)$.

In our transfer learning problem, there are only two tasks (source and target tasks), so the Euclidean projection in Eq.(12) can be efficiently calculated in $O(2p) = O(p)$. Therefore, the optimization procedure solves the optimization problem in Eq.(7) with a time complexity of $O(\frac{1}{\sqrt{\varepsilon}}(N_S + N_T)p)$ for achieving an accuracy of ε.

D. Solution Path and Strong Rule

Usually, in the learning process, the model has to be trained based on a series of values for λ, and the best λ is selected via cross-validation. In this paper, we employ the warm-start approach given in [26] to build the solution path; initialize λ to a sufficiently large number, which forces B to a zero matrix, and then gradually decreases λ in each learning iteration. For a new λ, the initial value of B is the estimated B learned from the previous λ, so the initial value of B is not far from the optimal value, and the algorithm will converge within a few iterations.

Firstly, λ_{max}, the smallest tuning parameter value which forces B to a zero matrix, needs to be calculated. From Eq.(13) we can see that if $\parallel G(S^{(0)})_j \parallel_2 < \lambda$ for all j, then $B = \mathbf{0}$ is the optimal solution. Thus, we set

$$\lambda_{max} = \max_j \parallel G(S^{(0)})_j \parallel_2 \quad (15)$$
$$= \max_j \parallel S_j^{(0)} - \frac{1}{\gamma_0} g^{'}(S^{(0)})_j \parallel_2 = \max_j \parallel g^{'}(\mathbf{0})_j \parallel_2$$

to be the first λ, where $g^{'}(\cdot)_j$ is the j^{th} row of $g^{'}(\cdot)$. If $\min(N_S, N_T) \geq p$ we set $\lambda_{min} = 0.0001\lambda_{max}$, else we set $\lambda_{min} = 0.05\lambda_{max}$. In our experiments, we search m different λ values in total, and for the k^{th} step $\lambda_k = \lambda_{max}(\lambda_{min}/\lambda_{max})^{k/m}$.

The FISTA based learning scheme is an efficient method to solve the transfer learning problem proposed in Eq.(7). However, if the feature dimensionality (p) is extremely large, the proposed optimization approach will still take substantial amount of time. *Screening* is a state-of-the-art technology which is able to efficiently identify features whose corresponding coefficients are guaranteed to be zero. Removal of these features will dramatically reduce the feature dimension; thus, screening is able to improve the efficiency of many sparse models [27]. Our optimization problem in Eq.(7) can be rewritten as

$$\min_B g(B) + \lambda \sum_{j=1}^{p} \parallel B_j \parallel_2 \tag{16}$$

where B_j stands for the j^{th} row of B. Eq.(16) belongs to the general Lasso-type problems, and based on the the Karush-Kuhn-Tucker (KKT) conditions, Tibshirani et al. proposed the strong rules for this type of problems [12]. The KKT conditions for Eq.(16) are

$$g^{'}(\hat{B})_j = \lambda \boldsymbol{\theta}_j \text{ for } j = 1, 2, \cdots, p \tag{17}$$

where $\hat{B}$ is the optimal solution and $\boldsymbol{\theta}_j$ is a subgradient of $\parallel \hat{B}_j \parallel_2$, which satisfies $\parallel \boldsymbol{\theta}_j \parallel_2 \leq 1$ and $\parallel \boldsymbol{\theta}_j \parallel_2 < 1$ implies $\hat{B}_j = 0$. Based on the above KKT conditions and [12, Section 6, page 17], for our problem the sequential strong rule for Eq.(16) to discard inactive features (corresponding coefficients are zero) is as follows.

Theorem 2: Given a sequence of parameter values $\lambda_{max} = \lambda_0 > \lambda_1 > \cdots > \lambda_m$, and suppose the optimal solution $\hat{B}(k-1)$ at λ_{k-1} is known. Then for any $k = 1, 2, \cdots, m$ the j^{th} feature will be discarded if

$$\parallel g^{'}(\hat{B}(k-1))_j \parallel_2 < 2\lambda_k - \lambda_{k-1} \tag{18}$$

and the corresponding coefficient $\hat{B}(k)_j$ will be set to 0.

However, based on the experimental analysis in [12], we know that, Theorem 2 might mistakenly discard active features (corresponding coefficients are nonzero), so we need to check KKT conditions of the discarded features. Let V^d and V^s denote the index set of discarded features and selected features, respectively. From Theorem 2, we get $\hat{B}(k)_j = 0,\ \forall j \in V^d$, and based on Eq.(13) we know that if

$$\parallel g^{'}(\hat{B}(k))_j \parallel_2 \leq \lambda_k \quad \forall j \in V^d$$

is true, then $\hat{B}(k)$ is the optimal solution at λ_k. Otherwise, V^s need to be updated via $V^s = V^s \cup V^v$ where

$$V^v = \left\{ j \middle| j \in V^d,\ \parallel g^{'}(\hat{B}(k))_j \parallel_2 > \lambda_k \right\} \tag{19}$$

is the index set of mis-discarded features.

Above all, Figure 1 summarizes our proposed model with solution path and strong rule. Firstly, λ_{max} will be calculated by Eq.(15) as the starting searching point. Next, the strong rule will be used to discard inactive features, and the model will be trained with the selected features. To prevent mis-discarding, we then have to check the KKT conditions. If there are any mis-discarded features, we have to update the set of selected features and retrain the model; if not, we can train a model based on a new λ. In order to ensure the reproducibility of our work, the codes of *Transfer-Cox* model with strong rule are made available at this github website [1].

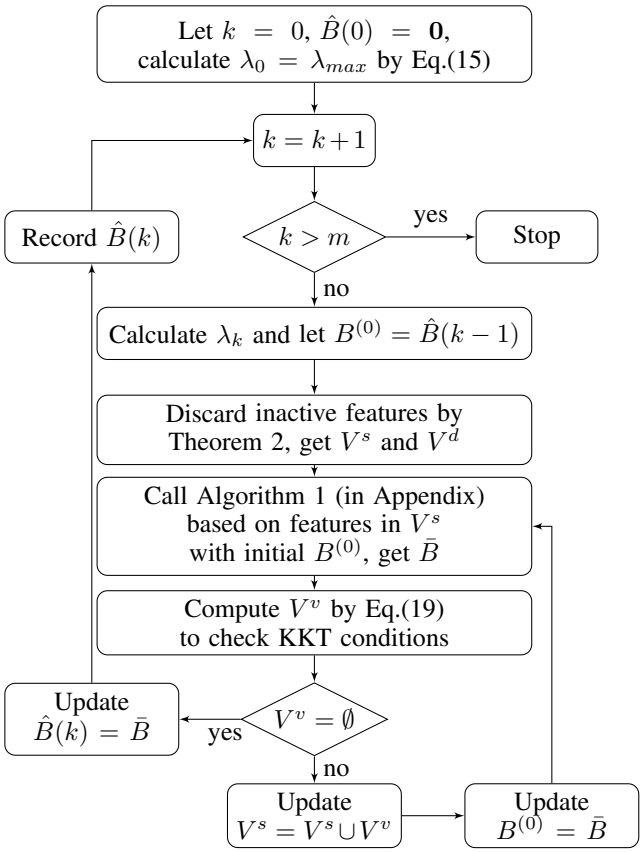


Fig. 1: Flowchart for *Transfer-Cox* algorithm with strong rule.

IV. Experimental Results

In this section, we will first describe the datasets used in our evaluation and demonstrate the prediction performance of the proposed *Transfer-Cox* model. Then we will experimentally demonstrate the efficiency of the screening methods and the scalability of the proposed algorithm. Finally, we perform a detailed study on the biomarkers selected by the proposed algorithm on different cancer types and show their biological significance as well.

A. Dataset Description

For our model evaluation, we use publicly available high-dimensional gene expression cancer survival benchmark datasets [2] from The Cancer Genome Atlas (TCGA) [28]. In this paper, we perform knowledge transfer in survival analysis by analyzing the microarray gene expressions for different cancer types. We have labeled data in all cancer types. The dataset

[1] https://github.com/MLSurvival/TransferCox

[2] Downloaded from https://cran.r-project.org/web/packages/dnet/index.html

contains somatic mutational profiles for $3,096$ cancer patients with survival information, and for each patient the relative activity of $19,171$ genes are measured. These gene values are considered to be the features in our data. The cancer patients belong to one of the 12 major cancer types: bladder urothelial carcinoma (BLCA), breast adenocarcinoma (BRCA), colon carcinoma (COAD), rectal carcinoma (READ), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), acute myeloid leukaemia (LAML), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian serous carcinoma (OV) and uterine corpus endometrial carcinoma (UCEC).

TABLE II: Basic statistics of the selected 8 cancer types.

Data	# **Instances**	# **Uncensored**	# **Censored**
BRCA	763	90	673
GBM	275	176	99
HNSC	300	119	181
KIRC	417	136	281
LAML	185	117	68
LUAD	155	50	105
LUSC	171	68	103
OV	315	181	134

In our experiments, the goal is to improve the performance of survival prediction for a particular cancer type. Hence, in our transfer learning setting, this specific cancer type is considered to be the target domain and the data from remaining types is considered to be the source domain. Table II shows the basic statistics of the cancer types considered for our analysis. The number of uncensored instances in four cancer types are too small and hence these four cancer types were eliminated for our evaluation. In our experiment, for the remaining 8 cancer types, each of them will be considered as the target domain and the data from the remaining cancer types will be considered as the source domain. In this table, the columns titled "# Uncensored" and "# Censored" correspond to the number of uncensored and censored instances in each cancer type, respectively. For these cancer types, the event of interest is patient death; therefore, an uncensored instance refers to the patient being dead during the study, while a censored instance refers to the corresponding patient is still alive at the last observed time (which will be the censored time).

B. Performance Comparison

To the best of our knowledge, neither transfer learning nor multi-task learning for survival analysis has been studied in the literature. Hence, we can only compare our proposed *Transfer-Cox* with standard related survival analysis methods. As our *Transfer-Cox* is a Cox-based model and $l_{2,1}$-norm is a Lasso-type penalty, we choose Cox model and two other popular regularized Cox models: LASSO-COX and EN-COX as comparison methods. In our experiments, these three survival analysis methods are applied both on the target dataset (the specified cancer type) and the entire dataset. For simplicity, they are referred to as "Local" models and "Global" models, respectively. It should be noted that, in "Global" models, although each model is built on the entire dataset, the performance is measured only on the target dataset. For a "Local" model, the training and testing are performed only on the target cancer type (using cross validation). For a global model, the training is done on the source+target samples and the testing is done on the target cancer samples.

The concordance index (C-index), or *concordance probability*, is used to measure the performance of prediction models in survival analysis [29]. 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

$$c = Pr(\hat{y}_1 > \hat{y}_2 | y_1 \geq y_2). \tag{20}$$

By definition, the C-index has the same scale as the area under the ROC curve (AUC) in binary classification, and if y_i is binary, then the C-index is same as the AUC. In the standard Cox and regularized Cox models, the hazard ratio is modeled to describe the time-to-event data. The instances with a low hazard rate should survive longer, so the C-index is calculated as follows:

$$c = \frac{1}{num} \sum_{i \in \{1 \cdots N | \delta_i = 1\}} \sum_{y_j > y_i} I[X_i \hat{\beta} > X_j \hat{\beta}] \tag{21}$$

where num denotes the number of comparable pairs and $I[\cdot]$ is the indicator function.

In Table III, we show the performance results of C-index values of different algorithms using 5-fold cross validation. The best results are highlighted in bold. The results show that our proposed *Transfer-Cox* model outperforms the other state-of-the-art models. We also notice that for 7 out of the 8 cancer types, the "Global" Cox model performs better than the "Local" Cox model, which indicates that having more samples from other cancer types will help in generalization and alleviate over-fitting. However, for 4 cancer types in the "Local" regularized Cox models perform better than the "Global" regularized Cox models; this phenomenon reflects that, from the genomics perspective, the preventable factors and reflections of different cancer types are clearly different.

C. Empirical Analysis of Efficiency

In this section, we will demonstrate the efficiency of the strong rule and also show the scalability performance of the proposed *Transfer-Cox* algorithm.

1) Efficiency of strong rule: To measure the efficiency of applying strong rule, we measure the *rejection ratio* and *screen ratio* which are defined as follows:

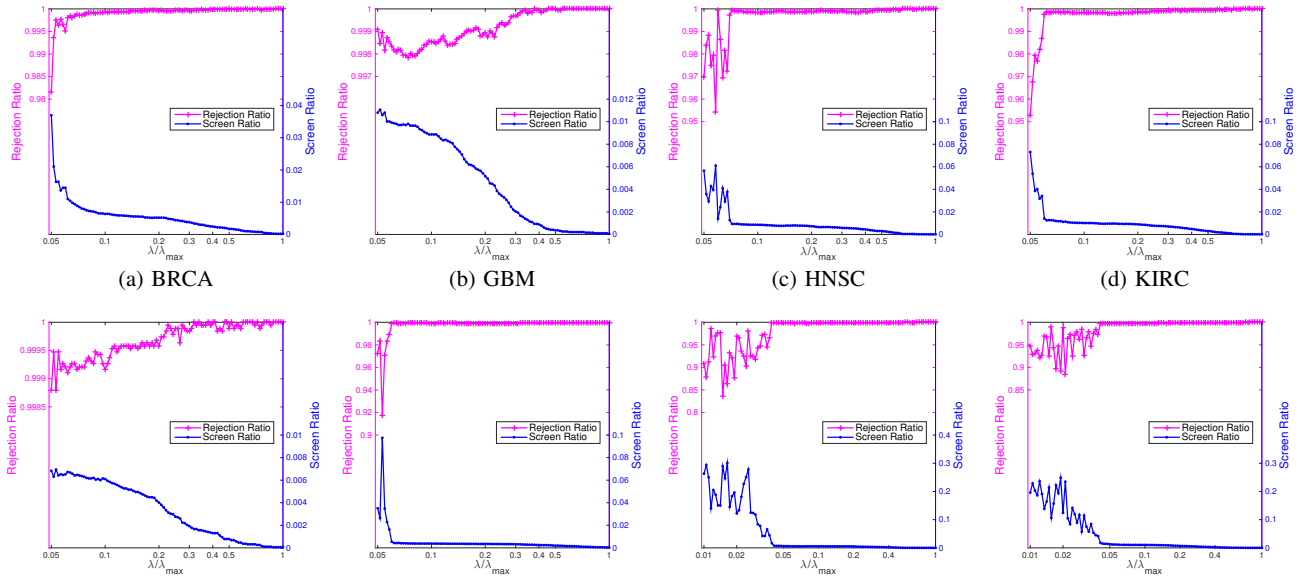
$$\text{rejection ratio} = \frac{\text{number of identified inactive features}}{\text{number of true inactive features}}$$

$$\text{screen ratio} = \frac{\text{number of selected features}}{\text{original feature dimension}}$$

Figure 2 shows the *rejection ratio* and *screen ratio* of the strong rule on gene expression data of 8 cancer types. In Figure 2(a)–(f) the λ_{min} is set equal to $0.05\lambda_{max}$, as mentioned in Section III-D. However, under this setting, less than one hundred features will be selected as active features in "LUSC" and "OV", so we set $\lambda_{min} = 0.01\lambda_{max}$ and draw the screening ratio in Figure 2(g) and Figure 2(h), for these two cancer types. All eight plots in Figure 2 reflect that the strong rule in the proposed *Transfer-Cox* model can successfully identify a majority of the inactive features (high rejection ratio) and dramatically decrease the feature dimensionality

TABLE III: Performance comparison of the proposed *Transfer-Cox* method and other existing related methods using C-index values (along with their standard deviations).

Dataset	Local			Global			*Transfer-Cox*
	COX	LASSO-COX	EN-COX	COX	LASSO-COX	EN-COX	
BRCA	0.4348 (0.0756)	0.3868 (0.0418)	0.4055 (0.0426)	0.5547 (0.0238)	0.5822 (0.0394)	0.5811 (0.0411)	**0.5869 (0.0456)**
GBM	0.5064 (0.0677)	0.5741 (0.0181)	0.5613 (0.0260)	0.5592 (0.0243)	0.5841 (0.0131)	0.5842 (0.0130)	**0.6136 (0.0300)**
HNSC	0.5663 (0.0759)	0.5591 (0.0527)	0.5788 (0.0491)	0.5794 (0.0059)	0.5528 (0.0776)	0.5542 (0.0789)	**0.6157 (0.0379)**
KIRC	0.5689 (0.0322)	0.6001 (0.0206)	0.6061 (0.0216)	0.5553 (0.0603)	0.5903 (0.0401)	0.5908 (0.0386)	**0.6255 (0.0393)**
LAML	0.5599 (0.0887)	0.6861 (0.0189)	0.6838 (0.0227)	0.6057 (0.0397)	0.6591 (0.0103)	0.6580 (0.0141)	**0.6939 (0.0305)**
LUAD	0.3832 (0.1371)	0.5327 (0.0840)	0.5435 (0.0337)	0.4463 (0.0443)	0.5354 (0.1026)	0.5378 (0.1040)	**0.5877 (0.0409)**
LUSC	0.5250 (0.0719)	0.4670 (0.1009)	0.4861 (0.0598)	0.5520 (0.0426)	0.5798 (0.0465)	0.5770 (0.0584)	**0.5905 (0.0374)**
OV	0.5132 (0.0260)	0.4991 (0.1043)	0.4971 (0.0911)	0.5438 (0.0826)	0.5708 (0.0850)	0.5697 (0.0825)	**0.6167 (0.0342)**



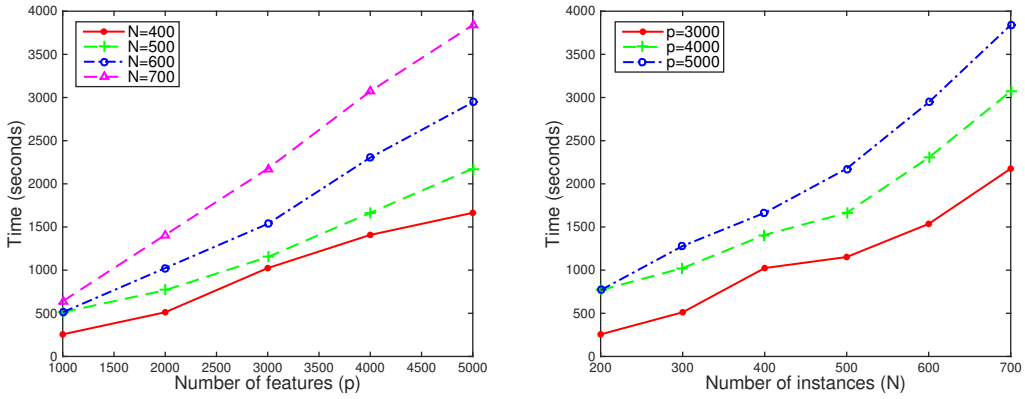
(a) BRCA (b) GBM (c) HNSC (d) KIRC

(e) LAML (f) LUAD (g) LUSC (h) OV

Fig. 2: *Efficiency of strong rule*: plots of the rejection ratio and screen ratio on gene expression data for 8 cancer types.

(low screen ratio) in the learning phase. Table IV presents the running time of the *Transfer-Cox* with and without strong rule and the *speedup* achieved. All the timing calculations are based on running the experiments on an Intel Xeon 3 GHz processor with 12 cores (24 threads). *Speedup* is the ratio of the running time of *Transfer-Cox* without screening to it's running time with screening. We only show the result on one cancer type "LUAD" because without the screening procedure, the computation of *Transfer-Cox* takes very long (more than one day). In addition to this cancer data, we also generated two synthetic datasets "Syn1" and "Syn2" using the the function "simple.surv.sim" in *survsim* package [30]. These two datasets have 500 instances in source domian, 100 instances in the target domian, and a maximum follow-up time of $1,000$ days. All the features are generated based on the uniform distribution, and each of them have a different random setted value interval. The coefficient vector is also randomly generated and remain in $[-1, 1]$. The observed time is assumed to follow a Log-logistic distribution and time to censorship follows a Weibull distribution. The datasets in scalability analysis are also generated in the same manner with different sample size and feature dimension. The results show that the screening method can dramatically speed up the algorithm and become more effective as the feature dimension increases (see Table IV).

2) Scalability of Transfer-Cox: We empirically evaluate the scalability of the proposed *Transfer-Cox* model with respect to the sample size ($N = N_S + N_T$) and the number of features



(a) Scalability w.r.t. p (b) Scalability w.r.t. N

Fig. 3: *Scalability resuts*: Plots of the runtimes for the *Transfer-Cox* model. The times denote total runtimes for 100 λ values averaged over five trials.

TABLE IV: Running time comparison for the *Transfer-Cox* model with and without screening rule for 100 λ values with default setting ($\lambda_{min} = 0.05\lambda_{max}$).

Data	p	**With screening**	**Without screening**	**speedup**
Syn 1	5,000	768 (s)	2944 (s)	3.83
Syn 2	10,000	1286 (s)	6084 (s)	4.73
LUAD	19,171	3.59 (hrs)	35.13 (hrs)	9.78

(p). In this experiment, we did not use strong rule as it will influence the the scalability analysis with respect to p. This is because the strong rule will discard features and make p unstable with different λ values, which is clearly shown in Figure 2. Figure 3(a) shows runtimes for fixed N and varying p, and Figure 3(b) shows runtimes for fixed p and varying N. These two plots suggest that the runtime of *Transfer-Cox* is close to being *linear with respect to both N and p.*

D. Biomarker Discovery

Biomarkers are important indicators used to diagnose a particular disease in a clinical setting. From a clinical perspective, it is known that different cancer types should share some common significant biomarkers such as one particular anticarcinogen, chemotherapy dose, and radiotherapy dose. However, at the genetic level, the preventable factors and reflections of different cancer types are clearly different. In Table V, we show a list of top 10 gene expression features for each cancer type based on their contributions (coefficient weights) in the *Transfer-Cox* model and find that most of the top-ranked gene expression features are usually related to the genetics of the corresponding cancer types. For example, in BRCA, CD6 is heterotypic adhesion with activated leukocyte cell adhesion molecule which is in breast cancer lines acting in melanoma tumor progression and resected breast tumors [31]. In GBM, AK5 affects the cyclophilin B depletion on GBM cell line[3]. In HNSC, the content of MRPL48 is one of the high expression genes to influence the HNSC as one essential mitochondrial ribosomal protein[4]. ITGB4 is the high expression of cell adhesion models as heterogeneous immunohistochemical feature in KIRC [32]. The lack of GSTM1 will increase the risk of LAML when GSTM1 gene is null genotype among LAML patients [33]. For LUSC, APOA1BP is in human serum, cerebrospinal and spinal fluid to help body's transport and metabolism related to lung cancer cell lines reported in human body fluids in pathological conditions [34]. For LUAD, MIR655 is one of the discovered class of small RNS which is linked to the development and progression of cancer in lung [35]. For OV, ST14 is in the papillary serous subtype of ovarian tumors reported by cytogenetic analysis of primary ovarian carcinomas and ovarian cancer cell lines [36]. It should be noted that some of the remaining features listed can be strong potential candidates for further biological testing for generating new hypotheses in the future. From this analysis, it is clear that the proposed *Transfer-Cox* algorithm not only provides better results in an efficient manner, but also inherently provides insights about the critical features for further analysis.

[3]http://www.ncbi.nlm.nih.gov/geoprofiles/104754733

[4]http://amp.pharm.mssm.edu/Harmonizome/gene/MRPL48

V. Conclusion

In this paper, we developed a novel transfer learning model for survival analysis. The proposed *Transfer-Cox* is a regularized Cox regression model that is able to efficiently select common hidden features in high-dimensional (right) censored data to transfer knowledge from the source domain to the target domain. The $l_{2,1}$-norm penalty is used to induce common sparseness into both source and target domains thus learning a shared low-dimensional feature representation for knowledge transfer and alleviating over-fitting the data, especially in high-dimensional scenarios. In order to speedup the learning scheme, we use the idea of screening and extend the strong rule to the proposed *Transfer-Cox* model. Thus, our model is able to efficiently identify most of the inactive features, and the computational cost of learning *Transfer-Cox* is dramatically reduced by the removal of the inactive features in the training phase. We compared the performance of the proposed *Transfer-Cox* algorithm with several state-of-the-art censored regression methods using publicly available high-dimensional microarray gene expression data from different cancer types. We also demonstrated the efficiency of the strong rule and showed linear scalability of the proposed model with respect to the

TABLE V: Top 10 gene expression features obtained for each cancer type using *Transfer-Cox* model.

Cancer Type: BRCA		**Cancer Type**: GBM	
Symbol	Description	Symbol	Description
SSBP4	single stranded DNA binding protein 4	CALR3	calreticulin 3
METTL22	methyltransferase like 22	AK5	adenylate kinase 5
SGCD	sarcoglycan (35kDa dystrophin-associated glycoprotein)	PSMF1	proteasome inhibitor subunit 1 (PI31)
LYPD6B	LY6/PLAUR domain containing 6B	SLC31A2	solute carrier family 31 (copper transporter), #2
FCGR1B	Fc fragment of IgG, high affinity Ib, receptor (CD64)	PPP1R12C	protein phosphatase 1, regulatory subunit 12C
CXCL14	chemokine (C-X-C motif) ligand 14	CHORDC1	cysteine and histidine-rich domain containing 1
RAB27A	RAB27A, member RAS oncogene family	PROX2	prospero homeobox 2
CD6	CD6 molecule	OR1A1	olfactory receptor, family 1, subfamily A, member 1
GOLGA8DP	golgin A8 family, member D, pseudogene	SRRM4	serine/arginine repetitive matrix 4
PTGR2	prostaglandin reductase 2	IL32	interleukin 32
Cancer Type: HNSC		**Cancer Type**: KIRC	
Symbol	Description	Symbol	Description
MRPL48	mitochondrial ribosomal protein L48	ITGB4	integrin, beta 4
MAP2K1	mitogen-activated protein kinase kinase 1	CCT8L2	chaperonin containing TCP1, subunit 8 (theta)-like 2
NSUN4	NOP2/Sun domain family, member 4	TP73-AS1	TP73 antisense RNA 1
OVOL1	ovo-like zinc finger 1	PGM1	phosphoglucomutase 1
SSX9	synovial sarcoma, X breakpoint 9	MEPE	matrix extracellular phosphoglycoprotein
INSIG1	insulin induced gene 1	CXorf40B	chromosome X open reading frame 40B
SP9	Sp9 transcription factor	BANK1	B-cell scaffold protein with ankyrin repeats 1
DDIT4	DNA-damage-inducible transcript 4	C10orf120	chromosome 10 open reading frame 120
EPN2	epsin 2	PRDX3	peroxiredoxin 3
EWSR1	EWS RNA-binding protein 1	C10orf91	chromosome 10 open reading frame 91
Cancer Type: LAML		**Cancer Type**: LUSC	
Symbol	Description	Symbol	Description
NPIPB15	nuclear pore complex interacting protein family, #B15	C10orf76	chromosome 10 open reading frame 76
LOC285696	uncharacterized LOC285696	APOA1BP	apolipoprotein A-I binding protein
GSTM1	glutathione S-transferase mu 1	MIR1267	microRNA 1267
IGKV2-24	immunoglobulin kappa variable 2-24	MIR509-2	microRNA 509-2
C2orf83	chromosome 2 open reading frame 83	IGLC7	immunoglobulin lambda constant 7
MIR1255A	microRNA 1255a	MIR1251	microRNA 1251
MTFR1L	mitochondrial fission regulator 1-like	P3H2	prolyl 3-hydroxylase 2
CCKBR	cholecystokinin B receptor	IMP4	IMP4, U3 small nucleolar ribonucleoprotein
SULT4A1	sulfotransferase family 4A, member 1	MAPKAPK3	mitogen-activated protein kinase-activated protein kinase 3
LINC00313	long intergenic non-protein coding RNA 313	CNTROB	centrobin, centrosomal BRCA2 interacting protein
Cancer Type: LUAD		**Cancer Type**: OV	
Symbol	Description	Symbol	Description
MIR655	microRNA 655	NADSYN1	NAD synthetase 1
GATC	glutamyl-tRNA(Gln) amidotransferase, subunit C	DNMT1	DNA (cytosine-5-)-methyltransferase 1
ZNF419	zinc finger protein 419	VMO1	vitelline membrane outer layer 1 homolog (chicken)
TRIM34	tripartite motif containing 34	OR2A1	olfactory receptor, family 2, subfamily A, member 1
PAK1IP1	PAK1 interacting protein 1	ST14	suppression of tumorigenicity 14 (colon carcinoma)
C11orf72	chromosome 11 open reading frame 72	FCRL4	Fc receptor-like 4
KLRC2	killer cell lectin-like receptor subfamily C, member 2	LRRTM1	leucine rich repeat transmembrane neuronal 1
WNT7A	wingless-type MMTV integration site family, # 7A	RFX7	regulatory factor X, 7
MST1P2	macrophage stimulating 1 (hepatocyte growth factor-like) pseudogene 2	SPRED1	sprouty-related, EVH1 domain containing 1
AP2S1	adaptor-related protein complex 2, sigma 1 subunit	BET1L	Bet1 golgi vesicular membrane trafficking protein-like

number of samples and the number of features. In the future, we will design new safe screening method for *Transfer-Cox* which does not need to check the KKT conditions. We also plan to develop other instance-based and feature-based transfer learning methods for survival analysis.

ACKNOWLEDGMENTS

This work was supported in part by the US National Science Foundation grants IIS-1231742, IIS-1527827, IIS-1646881, III-1539991, and III-1539722.

REFERENCES

[1] Y. Li, V. Rakesh, and C. K. Reddy, "Project success prediction in crowdfunding environments," in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, ser. WSDM '16, 2016, pp. 247–256. [Online]. Available: http://doi.acm.org/10.1145/2835776.2835791

[2] E. T. Lee and J. Wang, *Statistical methods for survival data analysis*. John Wiley & Sons, 2003, vol. 476.

[3] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Machine Learning*, vol. 73, no. 3, pp. 243–272, 2008.

[4] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient l2, 1-norm minimization," in *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. AUAI Press, 2009, pp. 339–348.

[5] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[6] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for cox's proportional hazards model via coordinate descent," *Journal of statistical software*, vol. 39, no. 5, pp. 1–13, 2011.

[7] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 193–200.

[8] T. Jebara, "Multi-task feature and kernel selection for SVMs," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 55.

[9] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *Neural Networks, IEEE Transactions on*, vol. 22, no. 2, pp. 199–210, 2011.

[10] D. Pardoe and P. Stone, "Boosting for regression transfer," in *Proceed-*

ings of the 27th international conference on Machine learning (ICML-10), 2010, pp. 863–870.

[11] J. Garcke and T. Vanck, "Importance weighted inductive transfer learning for regression," in *Machine Learning and Knowledge Discovery in Databases*. Springer, 2014, pp. 466–481.

[12] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. J. Tibshirani, "Strong rules for discarding predictors in lasso-type problems," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 74, no. 2, pp. 245–266, 2012.

[13] W. Pan, E. W. Xiang, N. N. Liu, and Q. Yang, "Transfer learning in collaborative filtering for sparsity reduction," in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI*, 2010.

[14] Y. Li, B. Vinzamuri, and C. K. Reddy, "Constrained elastic net based knowledge transfer for healthcare information exchange," *Data Mining and Knowledge Discovery*, vol. 29, no. 4, pp. 1094–1112, 2015.

[15] S. Al-Stouhi and C. K. Reddy, "Transfer learning for class imbalance problems with inadequate data," *Knowledge and Information Systems*, vol. 48, no. 1, pp. 201–228, 2016. [Online]. Available: http://dx.doi.org/10.1007/s10115-015-0870-3

[16] S. J. Pan and Q. Yang, "A survey on transfer learning," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 22, no. 10, pp. 1345–1359, 2010.

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

[18] 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*, 2007, pp. 759–766.

[19] A. Arnold, R. Nallapati, and W. W. Cohen, "A comparative study of methods for transductive transfer learning," in *Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on*. IEEE, 2007, pp. 77–82.

[20] Y. Li, K. S. Xu, and C. K. Reddy, "Regularized parametric regression for high-dimensional survival analysis," in *Proceedings of SIAM International Conference on Data Mining*, 2016, pp. 765–773.

[21] 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.

[22] 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.

[23] 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.

[24] Y. Li, J. Wang, J. Ye, and C. K. Reddy, "A multi-task learning formulation for survival analysis," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. ACM, 2016, pp. 1715–1724. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939857

[25] N. E. Breslow, "Contribution to the discussion of the paper by DR cox," *Journal of the Royal Statistical Society, Series B*, vol. 34, no. 2, pp. 216–217, 1972.

[26] 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.

[27] J. Wang, J. Zhou, P. Wonka, and J. Ye, "Lasso screening rules via dual polytope projection," in *Advances in Neural Information Processing Systems*, 2013, pp. 1070–1078.

[28] C. Kandoth, M. D. McLellan, F. Vandin, K. Ye, B. Niu, C. Lu, M. Xie, Q. Zhang, J. F. McMichael, M. A. Wyczalkowski *et al.*, "Mutational landscape and significance across 12 major cancer types," *Nature*, vol. 502, no. 7471, pp. 333–339, 2013.

[29] 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.

[30] D. Morina and A. Navarro, "The r package survsim for the simulation of simple and complex survival data," *Journal of Statistical Software*, vol. 59, no. 2, pp. 1–20, 2014.

[31] J. A. King, S. F. Ofori-Acquah, T. Stevens, A.-B. Al-Mehdi, O. Fodstad, and W. G. Jiang, "Activated leukocyte cell adhesion molecule in breast cancer: prognostic indicator," *Breast Cancer Res*, vol. 6, no. 5, pp. R478–R487, 2004.

[32] D. Andreadis, A. Nomikos, and C. Barbatis, "Metastatic renal clear cell carcinoma in the parotid gland: a study of immunohistochemical profile and cell adhesion molecules (cams) expression in two cases," *Pathology & Oncology Research*, vol. 13, no. 2, pp. 161–165, 2007.

[33] V. R. Arruda, C. S. P. Lima, C. R. E. Grignoli, M. B. De Melo, I. Lorand-Metze, F. L. Alberto, S. T. O. Saad, and F. F. Costa, "Increased risk for acute myeloid leukaemia in individuals with glutathione s-transferase mu 1 (gstm1) and theta 1 (gstt1) gene defects," *European journal of haematology*, vol. 66, no. 6, pp. 383–388, 2001.

[34] Z. Yousefi, J. Sarvari, K. Nakamura, Y. Kuramitsu, A. Ghaderi, and Z. Mojtahedi, "Secretomic analysis of large cell lung cancer cell lines using two-dimensional gel electrophoresis coupled to mass spectrometry," *Folia Histochemica et Cytobiologica*, vol. 50, no. 3, pp. 368–374, 2012.

[35] Y. Wang, W. Zang, Y. Du, Y. Ma, M. Li, P. Li, X. Chen, T. Wang, Z. Dong, and G. Zhao, "Mir-655 up-regulation suppresses cell invasion by targeting pituitary tumor-transforming gene-1 in esophageal squamous cell carcinoma," *J Transl med*, vol. 11, p. 301, 2013.

[36] M. Wan, T. Sun, R. Vyas, J. Zheng, E. Granada, and L. Dubeau, "Suppression of tumorigenicity in human ovarian cancer cell lines is controlled by a 2 cm fragment in chromosomal region 6q24-q25," *Oncogene*, vol. 18, no. 8, pp. 1545–1551, 1999.

APPENDIX

Algorithm 1: FISTA algorithm for *Transfer-Cox*

Input: Source dataset D_S, Target dataset D_T, Initial coefficient matrix $B^{(0)}$, w, μ, λ
Output: $\bar{B}$

1 **Initialize**: $B^{(1)} = B^{(0)}$, $d_{-1} = 0$, $d_0 = 1$,$\gamma_0 = 1$,$i = 1$;
2 **repeat**
3 Set $\alpha_i = \frac{d_{i-2}-1}{d_{i-1}}$, $S^{(i)} = B^{(i)} + \alpha_i(B^{(i)} - B^{(i-1)})$;
4 **for** $j = 1, 2, \cdots$ **do**
5 Set $\gamma = 2^j\gamma_{i-1}$;
6 Calculate $B^{(i+1)} = \pi_P(S^{(i)} - \frac{1}{\gamma}g^{'}(S^{(i)}))$;
7 Calculate $Q_\gamma(S^{(i)}, B^{(i+1)})$;
8 **if** $g(B^{(i+1)}) \leq Q_\gamma(S^{(i)}, B^{(i+1)})$ **then**
9 $\gamma_i = \gamma$, **break** ;
10 **end**
11 **end**
12 $d_i = \frac{1+\sqrt{1+4d_{i-1}^2}}{2}$;
13 $i = i + 1$;
14 **until** *Convergence of* $B^{(i)}$;
15 $\bar{B} = B^{(i)}$;

Algorithm 1 outlines the learning procedure of FISTA algorithm to solve optimization problem in Eq.(7). In lines 4-11, the optimal γ_i is chosen by the backtracking rule based on [5, Lemma 2.1, page 189], γ_i is greater than or equal to the Lipschitz constant of $g(\cdot)$ at $S^{(i)}$, which means γ_i is satisfied for $S^{(i)}$ and $\frac{1}{\gamma_i}$ is the possible biggest stepsize. In line 7, $Q_\gamma(S^{(i)}, B^{(i+1)})$ is the tangent line of $g(\cdot)$ at $S^{(i)}$, which can be calculated as

$$Q_\gamma(S^{(i)}, B^{(i+1)}) \tag{22}$$
$$= g(S^{(i)}) + \frac{\gamma}{2} \parallel B^{(i+1)} - S^{(i)} \parallel^2 + \langle B^{(i+1)} - S^{(i)}, g^{'}(S^{(i)}) \rangle$$