Survival Ensemble with Sparse Random Projections for Censored Clinical and Gene Expression Data

LIFENG ZHOU^{1,a)} HONG WANG^{1,b)} QINGSONG XU^{1,c)}

Received: April 6, 2016, Accepted: May 24, 2016

Abstract: Random projection is a powerful method for dimensionality reduction while its applications in high-dimensional survival analysis is rather limited. In this research, we propose a novel survival ensemble model based on sparse random projection and survival trees. Supported by the proper statistical analysis, we show that the proposed model is comparable to or better than popular survival models such as random survival forest, regularized Cox proportional hazard and fast cocktail models on high-dimensional microarray gene expression right censored data.

Keywords: survival ensemble, random projection, high-dimensional, gene expression, censored data

1. Introduction

In the past decade, interests have grown on studying the association between gene expression profiles and time-to-event outcomes and make prediction of subjects' future survival probabilities. Survival analysis in this scenario often faces two major challenges. First, censored data are usually observed instead of fully precise measurement of time-to event information. Second, the covariates including both genomic outcomes and clinical data are high-dimensional: the number of covariates is much larger than the number of observations. In the past two decades, various parametric and semi-parametric models have been proposed to deal with these two challenges. These approaches include regularized Cox-regression models [1], [2], [3], regularized accelerated failure time (AFT) models [4], [5], [6], partial least squares [7], [8], and supervised principal components [9], [10]. However, it is pointed out that when the underlying assumptions are not satisfied, these models may not work in high dimensional settings [11].

Recently, there has been a considerable interest in dealing with high dimensional survival data using non-parametric machine learning algorithms. Ensemble based approaches, in particular, has been considered in a variety of contexts and two major widely applied types of ensembles are boosting [11], [12], [13], [14], [15], [16] and random forest [17], [18]. The first instance of boosting for censored data was due to Ridgeway [19]. However, boosting with high dimensional censored data was not studied until Ref. [11] and they proposed a gradient boosting procedure using smoothing splines for estimating Cox base models. Due to the complexity of the problem, high dimensional survival analysis is not fully solved in previous studies and are still in need of

further research.

In high dimensional survival analysis, one practical issue we have to face is that most available algorithms are memory-based, namely, data operations and modelling should be done within the limits of the computer's memory. Consequently, when high performance computing facilities are not available for big datasets, memory shortage problems would be frequently encountered. To cope with such memory constraints, big data objects can be stored on hard disks as is done in R packages such as "ff" [20]. However, these sophisticated tools cannot be easily applied to current survival algorithms as many codes in the original algorithms need to be rewritten to cope with new data structures.

In this article, we introduce a novel survival tree ensemble based on sparse random projection for analyzing high-dimensional censored data with both clinical variables and gene expression information. We choose decision trees as the base learner for our survival ensemble because it has the flexibility to accommodate high dimensional covariates and it is the most widely used non-parametric survival model [21]. The reasons why random projection is introduced are two-fold. First, random projection is very unstable i.e., different random projections (mappings) may result in drastically different prediction results and this generates the diversity required in constructing a successful ensemble. Second, substantial computation reduction is achieved by sparse random projections as the projected data can get a significant speedup in processing time with little loss in prediction.

To alleviate the memory shortage problem, we adopt a file-based approach, namely, big data objects such as generated random projection matrixes and trained base models are stored on the hard disk as external files. And codes of the original base survival algorithm can keep unchanged and this is particularly important for a survival ensemble in that a new type of base survival model can be easily plugged into the ensemble without rewriting the whole codes.

The major contributions of this article can be summarized as

¹ School of Mathematics and Statistics, Central South University, China

a) lfzhou@csu.edu.cn

b) Corresponding Author, wh@csu.edu.cn

c) qsxu@csu.edu.cn

follows:

- We extend sparse random projection from classification and clustering to survival analysis for right censored data;
- We provide a divide-and-conquer file-based data objects management approach to store big survival data models on common desktop computers.

2. Methods

Survival analysis considers how a set of covariates or treatment $\mathbf{X} = (X_1, X_2, \dots, X_p)$ affects the time to disease or some other events τ . Hence, one major goal of survival analysis is to estimate the probability that a subject will survive past a certain time. If there is no censoring, standard regression could be applied. However, τ is usually not fully observed, i.e., $\tau = \min(U, C)$ typically contains both real survival times U and censored times C . In such cases, indicator variables $\delta = I(U \leq C)$ can be applied. We usually assume $(\mathbf{X}_i, \tau_i, \delta_i)$, $i = 1, 2, \dots, n$ to be n independent and identically distributed samples from $(\mathbf{X}, \tau, \delta)$.

In this section, we first give a brief introduction of sparse random projection, and then we propose to model the survival function using a non-parametric sparse random projection ensemble.

2.1 Sparse Random Projection

Projection techniques are widely used for dimension reduction in classification and regression problems. One class of such projections with low computational cost is random projection (RP) [22]. Given a p -dimensional data with n samples $X \subset \mathbb{R}^p$, a random projection from p dimensions to p' dimensions is a linear mapping $f: X \rightarrow X' \subset \mathbb{R}^{p'}$ ($p' < p$) and the newly reduced dimension data set for further processing is $X' = XR'$, where R' is the corresponding random matrix of $p \times p'$.

It has been pointed out that random projection respects local structures between data points and is computationally efficient with sufficient accuracy [23]. The main theoretical result behind the efficiency and feasibility of random projection is the following Johnson-Lindenstrauss Lemma [24].

Lemma 1 Given an integer n and let $\varepsilon > 0$, then for all positive integers $p' \geq p_0 = O(\varepsilon^{-2} \log n)$ and a set H of n points random selected from $\mathbb{R}^p$, there exists a mapping $f: \mathbb{R}^p \rightarrow \mathbb{R}^{p'}$ such that for all $u, v \in H$,

$$(1 - \varepsilon)\|u - v\|^2 \leq \|f(u) - f(v)\|^2 \leq (1 + \varepsilon)\|u - v\|^2 \quad (1)$$

The above Lemma asserts that any randomly selected n points in p -dimensional space can be projected down to $p' = O(\varepsilon^{-2} \log n)$ -dimensional space, with large probability that the squared distances between data points change by a factor of no more than $1 \pm \varepsilon$. In other words, the dimensionality of the original dataset can be greatly reduced to p' while all pairwise distances between data points are maintained without introducing significant distortions.

Here, mapping f (also called JL-embedding) can be constructed by random projections with positive probabilities [25]. For example, the projection in question can be realized by a sparse random matrix $R(i, j)$ with the following elements r_{ij} :

$$r_{ij} = \begin{cases} \sqrt{3}, & \text{with probability } 1/6; \\ 0, & \text{with probability } 2/3; \\ -\sqrt{3}, & \text{with probability } 1/6. \end{cases} \quad (2)$$

And the consequent random projection holds for Lemma 1 with probability at least $1 - n^{-\beta}$ [25].

From the above, we notice that random matrix does not depend on the original dimensionality. Random projection simply projects the high-dimensional data into a low-dimensional one without calculating eigenvalues and eigenvectors of data matrices. As a computationally efficient approach, (sparse) random projection can find many applications in high-dimensional survival analysis when instantaneous results are necessary.

2.2 Survival Ensemble with Sparse Random Projection

Given a high dimensional survival data matrix D of dimension $n \times (p + 2)$ and $D = (\tau_j, \delta_j, \mathbf{X}_j)$, $j = 1, 2, \dots, n$, where n is the number of samples observed, and $\mathbf{X}_j$ is a p -dimensional variable set which usually contains p_1 -dimensional gene expression data and p_2 -dimensional covariates. Here, the whole variable set $\mathbf{X}_{n \times p}$ contains of n p -dimensional observations, namely,

$$\begin{aligned} \mathbf{X}_{n \times p} &= (\mathbf{X}_{n \times p_1}, \mathbf{X}_{n \times p_2}) \\ &= \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p_1} & x_{1p_1+1} & x_{1p_1+2} & \cdots & x_{1p_1+p_2} \\ x_{21} & x_{22} & \cdots & x_{2p_1} & x_{2p_1+1} & x_{2p_1+2} & \cdots & x_{2p_1+p_2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np_1} & x_{np_1+1} & x_{np_1+2} & \cdots & x_{np_1+p_2} \end{bmatrix} \end{aligned} \quad (3)$$

We first give a high-level description of how the proposed Survival Ensemble with Sparse Random Projections (SESRP) algorithm train a base survival function in the ensemble:

- (1) Generate a bootstrap sample $D' = (\tau', \delta', \mathbf{X}')$ of size n from D and $\mathbf{X}'_{n \times p} = (\mathbf{X}'_{n \times p_1}, \mathbf{X}'_{n \times p_2})$.
- (2) Randomly generate two sparse random matrices M_1 of $p_1 \times p_1'$ and M_2 of $p_2 \times p_2'$ and we set $p_1' = \lceil \sqrt{p_1} \rceil$, $p_2' = \lceil \sqrt{p_2} \rceil$ for simplicity.
- (3) Multiply $\mathbf{X}'_{n \times p_1}, \mathbf{X}'_{n \times p_2}$ by the above two sparse random matrices M_1 and M_2 respectively. Combining the results together, we obtain the new variable set $\mathbf{X}''_{n \times (p_1' + p_2')} = (\mathbf{X}'_{n \times p_1} M_1, \mathbf{X}'_{n \times p_2} M_2)$.
- (4) Use the newly formed data $D'' = (\tau', \delta', \mathbf{X}'')$ as the training set, we can train a base survival tree learner.

The reasons why we use two sparse random matrices here is two fold. First and foremost, it could bring more diversity to the survival ensemble and will increase its prediction accuracy. Second, merging a big gene expression matrix with a covariate matrix and multiplying by a random matrix is time-consuming as genes and covariates are usually stored in different data structures. Merging two reduced matrices, however, is extremely fast.

A survival ensemble usually contains a certain number of base survival models to maintain moderate predication accuracy. However, in case of big or high dimensional survival data, training too many base models on common desktop computer simultaneously will cause the memory shortage problem. Here, instead of training the whole ensemble with all base models at a time, we

propose a divide-and-conquer approach, namely, we will store blocks of base models and related data objects to disk files and read such block models when necessary. In our algorithm, we use the parameter *blocksize* to control the size of base model blocks.

Next, we present the pseudo-code of the proposed SESRP algorithm in Algorithm 1:

Algorithm 1 Survival Ensemble with Sparse Random Projections (SESRP)

```

1: INPUT:
2:   Training set:  $D = (\tau_j, \delta_j, \mathbf{X}_j), j = 1, \dots, n, \mathbf{X}_{n \times p} = (\mathbf{X}_{n \times p_1}, \mathbf{X}_{n \times p_2})$ 
3:    $L$ : Ensemble size
4:    $Tree$ : Survival tree base learning algorithm
5:    $B$ : Block size
6: OUTPUT:
7:   Survival ensemble  $C$ 
8: procedure SESRP( $D, B$ )
9:   Create matrix arrays  $A_1, A_2$  of size  $B$  and tree list  $T$  also of size  $B$ 
10:  while  $i$  in  $1 : L$  do
11:    Randomly generate a bootstrap sample  $D' = (\tau', \delta', \mathbf{X}')$  from  $D$ .
12:    Generate two sparse random matrices  $M_1$  and  $M_2$ 
13:    Multiply  $\mathbf{X}'_{n \times p_1}, \mathbf{X}'_{n \times p_2}$  by  $M_1$  and  $M_2$  respectively, and the obtained variable set is  $\mathbf{X}''_{n \times (p_1 + p_2)}$ 
14:    Train a base survival tree  $C_i$  using  $D' = (\tau', \delta', \mathbf{X}'')$  with CART algorithm
15:    if  $(i \bmod B = 0)$  then
16:      Save data objects  $A_1, A_2, T$  to disk file  $F_k$  where  $k = \text{floor}(i/B)$ 
17:    else
18:       $A_1[i] = M_1, A_2[i] = M_2, T[i] = C_i$ 
19:    end if
20:  end while
21:  if  $(i \bmod B \neq 0)$  then
22:    Save data objects  $A_1, A_2, T$  to disk file  $F_k$  where  $k = \text{floor}(i/B) + 1$ 
23:  end if
24:  return The survival ensemble  $C = \{C_i\}$ 
25: end procedure
26: In prediction, for  $m$  new data samples  $\mathcal{X} = (\mathcal{X}_{m \times p_1}, \mathcal{X}_{m \times p_2})$ 
27: while  $i$  in  $1 : L$  do
28:   Load  $F_k$  and multiply  $\mathcal{X}_{m \times p_1}, \mathcal{X}_{m \times p_2}$  by corresponding  $M_1$  and  $M_2$  respectively and we obtain the transformed new data  $\mathcal{X}'$ 
29: end while
30: Hence, the average hazard rate  $h$ 

```

$$h = \frac{1}{L} \sum_{C_i \in C} C_i(\mathcal{X}')$$

3. Results & Discussion

In this section, we first describe the data set used in the experiments and then statistical tests that we used. Finally, we present and discuss the experimental results.

3.1 Datasets

In this study, we want to test our SESRP algorithm's performance on real high-dimensional survival data extensively analyzed in the statistical literature. In the experiments, distant metastasis-free survival (DMFS) was the primary survival endpoint, which is defined as the time elapsing between cancer diagnosis and date of local or systemic relapse, or death. When

DMFS information was unavailable, relapse-free or overall survival information is used.

A brief introduction and summary of the three used datasets are given below.

3.1.1 TransBig Dataset

This breast cancer dataset contains the gene expression and clinical data published in Ref. [26]. The original data contains 198 samples to independently validate a 76-gene prognostic breast cancer signature as part of the TransBig project. After omitting missing values and incomplete covariates, 9 out of 21 covariates are retained. Thus, in the experiment, 22283 gene features and 9 clinical covariates are provided for each sample.

3.1.2 VDX Dataset

The Veridex (VDX) dataset which contains 344 patients with primary breast cancer was published in Ref. [27]. After omitting missing values and incomplete covariates, 8 out of 21 covariates are retained. Thus, in the experiment, 22283 gene features and 8 clinical covariates are provided for each sample.

3.1.3 TCGA Dataset

The original TCGA dataset contains both clinical covariates and gene expression information of 3096 cancer patients covering 12 major types of cancers. It was provided by TCGA (The Cancer Genome Atlas) and presented in Refs. [28], [29]. In each patient sample, 19,420 gene state information and 5 clinical covariates are provided. It is a typical high-dimensional dataset with large samples.

3.2 Performance Metrics and Statistical Tests

We assess the performance of the proposed algorithm using 5-2 fold cross-validation, i.e., we randomly split each dataset into two halves, train the models using half of the dataset, and then test the models with the other half and vice versa. This validation procedure is repeated five times, each with a random half/half partition.

In survival analysis, doctors or decision makers are much concerned with the relative risks between patients with different genes and covariates information. Hence, as suggested by Ref. [30], we adopt Harrell's concordance index (C-index, CI) measure [31] to evaluate the accuracy of such relative risks in later experiments.

The C-index is a generalisation of the area under the ROC curve (AUC) and it is defined as the proportion of all usable patient pairs in which the predictions and outcomes are concordant [31]. The C-index can be calculated as follows:

- Create all pairs of observed times.
- For all valid time pairs, i.e., pairs where one time T_{j1} is greater than the other T_{j2} , test whether the corresponding predictions are concordant, i.e., $\eta_{j1} > \eta_{j2}$. If so, add 1 to the running sum s ; If $\eta_{j1} = \eta_{j2}$, add 0.5 to the sum s ; If $\eta_{j1} < \eta_{j2}$, add 0 to the sum s .
- Count the number n of valid time pairs. Divide the total sum s by the number of valid time pairs n and we obtain $CI = s/n$.

Similar to AUC, $CI = 1$ indicates the model has a perfect prediction accuracy and $CI = 0.5$ implies that the model is as good as a random predictor.

We use the non-parametric Friedman test [32] to test the statistical significance of various survival models. Friedman's test statistic is based on the average ranked performance of the algorithms on all runs of the datasets and can be calculated according to the following formula:

$$FT = \frac{12}{nm(m+1)} \sum_{j=1}^m \left(\sum_{i=1}^n r_i^j \right)^2 - 3n(m+1) \quad (4)$$

where m denotes the number of survival models, n the number of datasets (different splits of the same dataset are regarded as different datasets), and r_i^j the rank of survival models j on the i -th dataset. If the value of FT is large enough, the null hypothesis that there is no significant difference among the different survival models can be rejected and a Nemenyi post-hoc test can be adopted to find where the difference lies.

For two survival models C_1 and C_2 , the Nemenyi statistic z is calculated as follows:

$$z = \frac{R_{j1} - R_{j2}}{\sqrt{\frac{m(m+1)}{6n}}} \quad (5)$$

where R_j denotes the mean rank of survival models C_j on all runs of the dataset, namely, $R_j = \frac{1}{n} \sum_{i=1}^n r_i^j$. The performance of two survival models is significantly different if the z value is larger than a certain critical value [32].

3.3 Results and Discussions

3.3.1 Performance Comparison Result

Here, we want to compare the proposed method with three state-of-the-art survival models for high-dimensional censored data. The first method is random survival forest (RSF); the second method is the regularized Cox model (CoxNet) [33] and the third method is fast cocktail model (CockTail) [34]. Comparisons with these models are conducted with corresponding “randomForestSRC”, “glmnet”, and “fastcox” packages in R. For the ease of notation, survival models SESRP, RSF, CoxNet and CockTail are denoted by A, B, C, D, respectively when necessary.

In the experiments, we want all models to have the same opportunities to achieve the best results, thus the default settings are adopted. For ensemble methods, i.e., SESRP, RSF, 500 trees are built. For CoxNet, L2-penalization is chosen instead of L1-penalization in that lasso-type Cox models's performance is rather poor on TransBig and VDX datasets.

The following Figs. 1, 2 and 3 report the performance of SESRP, RSF, CoxNet and CockTail algorithms in term of C-index on 5-2 fold cross-validations of TransBig, VDX and TCGA.

From these figures, we may notice that survival models's performance vary on these different benchmark datasets. In term of averaged C-Index, the proposed SESRP approach takes the first place on both TransBig and VDX datasets and ranks the second on the TCGA dataset; RSF always takes the third place on all three datasets; CoxNet ranks the second on TransBig and VDX and the last on TCGA; Cocktail ranks the third and the last on TransBig and VDX, respectively but takes the lead on TCGA.

The reason why SESRP is beaten by Cocktail on TCGA might be that random projection approach may not work well when the

Fig. 1 Performance in terms of C-index on TransBig data.

Fig. 2 Performance in terms of C-index on VDX data.

Fig. 3 Performance in terms of C-index on TCGA data.

embedding dimension ($p1'$ or $p2'$) is smaller than the so-called statistical dimension [35]. As in the experiments, we simply set the embedding dimension to be the root of dimension of the variables, a fine tuning of parameter $p1'$ or $p2'$ would improve SERSP in this case.

Next, we discuss the overall performance of all compared models. The Friedman rank sum test statistic across these datasets is 13.8706 and is significant as the p -value is $p = 0.003087$. The average ranks of survival models A, B, C and D are:

$$R_{jA} = 1.733, R_{jB} = 3.1, R_{jC} = 2.4, R_{jD} = 2.767$$

Thus, to find out which pairs of algorithms are significantly different, we compute the Nemenyi test statistics for different pairs of survival models and obtain $z_{BA} = 4.1$, $z_{CA} = 2$ and $z_{DA} = 3.1$.

For $\alpha = 0.05$, the critical value of Nemenyi's test is 2.619. It can be seen that all Nemenyi statistics except the z_{CA} value exceed 2.619. Thus, in terms of C-index, there exists significant differences between SERSP and RSF, CockTail algorithms on these datasets. In other words, from the performance on different datasets, SERSP is significantly better than RSF and CockTail.

Tough SESRP beats CoxNet in terms of average rank, the difference between SESRP and CoxNet is not significantly different according to the Nemenyi statistic. However, the performances of CoxNet on TCGA and VDX datasets are less stable than the proposed SESRP algorithm as its worst cases got zero values on these datasets.

Table 1 Comparison on running time.

	TransBig	VDX	TCGA
SESRP	2,074.8	1,585	3,073
RSF	1,087	660.4	16,434.1

Finally, we also consider the time efficiency of the proposed survival model. Since SESRP is a tree ensemble, here, we only compare it with RSF. Average running time of both tree ensembles on three datasets is reported in the following **Table 1**.

Looking at Table 1, one may observe that SESRP is a bit slower than RSF on small sample-sized datasets such as TransBig and VDX and this is mainly due to overheads caused by hard disk and memory swapping. However, unlike RSF, SESRP's running time does not vary too much across datasets and its superiority on efficiency stands out when applied to survival data with larger sample size as we see that SESRP is four times faster than RSF on the TCGA dataset.

4. Conclusions

In this study, we have presented a new survival ensemble with sparse random projection algorithm for high dimensional gene features with clinical covariates. By studying the well-known benchmark datasets, we have found that SESRP generally outperforms or comparable to state-of-the-art survival models such as rotation survival forest, regularized Cox proportional hazard and fast cocktail models in terms of C-index metric. As a non-parametric approach, SESRP does not impose parametric assumptions on hazard functions, and it extends the well-known sparse random projection methodology to survival analysis.

The R code and the supplementary material are available at url: <https://github.com/whcsu/sesrp> and we are working hard to provide an R package for the proposed SESRP algorithm as soon as possible.

Although motivated by a gene expression dataset, our approach can be applied to other genomic data as well such as SNP data. A limitation of our proposed algorithm is the lack of interpretability as compared to regression-based models which is counterbalanced by the ability to model complex interaction effects among the gene features and clinical covariates and hence gains in the prediction accuracy.

Acknowledgments This work was supported in part by the Social Science Foundation for Young Scholars of Ministry of Education of China Under Grant No.15YJCZH166. The authors want to all reviewers for their valuable and constructive comments, which greatly improves the quality of this paper.

References

- [1] Fan, J. and Li, R.: Variable Selection for Cox Proportional Hazards Model and Frailty Model, *The Annals of Statistics*, Vol.30, No.1, pp.74–99 (online), DOI: 10.2307/2700003 (2002).
- [2] Hastie, T. and Tibshirani, R.: Efficient quadratic regularization for expression arrays, *Biostatistics*, Vol.5, No.3, pp.329–340 (2004).
- [3] Gui, J. and Li, H.: Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data, *Bioinformatics*, Vol.21, No.13, pp.3001–3008 (2005).
- [4] Huang, J., Ma, S. and Xie, H.: Regularized estimation in the accelerated failure time model with high-dimensional covariates, *Biometrics*, Vol.62, No.3, pp.813–820 (2006).
- [5] Datta, S., Le-Rademacher, J. and Datta, S.: Predicting patient survival

- from microarray data by accelerated failure time modeling using partial least squares and LASSO, *Biometrics*, Vol.63, No.1, pp.259–271 (2007).
- [6] Cai, T., Huang, J. and Tian, L.: Regularized estimation for the accelerated failure time model, *Biometrics*, Vol.65, pp.394–404 (online), DOI: 10.1111/j.1541-0420.2008.01074.x (2009).
- [7] Nguyen, D.V. and Rocke, D.M.: Partial least squares proportional hazard regression for application to DNA microarray survival data, *Bioinformatics*, Vol.18, No.12, pp.1625–1632 (2002).
- [8] Huang, J. and Harrington, D.: Iterative Partial Least Squares with Right-Censored Data Analysis: A Comparison to Other Dimension Reduction Techniques, *Biometrics*, Vol.61, No.1, pp.17–24 (2005).
- [9] Li, L. and Li, H.: Dimension reduction methods for microarrays with application to censored survival data, *Bioinformatics*, Vol.20, No.18, pp.3406–3412 (2004).
- [10] Bair, E., Hastie, T., Paul, D. and Tibshirani, R.: Prediction by supervised principal components, *Journal of the American Statistical Association*, Vol.101, No.473, pp.119–137 (2006).
- [11] Li, H. and Luan, Y.: Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data, *Bioinformatics*, Vol.21, No.10, pp.2403–2409 (online), DOI: 10.1093/bioinformatics/bti324 (2005).
- [12] Hothorn, T. and Bühlmann, P.: Model-based boosting in high dimensions, *Bioinformatics (Oxford, England)*, Vol.22, No.22, pp.2828–9 (online), DOI: 10.1093/bioinformatics/btl462 (2006).
- [13] Ma, S. and Huang, J.: Clustering threshold gradient descent regularization: with applications to microarray studies, *Bioinformatics*, Vol.23, No.4, pp.466–472 (2007).
- [14] Binder, H. and Schumacher, M.: Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models, *BMC bioinformatics*, Vol.9, No.1, p.14 (2008).
- [15] Binder, H., Allignol, A., Schumacher, M. and Beyersmann, J.: Boosting for high-dimensional time-to-event data with competing risks, *Bioinformatics*, Vol.25, No.7, pp.890–896 (online), DOI: 10.1093/bioinformatics/btp088 (2009).
- [16] Wang, Z. and Wang, C.: Buckley-James boosting for survival analysis with high-dimensional biomarker data, *Statistical Applications in Genetics and Molecular Biology*, Vol.9, No.1, Article 24 (2010).
- [17] Ishwaran, H., Kogalur, U.B., Gorodeski, E.Z., Minn, A.J. and Lauer, M.S.: High-dimensional variable selection for survival data, *Journal of the American Statistical Association*, Vol.105, No.489, pp.205–217 (2010).
- [18] Ishwaran, H., Kogalur, U.B., Chen, X. and Minn, A.J.: Random survival forests for high-dimensional data, *Statistical Analysis and Data Mining*, Vol.4, No.1, pp.115–132 (online), DOI: 10.1002/sam.10103 (2011).
- [19] Ridgeway, G.: The state of boosting, *Computing Science and Statistics*, pp.172–181 (1999).
- [20] Daniel, A., Christian, G., Oleg, N., Jens, O. and Walter, Z.: ff: memory-efficient storage of large data on disk and fast access functions, Technical report (2014).
- [21] Bou-Hamad, I., Larocque, D. and Ben-Ameur, H.: A review of survival trees, *Statistics Surveys*, Vol.5, pp.44–71 (2011).
- [22] Bingham, E. and Mannila, H.: Random projection in dimensionality reduction: Applications to image and text data, *Proc. 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.245–250, ACM (2001).
- [23] Wang, J.: *Geometric structure of high-dimensional data and dimensionality reduction*, Springer (2012).
- [24] Johnson, W.B. and Lindenstrauss, J.: Extensions of Lipschitz mappings into a Hilbert space, *Contemporary Mathematics*, Vol.26, No.189-206, p.1 (1984).
- [25] Achlioptas, D.: Database-friendly random projections: Johnson-Lindenstrauss with binary coins, *Journal of Computer and System Sciences*, Vol.66, No.4, pp.671–687 (2003).
- [26] Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., Viale, G., Delorenzi, M., Zhang, Y., d'Assignies, M.S., et al.: Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series, *Clinical Cancer Research*, Vol.13, No.11, pp.3207–3214 (2007).
- [27] Wang, Y., Klijn, J.G., Zhang, Y., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M.E., Yu, J., et al.: Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer, *The Lancet*, Vol.365, No.9460, pp.671–679 (2005).
- [28] Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al.: Mutational landscape and significance across 12 major cancer types, *Nature*, Vol.502, No.7471, pp.333–339 (2013).
- [29] Fang, H. and Gough, J.: The 'dnet' approach promotes emerging re-

- search on cancer patient survival, *Genome Medicine*, Vol.6, p.64 (online), DOI: 10.1186/s13073-014-0064-8 (2014).
- [30] Ishwaran, H., Kogalur, U.B., Blackstone, E.H. and Lauer, M.S.: Random survival forests, *The Annals of Applied Statistics*, Vol.2, No.3, pp.841–860 (2008).
 - [31] Harrell, F.E., Lee, K.L. and Mark, D.B.: Tutorial in biostatistics multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, *Statistics in Medicine*, Vol.15, pp.361–387 (1996).
 - [32] Demšar, J.: Statistical comparisons of classifiers over multiple data sets, *The Journal of Machine Learning Research*, Vol.7, pp.1–30 (2006).
 - [33] Simon, N., Friedman, J.H., Hastie, T. and Tibshirani, R.: Regularization Paths for Cox’s Proportional Hazards Model via Coordinate Descent, *Journal of Statistical Software*, Vol.39, No.5, pp.1–13 (2011).
 - [34] Yang, Y. and Zou, H.: 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).
 - [35] Oymak, S. and Tropp, J.A.: Universality laws for randomized dimension reduction, with applications, *arXiv preprint arXiv:1511.09433* (2015).

Lifeng Zhou was born in 1981. She received her M.S. degree from Central South University in 2013 and currently is a Ph.D. Candidate in Statistics at Central South University. Her research interests are bio-data mining and bioinformatics.

Hong Wang was born in 1977. He received his Ph.D. from Central South University, China in 2015. He has been with Central South University since 2000. His current research interests are ensemble learning, survival analysis, bioinformatics and biostatistics. He is a member of Institute of Mathematics & Statistics.

Qingsong Xu was born in 1961. He received his Ph.D. degree from Hunan University in 2001 and has been a professor in statistics with Central South University since 2004. His research interest is applied statistics.

(Communicated by *Takeshi Obayashi*)