SCAD Support Vector Machine for Survival Analysis

Boyi Guo
Department of Statistics
University of Illinois at Urbana-Champaign
Champaign, IL 61820
boyiguo1@illinois.edu

1 Introduction

In many medical studies, a significant portion of data is a collection of time-to-event observations. Estimating the failure time, as a function of patient demographic and prognostic variables, is of central importance for risk assessment and health planing. Sometimes,the endpoint is something different from the the time event of interest happens, due to length of study, etc. Most of such data is subject to right censoring. The goal of this paper is to develop a tool for analysing such data using machine learning techniques.

Traditional research conducted on survival analysis, especially right-censored failure time analysis, mainly focuses on estimating the failure time distribution or hazard function. The main approaches include using parametric models, such as the Weibull distribution, and semi-parametric models, such as proportional hazard models[1]. Usually restrictive assumptions are made for these models, such as the distribution function is smooth in both time and covariates. They become burdens when considering today's high-dimensional data setting, such as micoarray data.

Recently, lots of papers propose using support vector machine(SVM) for survival data. The choice of SVM is motivated by the fact that SVM is easy-to-compute technique that enable estimation or classification under weak or no assumptions on distribution. It typically minimizes a regularized version of the empirical risk over some reproducing kernel Hilbert space. We note that a number of SVMs have been suggested for survival data. Most of these efforts formalize the problem under the regression setting[2]-[4], while others formalize the problem under the classification setting[5]-[6]. However, formalization using regression setting is intrinsically more difficult than classification. Further practitioners generally use the modelling outputs as a reference and they are usually concerned with what the factors impacting on the event of interest most are. Under the classification setting, we can also compare the ranking of patients.

In this paper, we propose to use a special classification formulation that addresses the issues of incomplete information in the survival time. Instead of predicting the survival time, We try to classify the ranking of the observation. We adapt the smoothly clipped absolute deviation penalty(SCAD) SVM[7] framework to non-censored and right-censored data as follows. First, we represent the distributions quantity of interest as a hinge loss function, i.e., a function that minimizes the risk with respect to a loss function. Afterwards, We construct a data-dependent version of this loss function using inverse-probability of censoring weighting[8]. We then minimize a regularized empirical risk with respect to this data-dependent loss function to obtain an SVM decision function for non-censored or right-censored data. Finally, we define the SVM learning method for censored data as the mapping that assigns for every survival data set its corresponding SVM decision function.

The paper is organized as follows. In Section 2 we review the standard SCAD SVM and discuss the integration of inverse probability weighting. Section 3 introduced the implementation of the method. Simulation and application on real data appear in Section 4. Concluding remarks appear in Section 5.

2 Method

2.1 Survival Data Notation

This paper considers failure time data according to the following specifications. Let $T \geq 0$ be a random variable denoting the time to failure. Assume that (Z, T) is a random vector with a joint distribution function F_{ZT}. Here $X \subset \mathbb{R}^d$ denote the $d \geq 0$ covariates of the subject under study. For notational convenience, the case of ties is excluded by assumption. Let $C > 0$ denote the time of censoring following F_C, which is assumed to be independent of both Z and T. Now let $\delta \in \{0, 1\}$ denote a binary random variable indicating whether the subject is withdrawn from the study before its actual failure (right censoring). If $\delta = 1$, let $U \geq 0$ denote the time of failure. Formally

$$U = \begin{cases} T & \text{if } \delta = 1 \\ C & \text{if } \delta = 0 \end{cases} \tag{1}$$

let the set $\mathcal{D} = \{(Z_i, U_i, \delta_i)\}_{i=1}^n$ denote the observed i.i.d. samples from a distribution $F_{ZU\delta}$. Let the random variable Z denote the random triple (Z, U, δ) with i.i.d. samples $\{D_i\}_{i=1}^n$

2.2 SCAD SVM for survival data

Through the L_1 penalty gives sparse solutions, the estimates can be biased for large coefficients since larger penalties are imposed on larger coefficients. Zhang et al.[7] proposed the smoothly clipped absolute deviation SVM which overcomes the biasness problem of the L_1 penalty. They showed that the SCAD penalty produces sparse solution by thresholding small estimates to zero, provides nearly unbiased estimates for large coefficients and gives a model continuous in data. It can be perfectly performed on high-dimensional low sample size data and conduct variable selection.

Though having the same form as the L_1 penalty at the neighbourhood of zero, the SCAD applies a constant penalty for large coefficients while the L_1 penalty increases linearly as the coefficient increases. It is the distinct feature that guards the SCAD penalty against producing biases for estimating large coefficients. Mathematically, the SCAD penalty has the expression

$$p_\lambda(|\beta|) = \begin{cases} \lambda|\beta| & \text{if } |\beta| \leq \lambda \\ -\frac{(|\beta|^2 - 2a\lambda|\beta| + \lambda^2)}{2(a-1)} & \text{if } \lambda < |\beta| \leq a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\beta| > a\lambda \end{cases} \tag{2}$$

where $\lambda > 0$ are tuning parameter. Fan and Li[9] showed that Bayes risks are not sensitive to the choice of a, and $a = 3.7$ is a good choice for various problems. We also use $a = 3.7$ in our examples.

Interestingly, when $p \gg n$, linear classifiers often give better performances than non-linear ones in many applications, even though non-linear methods are known to be more flexible. Since we are focus on classifying high-dimensional low sample size data, only linear SVMs are considered in this paper. In other words, we use the input vector x as basis functions, i.e. h(x)=x and q=p.

The original objective function for SCAD SVM describes as

$$\min_{\beta_0, \beta} \quad \sum_{i=1}^{n} \Big[1 - y_i\Big(\beta_0 + \sum_{j=1}^{q} \beta_j h_j(x_i)\Big)\Big]_+ + \sum_{j=1}^{q} p_\lambda(|\beta_j|) \tag{3}$$

consisting of the hinge loss part and the SCAD penalty on β. To be noted, when $p \gg n$, linear classifiers often give better performances than non-linear ones in many applications[10], even though non-linear methods are known to be more flexible. Since we are focus on classifying high-dimensional low sample size data, only linear SVMs are considered in this paper. In other words, we use the input vector x as basis functions, i.e. $h(x) = x$ and $q = p$.Adapting censoring into SCAD SVM context, we set up our new objective function as

$$\min_{\beta_0, \beta} \quad \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij}\big[1 - \mathbb{I}(i = j)\big]\Big[1 - Y_{ij}\Big(\beta_0 + \sum_{l=1}^{q} \beta_l h_l(x_{ij})\Big)\Big]_+ + \sum_{l=1}^{q} p_\lambda(|\beta_l|) \tag{4}$$

where W_{ij} is inverse probability weights for censored data, and Y_{ij} is the classification result regarding the ranking of two samples, and X_{ij} is the difference of covariates of two samples. Mathematically, they are

$$Y_{ij} = \mathbb{I}(U_i < U_j) - \mathbb{I}(U_i > U_j) \quad \text{and} \quad X_{ij} = Z_i - Z_j \tag{5}$$

To integrate censoring into the classification model, we introduce the inverse probability weighing proposed by Cai and Cheng[8]. Conditioning on Z_i and Z_j, the expectation of $\delta_j \mathbb{I}(U_i > U_j)/G(X_j)^2$ is the same as the expectation of $\mathbb{I}(T_i > T_j, T_j < \tau_0)$. Specifically, $\hat{\beta} = \text{argmax}\hat{Q}_n(\beta)$ where

$$\hat{Q}_n(\beta) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\delta_j \mathbb{I}(X_i > X_j)}{\hat{G}(x_j)^2} \mathbb{I}(\beta^T Z_i < \beta^T Z_j). \tag{6}$$

They show that under regularity conditions, $\hat{\beta}$ is consistent for $\bar{\beta} = \text{argmax}\hat{Q}_{\tau_0}(\beta)$ and $n^{\frac{1}{2}}(\hat{\beta} - \bar{\beta})$ converges in distribution to a zero-mean multivariate normal, where $\hat{Q}_{\tau_0}(\beta) = \mathbb{P}(T_i > T_j, \beta^T Z_i < \beta^t Z_j, T_j < \tau_0)$. We implement this idea in our classification model in the form of

$$W_{ij} = \mathbb{I}(U_i < U_j)\delta_i + \mathbb{I}(U_i > U_j)\delta_j \tag{7}$$

In the non-censoring case, $W_{ij} = 0$ all the time, which means every pair of observations in dataset will be fed into learning machine. It will automatically become a normal classification. In the right-censoring case, $W_{ij} \in \{0, 1\}$ depends on if the early finished observation is censored. Only if the early finished observation is not censored, this pair of observations will be fed into learning machine. Otherwise, the pair will be dropped.

3 Implementation

The main frame work of the program is built upon well developed R package *penalizedSVM*[11]. Before feeding survival to the function *svm.fs*, we transform the data with the help of R package *foreach*. It efficiently reduces the time of transform dataset in appropriate format and setting up the weights for each pair of observations. More detailed information are enclosed in attached R files.

4 Result

4.1 Simulation

Simulation studies were conducted to evaluate the performance of the proposed method. The following describes the method to generate input data with censored survival outcomes that emulate the mechanisms presented by the actual data. We first sample 12-dimensional input data with 200 observations from a multi-variate normal distribution with zero means and variance-covariance matrix Σ. $Sigma$ is set to have the same correlation coefficient $\rho = 0.5$ for all input variables. We randomly assign first 7 model parameters, while last 5 of them are 0. And we calculate $logT = \beta^T * Z + \epsilon$. Hence, this model is only associated with first 7 variables plus random noise. Finally, we compute random sample censored time from exponential distribution and calculate the $delta$ by comparing T and C.

We analyse the simulation data with SCAD SVM and build the model with the training data and evaluate the performance of the model with the test data. The regularization parameter λ is determined with 5-fold cross-validation with training data only.

Due to the computation limit of my computer and the configuration of *penelizedSVM*, we only tune $\lambda \in \{0.01, 0.1, 0.3, 0.5\}$. The training result are as follows. "Est. Var #" represents the number of variables left with corresponding lambda value; "Correct Var #" represent the number of correct variables selected by learning machine divided by the number of variables left by learning machine or in the original model; "GACV" represent the generalized approximate cross validation error estimation.

λ	0.01	0.1	0.3	0.5
Est. Var #	9	6	5	5
Correct Var #	$\frac{7}{9}$	$\frac{6}{7}$	$\frac{5}{7}$	$\frac{5}{7}$
GACV	0.016	0.0155	0.0229	0.0224

SVM recommends 0.1 as λ, since it gives the lowest cross validation error. Then we feed the test data in SVM with $\lambda = 0.1$. The classification result shows as follow

	predict label	
label	-1	1
-1	1864	27
1	32	1920

The simulation result shows the method can not only reduce the dimension of the data correctly,from 12 to 6, but also classify the ranking of observation pairs correctly.

4.2 Real Data

In this paper, we study a manually curated ovarian cancer data for gene expression meta-analysis. This resource provides uniformly prepared microarray data for 2970 patients from 23 studies with curated and documented clinical metadata. Due to some limitation, we will pick one dataset from all these 23 studies and run SCAD SVM on it.The data is coming from *curatedOvarianData* Bioconductor package[12].

The data set contains 228 patients, and it consists of 100 gene expression covariates. We separates the data into train data and test data by half. Transform the data and feed into the SVM, it gives the following results.

λ	0.01	0.1	0.3	0.5
Est. Var #	85	51	13	7
GACV	0.00849	0.00849	0.187	0.239

SVM recommends 0.1 as λ, since it gives the lowest cross validation error. Then we feed the test data in SVM with $\lambda = 0.1$. The classification for test data shows as follow

	predict label	
label	-1	1
-1	811	8
1	5	799

The simulation result shows the method can not only reduce the dimension of the data correctly,from 100 to 51, but also classify the ranking of observation pairs correctly.

5 Conclusion

Analysis of censored failure time data containing a huge number of dimensions is important in medical research, especially for currently genetic studies. How to mine these databases and identify important prognostic factors presents a class of interesting and challenging questions. In this paper, we proposed a SCAD SVM method for survival data which can estimate and select variable simultaneously. The proposed method can effectively reduce the dimension of the input variables and select important survival-associated prognostic factors while providing satisfactory estimation and prediction.

Comparing to previous SVM method on survival data, our proposed method can efficiently reduce the dimension of the dataset. Besides, traditional SVM classification survival has the limitation that it can only predict at certain time point. By classifying the rank of the observations, we can have the idea of if the interest of even will happen at any time point.

However, there is more work can be done regarding improvement and validation of the proposed method. When running simulation study, we find that increasing sample size is really a factor that limit the speed of the program. Transforming the original survival data into ranking classification result will square the sample size, makes the running time for the software, from $O(n)$ to $O(n^2)$. This constitutes our motivation to further research.

References

[1] Lawless, J. F. (2011). Statistical models and methods for lifetime data. *John Wiley & Sons.*

[2] Khan, F. M., & Zubek, V. B. (2008, December). Support vector regression for censored data (SVRc): a novel tool for survival analysis. *In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on* (pp. 863-868). *IEEE.*

[3] Shim, J., & Hwang, C. (2009). Support vector censored quantile regression under random censoring. *Computational Statistics & Data Analysis*, 53(4), 912-919.

[4] Shivaswamy, P. K., Chu, W., & Jansche, M. (2007, October). A support vector approach to censored targets. *In Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on* (pp. 655-660). *IEEE.*

[5] Shiao, H. T., & Cherkassky, V. (2013, January). SVM-based approaches for predictive modeling of survival data. *In The 2013 International Conference on Data Mining.*

[6] Van Belle, V., Pelckmans, K., Suykens, J. A. K., & Van Huffel, S. (2007, July). Support vector machines for survival analysis. *In Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare* (CIMED2007) (pp. 1-8).

[7] Zhang, H. H., Ahn, J., Lin, X., & Park, C. (2006). Gene selection using support vector machines with non-convex penalty. *Bioinformatics*, 22(1), 88-95.

[8] Cai, T., & Cheng, S. (2008). Robust combination of multiple diagnostic tests for classifying censored event times. *Biostatistics*, 9(2), 216-233.

[9] Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties.*Journal of the American statistical Association*, 96(456), 1348-1360.

[10] Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1). *Springer, Berlin: Springer series in statistics.*

[11] Becker, N., Werft, W., Toedt, G., Lichter, P., & Benner, A. (2009). penalizedSVM: a R-package for feature selection SVM classification. *Bioinformatics*, 25(13), 1711-1712.

[12] Ganzfried, B. F., Riester, M., Haibe-Kains, B., Risch, T., Tyekucheva, S., Jazic, I., ... & Huttenhower, C. (2013). curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome. Database, 2013, bat013.