Chapter 2
Efficient Support Vector Machine Method for Survival Prediction with SEER Data

Zhenqiu Liu, Dechang Chen, Guoliang Tian, Man-Lai Tang, Ming Tan, and Li Sheng

Abstract Support vector machine (SVM) is a popular method for classification, but there are few methods that utilize SVM for survival analysis in the literature because of the computational complexity. In this paper, we develop a novel L_1 penalized SVM method for mining right-censored survival data (L_1 SVMSURV). Our proposed method can simultaneously identify survival-associated prognostic factors and predict survival outcomes. It is easy to understand and efficient to use especially when applied to large datasets. Our method has been examined through both simulation and real data, and its performance is very good with limited experiments.

Keywords Support vector machine · SVM · Survival analysis · Prognostic factors · SEER

2.1 Introduction

Survival prediction and prognostic factor identification play an important role in medical research. Survival data normally include a variable indicating whether some outcome under consideration (such as death or recurrence of a disease) has occurred within a specific follow-up time. The modeling technique has to consider that for some patients the follow-up may end before the event occurs. In other words, we must take into account patients for whom the event has not occurred during the follow-up period but might have occurred just after it. This makes it more difficult to apply a standard machine learning method for survival prediction.

Z. Liu
University of Maryland at Baltimore, Baltimore, MD, USA
e-mail: zliu@umm.edu

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978-1-4419-5913-3_2,

Many models for survival prediction have been proposed in the statistical literature. However, most of them are designed for small datasets and not suitable for large data mining. The most popular one is Cox proportional hazards model [1, 2, 7, 14], in which model parameters are estimated with partial log likelihood maximization. Another one is the accelerate failure time (AFT) model [5, 15, 16]. AFT is a linear regression model in which the response variable is the logarithm or a known monotone transformation of a failure (death) time. Even though the semi-parametric estimation of an AFT model with an unspecific error distribution has been studied extensively in the literature, the model has not been widely used in practice, mainly due to the difficulties in computing model parameters [4]. Recently, AFT has been applied to gene expression data with a small size but a large dimension [9, 11]. Survival prediction with the area under the ROC curve (AUC) maximization has also been proposed [8] for high-dimensional gene expression data with a small size. However, it is much more difficult to apply these methods for survival data with a large size and a high dimension.

The size in a retrospective surveillance, epidemiology and end results (SEER) dataset is usually very large. For example, the lung cancer data set from SEER that contains the records of lung cancer patients who were diagnosed from 1973 to 2002 has more than 500,000 patients; and the breast cancer data set also has more than 500,000 patients at the same time span. Cancer data also contain a high percentage of censored observations. With the above mentioned lung cancer data set, for instance, 34% of total records involve censored times. Ignoring records that contain censored observations or treating censored data as the actual life-times will produce biases in survival modeling. Traditional statistical methods fail to deal with large survival data sets. New methods are required for mining the SEER data efficiently.

Though there are many publications for L_1SVM in the classification framework [10], there are few SVM methods for survival analysis because of the computational complexity. In this paper, we propose a novel L_1 [12, 13] SVM approach for survival outcome predictions (L_1 SVMSURV). At the same time, L_1 SVMSURV is utilized to automatically identify survival-associated prognostic factors. The proposed models are evaluated with simulation and real data using the global AUC summary (GAUCS) measure [3]. The paper is organized as follows. In Sect. 2.2, we formulate L_1 SVMSURV under the accelerate failure time framework and introduce the GAUCS measure. Experiments with simulation and real survival data for model performance evaluation are given in Sect. 2.3. Finally, conclusions and remarks are provided in Sect. 2.4.

2.2 L_1 SVMSURV for Censored Survival Outcomes

Consider a set of n independent observations $\{T_i, \delta_i, \mathbf{x}_i\}_{i=1}^n$, where δ_i is the censoring indicator and T_i is the survival time (event time) if $\delta_i = 1$ or censored time if $\delta_i = 0$, and $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})^t$ is the m-dimensional input vector of the ith sample. Let

$\mathbf{w} = (w_1, w_2, \ldots, w_m)^t$ be a vector of regression coefficients and $\phi(\mathbf{x}_i)$ be the nonlinear transformation of $\mathbf{x}_i$ in the feature space. The AFT model is defined as

$$M(\mathbf{x}_i) = \mathbf{w}^t \phi(\mathbf{x}_i), \quad i = 1, \ldots, n, \tag{2.1}$$

where $M(\mathbf{x}_i) > logT_i$ if $\delta_i = 0$ and $M(\mathbf{x}_i) = logT_i$ if $\delta_i = 1$. Because there are both equality and inequality constraint in the model, simple least square solution will fail to work. We use $L_1 = \sum_{i=1}^{n} |w_i|$ as the penalty for the sparse solution and have the quadratic L_1 SVMSURV:

$$\begin{gathered} min \frac{1}{2n} \sum_{i=1}^{n} \xi_i^2 + \lambda \sum_{i=1}^{n} |w_i| \\ \text{s.t. } |\mathbf{w}^t \phi(\mathbf{x}_i) - logT_i| < \xi_i, \quad \text{if} \quad \delta_i = 1, \\ \mathbf{w}^t \phi(\mathbf{x}_i) > logT_i - \xi_i, \quad \text{if} \quad \delta_i = 0 \\ \xi_i \geqslant 0, \quad \forall\, 1 \leqslant i \leqslant n. \end{gathered} \tag{2.2}$$

When ties in the event times are presented, variables associated with each tied time appear in the constraints independently. We can define an index function $I(\delta_i) = 1$ if $\delta_i = 1$, and I is defined as $I(\delta_i) = 1$ if $logT_i \geqslant \mathbf{w}^t \phi(\mathbf{x}_i)$ and 0, otherwise, when $\delta_i = 0$. Then, we have

$$J(\mathbf{w}; \lambda) = \frac{1}{2n} \sum_{i=1}^{n} I(\delta_i) \{\mathbf{w}^t \phi(\mathbf{x}_i) - logT_i\}^2 + \lambda \sum_{i=1}^{n} |w_i|. \tag{2.3}$$

Since $|w_i|$ does not have the first order derivative at 0, we will use the similar procedure proposed by Liu et al. [8] for parameter estimation. Rewrite $J(\mathbf{w}, 0)$ as a function of the kth parameter w_k, and let the remaining parameters $\mathbf{w}_{-k}$ be fixed. We have

$$\begin{aligned} J(\mathbf{w}; 0) &= \frac{1}{2n} \sum_{i} I(\delta_i) (\mathbf{w}^t \phi(\mathbf{x}_i) - logT_i)^2, \\ &= \frac{1}{2} b_k w_k^2 + c_k w_k + d_k, \end{aligned} \tag{2.4}$$

where

$$\begin{aligned} b_k &= \tfrac{1}{n} \sum_{i} I(\delta_i) \phi_k^2(\mathbf{x}_i), \\ c_k &= \tfrac{1}{n} \sum_{i} I(\delta_i) (logT_i - \mathbf{w}_{-k}^t \Phi_{-k}(\mathbf{x}_i(\mathbf{x}_i))), \\ d_k &= \tfrac{1}{2n} \sum_{i} I(\delta_i) (logT_i - \mathbf{w}_{-k}^t \Phi_{-k}(\mathbf{x}_i))^2. \end{aligned}$$

Equation (2.4) is a quadratic function of w_k, and we have the following first order derivative w.r.t. w_k:

$$\frac{\partial J(\mathbf{w};0)}{\partial w_k} = b_k w_k + c_k.$$

Since $\mathbf{w}$ is not differentiable at 0, the first derivative of $J(\mathbf{w},\lambda)$ is a step function:

$$\partial_{w_k} J(\mathbf{w};\lambda) = \begin{cases} \{(b_k w_k - c_k) - \lambda\}, & w_k < 0 \\ [-c_k - \lambda, -c_k + \lambda], & w_k = 0 \\ \{(b_k w_k - c_k) + \lambda\}, & w_k > 0 \end{cases} \tag{2.5}$$

We therefore can update each coefficient w_i with the following equation:

$$w_k(c_k) = \begin{cases} (\lambda + c_k)/b_k, & c_k < -\lambda \\ 0, & c_k \in [-\lambda, \lambda] \\ (-\lambda + c_k)/b_k, & c_k > \lambda \end{cases} \tag{2.6}$$

The model performance is evaluated by the GAUCS measure. In survival analysis, the AUC is a time dependent measure. We will utilize the GAUCS measure for the comparison of model performance. The GAUCS is defined by averaging over t: $\text{GAUCS} = 2\int AUC(t)g(t)S(t)\,\mathrm{d}t = Pr(M_j > M_k | t_j < t_k)$, which indicates the probability that the subject who died (case) at an earlier time has a larger value of the risk score. In the above equation, $S(t)$ and $g(t)$ are the survival and corresponding density functions, respectively.

2.3 Computational Results

Simulation data: Simulation studies were conducted to evaluate the performance of the proposed method under different assumptions. 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 $\mathbf{x}$ with 10,000 training and test samples, respectively, from a multivariate normal distribution with zero means and variance–covariance matrix Σ. Σ is set to have the same correlation coefficient ρ for all input variables and different $\rho = 0, 0.2, 0.4, 0.6,$ and, 0.8 will be chosen to assess the performance of the proposed method. We then choose model parameters $\mathbf{w} = [-1.9, 2.8, 1.7, 0, 0, 0, 0, 0, 0, 0, 0, 0]^T$ and calculate $Z = \mathbf{w}^t\mathbf{x} + \varepsilon$. Hence, this model is only associated with the first three input variables plus random noise. Finally, we compute $H = Aexp(-0.5Z)$ with $A = 100$, sample the survival time T_i from Weibull random number generating function, and build $d_i = (\text{rand}(1,1) + C) * T_i$, such that the censoring status is $\delta_i = T_i < d_i$. Different Cs give different portions of censored data.

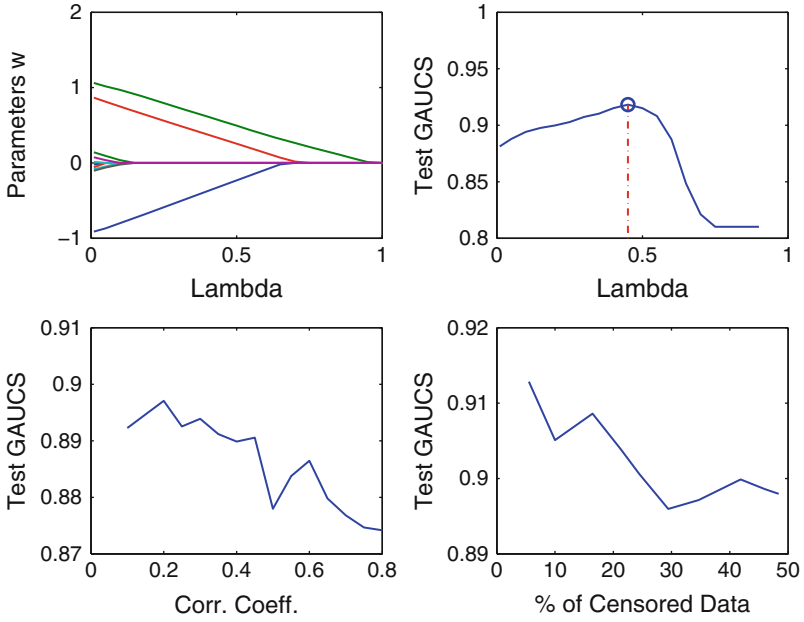

Fig. 2.1 Regularization path and test GAUCS

We analyze the simulation data with L_1 SVMSURV 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 tenfold cross-validation with training data only. To prevent bias arising from the specific data, we simulate the data 100 times. The average computational results are reported in Fig. 2.1.

The upper left subfigure is the regularization path with different λs. It shows that only three nonzero w_is are left when $\lambda \geqslant 0.1$ and all w_is go to zero when $\lambda = 1$. The upper right subplot shows that the value of the optimal test GAUCS is 0.9183 with $\lambda = 0.45$ for simulation data with no censoring. Combining both upper subplots, we conclude that our proposed method correctly selects the three variables. The lower left subplot shows the average value of the test GAUCS with different correlation coefficients ρ among input variables. As correlations among input variables increase from 0.1 to 0.8, the average value of the test GAUCS over 100 test data sets decreases from 0.898 to 0.873. Finally, the lower right subplot shows that the value of the test GAUCS also decreases with the increase of the percent of censored data, but the decrease is not statistical significant, which indicates that the proposed model is robust.

SEER prostate data: Prostate cancer is the most common cancer, other than skin cancers, and the second leading cause of cancer death in American men, behind only lung cancer. In this paper, we study the SEER prostate cancer registry data from 2000 to 2003 in the states of Greater California, Kentucky, Louisiana, and New Jersey. There are in total 104,363 patients in the database. For the comparison

purpose, we only study 13,975 samples which do not contain missing tumor size information. Fifteen potential prognostic factors are included in the model, including age, race, married status, grade, size of tumor, clinical extension of tumor, lymph node involved, number of positive nodes examined, number of nodes examined, surgery performed, radiation, radiation sequence with surgery, stage (TNM), PSA marker, and number of primaries. We divide the data into training and test data with a roughly equal size. The regularization parameter is again determined by tenfold cross-validation. To prevent bias and overfitting from a specific grouping, we partition the data 100 times and the average λ and test GAUCS values are shown in Fig. 2.2.

The left subplot shows that the optimal test GAUCS $= 0.7707$ is reached at $\lambda = 0.05$. There are five prognostic factors associated with survival, i.e., age (0.179 ± 0.002), tumor size (-0.002 ± 0.0003), the clinical extension of tumor (-0.001 ± 0.0002), radiation (0.1 ± 0.007), and stage (-0.06 ± 0.0008), where the values in the parenthesis are the mean and standard deviation of w_i for each prognostic factor. They indicate that patients who have cancer in their later age and those who have been exposed to radiations may survive longer and patients with a large tumor size, clinical extension of tumor, and late stage may die earlier. Our conclusion that younger men with prostate cancer have shorter survival times is consistent with the finding in [6]. While the reasons for this unexpected but excited finding are not clear, one explanation may be that young men with prostate cancer may have biologically more aggressive forms of the disease than the forms

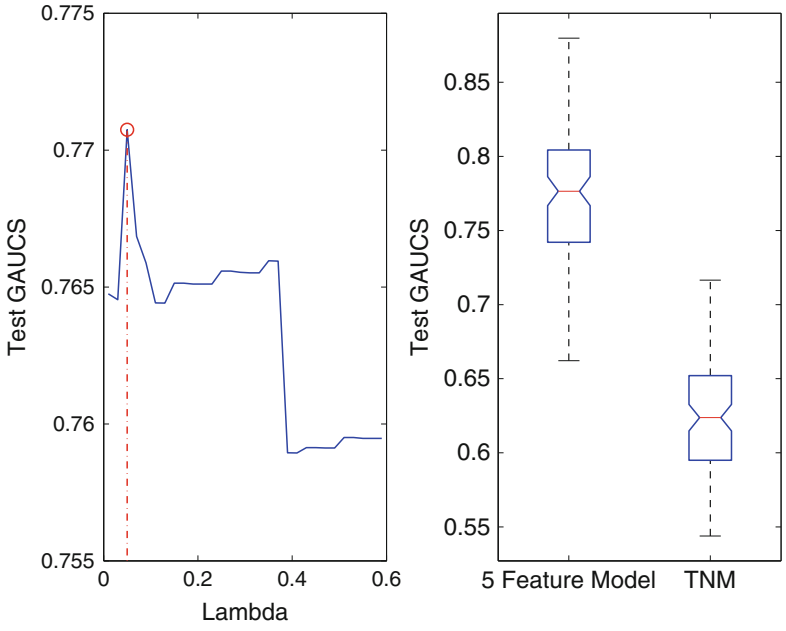

Fig. 2.2 Average test GAUCS with SEER data

diagnosed in older men. Additional studies are needed to determine what, if any, underlying differences exist between prostate cancer found in young men and that found in older men. These studies may help clinicians improve screening in young men and could ultimately lead to the development of better treatment strategies for young patients. Finally, the right subplot shows the performance comparison of our proposed model with the TNM stage system. It is seen that the average value of the test GAUCS increases roughly 22% from 0.62 to 0.77.

2.4 Conclusions and Remarks

Analysis of censored failure time data containing a huge number of samples is important in practice, especially for data containing millions of patients' records. How to mine these databases and identify important prognostic factors presents a class of interesting and challenging questions. In this paper, we propose a L_1 penalized SVM method for simultaneous variable selection and estimation. The simulation studies and the real example on prostate cancer illustrate that 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. More work can be done regarding improvement and validation of the proposed method. This constitutes our future work. For example, the asymptotic properties of the model will be studied in our future work, and we will apply the proposed method to other cancer data sets.

Acknowledgment This work was partially supported by NIH Grant 1R03CA133899-01A210 and NSF CCF-0729080.

References

1. Cox DR (1972). Regression models and life-tables (with discussion). Journal of Royal Statistical Society, Series B 34:187–220.
2. Gui J, Li H (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics 21:3001–3008.
3. Heagerty PJ, Zheng Y (2005). Survival model predictive accuracy and ROC curves. Biometrics 61(1):92–105.
4. Jin Z, Lin DY, Wei LJ, Ying ZL (2003). Rank-based inference for the accelerated failure time model. Biometrika 90:341–353.
5. Kalbfleisch JD, Prentice RL (1980). The Statistical Analysis of Failure Time Data. New York: John Wiley.
6. Lin DW, Porter M, Montgomery B (2009). Treatment and survival outcomes in young men diagnosed with prostate cancer: a Population-based Cohort Study. Cancer 115 (13):2863–2871.
7. Liu Z, Jiang F (2009). Gene identification and survival prediction with Lp penalty and novel similarity measure. International Journal of Data Mining and Bioinformatics 3(4):398–408.

8. Liu Z, Gartenhaus RB, Chen X, Howell C, Tan M (2009). Survival prediction and gene identification with penalized global AUC maximization. Journal of Computational Biology 16(12):1661–1670.
9. Ma S, Huang J (2007). Additive risk survival model with microarray data. BMC Bioinformatics 8:192.
10. Mangasarian OL (2006). Exact 1-norm support vector machines via unconstrained convex differentiable minimization. Journal of Machine Learning Research 7:1517–1530.
11. Sha N, Tadesse MG, Vannucci M (2006). Bayesian variable selection for the analysis of microarray data with censored outcomes. Bioinformatics 22(18):2262–2268.
12. Tibshirani R (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58(1):267–288.
13. Tibshirani R (1997). The lasso method for variable selection in the Cox model. Statistics in Medicine 16(4):385–395.
14. Van Houwelingen HC, et al. (2006). Cross-validated Cox regression on microarray gene expression data. Statistics in Medicine 25:3201–3216.
15. Wei LJ (1992). The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. Statistics in Medicine 11:1871–1879.
16. Ying ZL (1993). A large sample study of rank estimation for censored regression data. Annals of Statistics 21:76–99.

http://www.springer.com/978-1-4419-5912-6

Advances in Computational Biology
Yen, N. (Ed.)
2010, XVII, 759 p., Hardcover
ISBN: 978-1-4419-5912-6