Feature Selection in Survival Least Squares Support Vector Machines with Maximal Variation Constraints

V. Van Belle *, K. Pelckmans, J.A.K. Suykens, and S. Van Huffel

Katholieke Universiteit Leuven, ESAT-SCD
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
{vanya.vanbelle,kristiaan.pelckmans,johan.suykens,sabine.vanhuffel}
@esat.kuleuven.be

Abstract. This work proposes the use of maximal variation analysis for feature selection within least squares support vector machines for survival analysis. Instead of selecting a subset of variables with forward or backward feature selection procedures, we modify the loss function in such a way that the maximal variation for each covariate is minimized, resulting in models which have sparse dependence on the features. Experiments on artificial data illustrate the ability of the maximal variation method to recover relevant variables from the given ones. A real life study concentrates on a breast cancer dataset containing clinical variables. The results indicate a better performance for the proposed method compared to Cox regression with an L_1 regularization scheme.

Key words: failure time data, feature selection, LS-SVM

1 Introduction

Survival analysis studies the time until a certain event is observed. A typical problem within these studies is the presence of censored data, or data for which the exact failure time is not observed exactly. The most common censoring scheme is right censoring. In this case the event time is known to be later than the last time for which information is available. A second type of censoring is left censoring, occurring in cases where it is known that the failure occurred before a certain time. Another type of censoring is interval censoring, a combination of left and right censoring. This type is often seen in clinical studies in which patients are scheduled for regular check-ups.

An important goal within the analysis of survival data is the construction of prognostic indices. A prognostic index is a scoring function indicating the risk of each patient (in medical studies) or component (in electrical studies). The construction of a prognostic index is associated with a financial cost and

* Research supported by GOA-AMBioRICS, CoE EF/05/006, FWO G.0407.02 and G.0302.07, IWT, IUAP P6/04, eTUMOUR (FP6-2002-LIFESCIHEALTH 503094)

effort depending on the number of covariates used within the index and the difficulty of obtaining that covariate. By selecting the most relevant variables, this cost can be reduced. Due to this fact and the increasing availability of high dimensional data, as there are genomics and proteomics data [1], feature selection becomes more and more an issue in current data analysis. Different feature selection procedures have been proposed [2,3,4,5], and can be categorized as filter or wrapper methods [6]. Filter methods select variables independent to the predictor, whereas wrapper methods are methods for which the selection procedure is related to the predictor performance. Most feature selection methods are based on forward selection or backward elimination processes. Forward selection starts with the most relevant feature and adds the best feature in each step. The subset of features is no longer extended once the addition of an extra feature does not improve the performance. In backward selection the search is started with the largest set of variables and the least promising one is eliminated in each step. Disadvantages of these methods are the need to train the model for each subset of selected variables and their large variability: small changes in the data can result in the selection of different subsets of variables. An alternative approach to reveal the most relevant feature is found in LASSO [2], which shrinks some coefficients and sets others to 0. To obtain zeros for certain coefficients the sum of the absolute values of the coefficients is constrained to be less than a certain constant.

In this paper we select variables as proposed in [7,8] and investigate it for a model in survival analysis. Instead of constraining the sum of the coefficients, the maximal variation of each component is minimized. The reason for this is that non-relevant variables will result in very small maximal variations, whereas the variation for relevant variables will remain large. The goal of this paper is to combine the above feature selection procedure with least squares support vector machines (LS-SVM) [9,10] designed for survival analysis [11], resulting in a convex quadratical optimization (QP) problem.

This work is organized as follows. In Section 2 we summarize the principle of survival LS-SVM. Afterwards this model is adapted towards feature selection. Section 3 describes results obtained on artificial and clinical datasets. Section 4 concludes this work.

2 Additive Survival Least Squares LS-SVM

In previous work we presented a Least Squares Support Vector Machine (LS-SVM) as a flexible, non-linear, kernel based model for survival analysis [11], which builds further on the work presented in [12,13]. The model is built from the idea that in practice one is interested in the ranking of samples according to their risk on experiencing the event. The final goal is to create a model for which the predicted risk $u(x)$ correlates with the observed failure times t. To measure the concordance between the estimated risk and the observed failure time, the concordance index (c-index) [14] is used:

$$CI(u) = \frac{\sum_{i=1}^{n_t} \sum_{j \neq i} \Delta(x_i, t_i, \delta_i; x_j, t_j, \delta_j) I[(u(x_j) - u(x_i))(t_j - t_i) > 0]}{\sum_{i=1}^{n_t} \sum_{j \neq i} \Delta(x_i, t_i, \delta_i; x_j, t_j, \delta_j)}, \quad (1)$$

where n_t is the number of test points, x_i, t_i and δ_i are the covariate vector, failure time and censoring indicator (1 for an observed event and zero for right censored data) of sample i and $I[z] = 1$ if $z > 0$, and zero otherwise. $\Delta(x_i, t_i, \delta_i; x_j, t_j, \delta_j)$ indicates whether the observations i and j are comparable and depends on the censoring mechanism present in the data. Without censoring $\Delta(x_i, t_i, \delta_i; x_j, t_j, \delta_j)$ equals 1. For right censoring $\Delta(x_i, t_i, \delta_i; x_j, t_j, \delta_j)$ equals zero if $t_i < t_j$ & $\delta_i = 1$ or $t_j < t_i$ & $\delta_j = 1$, and zero otherwise.

2.1 Pairwise Kernel Machine

A possible approach for maximizing the c-index with regard to $u(x)$ is as follows [11]:

$$\boxed{\text{P1}} \quad \min_{w, \xi_{ij}} \ \tfrac{1}{2} w^T w + \tfrac{1}{2} \gamma \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_{ij} \xi_{ij}^2$$
$$\text{s.t.} \ \ w^T \varphi(x_j) - w^T \varphi(x_i) = 1 + \xi_{ij}, \ \forall \, i, j = 1, \ldots, n, \quad (2)$$

where $x_i \in \mathbb{R}^d$ represents the feature vector for the ith datapoint, $w \in \mathbb{R}^{n_\varphi}$ is an unknown vector for the model $u(x) = w^T \varphi(x)$, with $\varphi(\cdot) : \mathbb{R}^d \to \mathbb{R}^{n_\varphi}$. n is the number of datapoints, $\gamma > 0$ a regularization constant, ξ_{ij} slack variables allowing for incorrect rankings and $\Delta_{ij} = \Delta(x_i, t_i, \delta_i; x_j, t_j, \delta_j)$. Here it is assumed that $t_j > t_i$ for $j > i$ and the 1 at the right hand side is used as a target value. To visualize the effect of one covariate on the estimated risk, it can be shown how risk changes when varying a single covariate while keeping the rest constant. However, this only gives information on the effect of the covariate on the ranking of the failure time and not on the failure time itself. To obtain the latter information, an extra constraint is added and a componentwise kernel is used [11]:

$$\boxed{\text{P2}} \quad \min_{w, \xi, b, \chi} \ \tfrac{1}{2} \textstyle\sum_{p=1}^{d} w_p^T w_p + \tfrac{1}{2} \gamma \displaystyle\sum_{i=1}^{n} \sum_{j=1}^{n} \Delta_{ij} \xi_{ij}^2 + \frac{1}{2} \mu \sum_{j=1}^{n} \chi_j^2$$

$$\text{s.t.} \ \begin{cases} \sum_{p=1}^{d} w_p^T \varphi_p(x_{j(p)}) - \sum_{p=1}^{d} w_p^T \varphi_p(x_{i(p)}) = 1 + \xi_{ij}, \forall \, i, j = 1, \ldots, n \\ \delta_j t_j = \delta_j (\sum_{p=1}^{d} w_p^T \varphi_p(x_{j(p)}) + b) - \chi_j, \qquad \forall \, j = 1, \ldots, n, \end{cases} \quad (3)$$

where $x_{i(p)}$ represents the pth covariate of the ith datapoint, d is the number of covariates, $w_p \in R^{n_{\varphi_p}}$ represents the unknown vector of the pth covariate in the model and $\varphi_p(x_{(p)})(\cdot) : \mathbb{R} \rightarrow \mathbb{R}^{n_{\varphi_p}}$ represents the feature map corresponding to the pth covariate $x_{(p)}$. $\gamma > 0$ and $\mu > 0$ are two regularization constants and ξ_{ij} and χ_j are slack variables allowing for incorrect rankings and regression errors, respectively.

In [11] we proposed to use the above method to estimate functional forms of covariates in a univariate setting and to combine these terms linearly in order to obtain an interpretable utility function (Figure 1(a)). However, certain covariates used in this model, are possibly of little or no importance to the development of an optimal utility function. The model is therefore adapted to incorporate the selection of relevant variables, as illustrated in Figure 1(b).

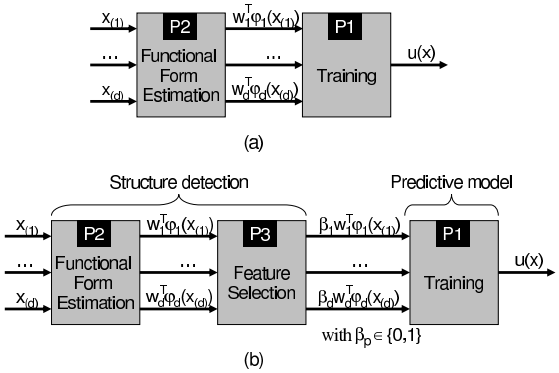


Fig. 1. Illustration of the training phase. **(a)** In previous work we proposed to use model P2 to estimate functional forms of covariates in a first layer and to combine these estimated linearly in a second layer with model P1. **(b)** To select relevant features a third layer is introduced, in which a new model (P3) is used for feature selection. The training of the final prognostic index $u(x)$ is then done only incorporating the selected features.

2.2 Feature Selection

The approach proposed in [7] uses a componentwise model (as we do, see equation (3)) and additionally penalizes large variations for each component. This results in variations which are very low for non-relevant variables and larger for relevant variables. In the methodology followed in our previous work [11] we use model P2 with a polynomial kernel $K(x, z) = (\nu + x^T z)^d, \nu \geq 0$, in a componentwise way, to estimate the functional form for each covariate in a first step (Figure 1). In a second step, model P1 with a linear kernel is used to create the prognostic index as a linear combination of the estimated functional forms. This work concentrates on the adaptation of this second layer, where we not only

want to produce a prognostic index, but we want to do so with a small subset of covariates. Therefore the sum of the maximal variations of all components is added to the loss function. Additional constraints indicate the restriction on the componentwise maximal variations. The model formulation then becomes:

$$\boxed{\text{P3}} \min_{w,\xi,m} \tfrac{1}{2}\sum_{p=1}^{d} w_p^T w_p + \tfrac{1}{2}\gamma \sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_{ij}\xi_{ij}^2 + \mu \sum_{p=1}^{d} m_p$$

$$\text{s.t.} \begin{cases} \sum_{p=1}^{d} w_p^T \left(\varphi_p(x_j) - \varphi_p(x_i)\right) = 1 + \xi_{ij}, \forall\ i,j = 1,\ldots,n \\ -m_p \leq w_p^T \varphi_p(x_{i(p)}) \leq m_p, \qquad\qquad \forall\ i = 1,\ldots,n,\ \forall\ p = 1,\ldots,d, \end{cases} \tag{4}$$

where m_p is the variation of the pth covariate. The prognostic index for a datapoint with covariates x is defined as $u(x) = \sum_{p=1}^{d} w_p^T \varphi_p(x_{(p)})$, where $\varphi_p(x_{(p)}) = x_{(p)}$ for a linear kernel.

This problem is a quadratic programming problem and can therefore be solved efficiently. In our application, model P3 is used in a linear setting, since all non-linear effects are estimated in the first layer of the model. However, the formulation can be solved in the dual form for applications with other kernels.

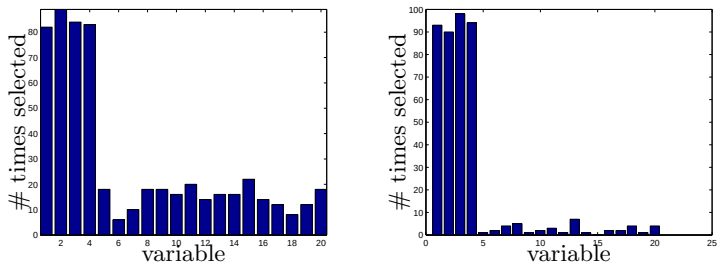
3 Results

This section summarizes results on artificial and clinical datasets. The artificial data shows a clear difference in maximal variation for relevant versus non-relevant variables. Applying this approach to a dataset with clinical features [15] on breast cancer results in a better prognostic index then when using L_1 regularization with Cox' proportional hazard model [16]. Model selection was performed using 10-fold cross-validation. The model selection criterion was the concordance of new samples relative to the training samples, as defined in equation (1). The tuning parameters $\gamma(P1)$, $\gamma(P2)$, $\gamma(P3)$, $\mu(P2)$ and $\mu(P3)$ where found using 10-fold cross-validation combined with a grid search, where CI^u was used as model selection criterion.

3.1 Artificial Data

In a first example we generated 100 datasets, each containing 100 training points, 100 test points and 20 variables of which only 4 contributed to the survival time. All covariates were normally distributed and the survival time was Weibull distributed, depending on the covariates as $\sum_{p=1}^{20} w_p x_{(p)}$ where w_p was a randomly chosen value for $p = 1,\ldots,4$ and 0 otherwise. Conditional independent censoring is added as follows: the censoring time is distributed exponentially, dependent on the first covariate. A point is considered to be censored in case the survival time was higher than the censoring time.

When only considering model P3, Figure 2(a) shows the frequency at which variables were selected for each of the 100 models. Features for which the maximal variation was larger than one fifth of the largest maximal variation were selected for each model. We clearly see that the relevant features are selected for nearly every model. The non-relevant variables are only sporadically selected. Results are further improved after 10-fold cross-validation (Figure 2(b)).



(a) (b)

Fig. 2. (a) Frequency of selection for each variable in an artificial example where only the first 4 covariates contribute to the failure time (Weibull distributed). 100 models were trained with randomly chosen weights for the first 4 covariates. Variables with a maximal variation larger than one fifth of the largest maximal variation were selected. The relevant features are selected in most models, whereas the non-relevant features are only sporadically selected. **(b)** Additional 10-fold cross-validation: features selected in more than 8 folds were retained. On all 100 models, the relevant features were significantly more often selected.

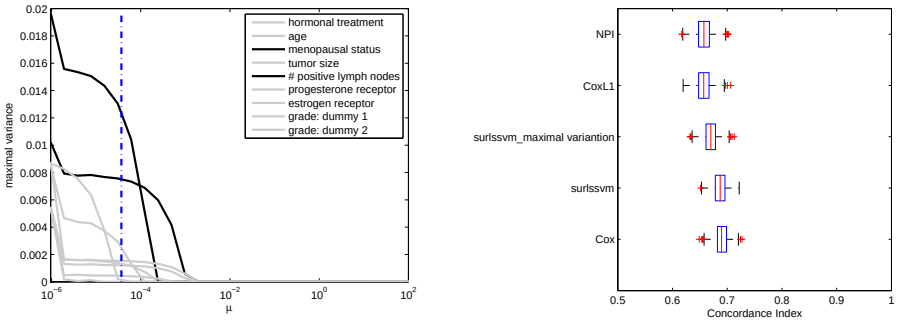
3.2 Breast Cancer Dataset

This example illustrates the selection of a subset of clinical variables on the German Breast Cancer Study Group data [15][1], containing information on 686 patients and 8 variables. Available variables are: hormone treatment, age, menopausal status, tumor size, tumor grade, the number of positive lymph nodes, the progesterone receptor (fmol) and the estrogen receptor (fmol). Two third of the data were used for training, the rest for testing the models. 299 (43.6%) patients had a breast cancer related event within the study time, leaving all other patients with a right censored failure time.

We compare the selected variables and the performances of our model (P1-P3-P3) (Figure 1(b)) with an L_1 regularization scheme with Cox regression (CoxL1) and the Nottingham Prognostic Index (NPI), a linear prognostic model used in clinical practice. Figure 3(a) illustrates how the variation of the parameter μ influences the maximal variation for the different components. The vertical line indicates the optimal value of this parameter. A clear difference in variation between the black and gray variables is noted at this point. According to our model

[1] `http://www.blackwellpublishers.com/rss/Volumes/A162p1.htm`

the number of positive lymph nodes and whether the patient received hormonal treatment or not, are relevant features. When using CoxL1, the variables with a non-zero coefficient are the number of positive lymph nodes, the progesterone receptor and the grade of the tumor. The NPI on the other hand considers the tumor size, grade and the number of positive lymph nodes as relevant variables. All kernel-based models used a polynomial kernel to fit non-linearities. Figure 3(b) compares performance of different models: Cox (Cox) and a two layer model (P2-P1) (Figure 1(a)) (SURLSSVM) using all covariates, CoxL1 and our presented model (P2-P3-P1) (SURLSSVM_maximal variation). The models with all variables performs best. The LSSVM-based model with a subset of variables performs better than CoxL1.



(a) (b)

Fig. 3. Feature selection with SURLSSVM. **(a)** Influence of μ on the maximal variation for each component, for optimal γ. At the optimal value of μ (vertical line) hormonal treatment and the number of positive lymph nodes are selected as relevant variables. **(b)** Comparison of performances of Cox and SURLSSVM(P2-P1), CoxL1 and model SURLSSVM_maximal variation (P2-P3-P3). The models with all variables performs best. The LSSVM-based model with a subset of variables performs better than CoxL1.

4 Conclusions

In this work we presented a method to select relevant features in survival analysis within an LS-SVM based model. Results on an artificial dataset show the selection of relevant variables and the rejection of non-relevant variables. In the clinical dataset used in this paper, the proposed method performs better than Cox regression with L_1 norm and the the NPI, which is used in clinical practice. Both the NPI and Cox-L1 use three different variables, whereas our method selected only two variables. Although further research is necessary, these preliminary results are promising.

References

1. Bøvelstad H. M. M., Nygård S., Størvold H. L. L., Aldrin M., Borgan O., Frigessi A., and Lingjærde O. C. C. *Bioinformatics*, 23(16):2080–2087.
2. Tibshirani R. Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.
3. Guyon I. and Elisseeff A. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, March 2003.
4. Rakotomamonjy A. Variable selection using SVM-based criteria. *Journal of Machine Learning Research*, 3:1357–1370, 2003.
5. Ojeda F., Suykens J.A.K., and De Moor B. Low rank updated LS-SVM classifiers for fast variable selection. *Neural Networks*, 21(2-3):437–449, 2008.
6. Kohavi R. and John G.H. . Wrappers for feature subset selection. *Artif. Intell.*, 97(1-2):273–324, 1997.
7. Pelckmans K., Suykens J.A.K., and De Moor B. Componentwise support vector machines for structure detection. In *Artificial Neural Networks: Formal Models and Their Applications - ICANN 2005*. Springer Berlin / Heidelberg, 2005.
8. Pelckmans K., Goethals I., De Brabanter J., Suykens J.A.K., and De Moor B. *Componentwise Least Squares Support Vector Machines*, chapter Support Vector Machines: Theory and Applications, pages 77–98. (Wang L., ed.), Springer, 2005.
9. Suykens J.A.K. and Vandewalle J. Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300, 1999.
10. Suykens J.A.K., Van Gestel T., De Brabanter J., De Moor B., and Vandewalle J. *Least Squares Support Vector Machines.* World Scientific, Singapore, 2002.
11. Van Belle V., Pelckmans K., Suykens J.A.K., and Van Huffel S. Additive survival least squares support vector machines. Technical report, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2008, submitted for publication.
12. Van Belle V., Pelckmans K., Suykens J.A.K., and Van Huffel S. Support Vector Machines for Survival Analysis. In *Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007)*, Plymouth (UK), July 2007.
13. Van Belle V., Pelckmans K., Suykens J.A.K., and Van Huffel S. Survival SVM: a Practical Scalable Algorithm. In *Proceedings of the 16th European Symposium on Artificial Neural Networks (ESANN08)*, pages 89–94, Bruges (Belgium), April 2008.
14. Harrell F.E., Lee K.L., and Pollock B.G. Regression models in clinical studies: Determining relationships between predictors and response. *Journal of the National Cancer Institute*, 80, 1988.
15. Schumacher M., Basert G., Bojar H., Huebner K., Olschewski M., Sauerbrei W., Schmoor C., Beyerle C., Neumann R.L.A., and Rauschecker H.F. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *Journal of Clinical Oncology*, 12, 1994.
16. Cox D.R. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society*, 34:197–220, 1972.