SUPPORT VECTOR MACHINES FOR SURVIVAL ANALYSIS

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

Katholieke Universiteit Leuven, Department of Electrical Engineering (ESAT), Division SCD, Belgium
{vanya.vanbelle, kristiaan.pelckmans, johan.suykens, sabine.vanhuffel}@esat.kuleuven.be

Abstract: This paper studies a learning machine designed for predictive modeling of independently right censored, and observed failure time data. This is achieved through the introduction of a *health* index which serves as a proxy between the subject's covariates and the observed failure times. The corresponding risk is defined in terms of the concordance index (c-index) measuring the predictive performance and discriminative power of the health index with respect to failure times. The resulting machine and its nonlinear kernel version are derived and are related to the literature on Support Vector Machines (SVMs) and ranking SVMs. Artificial datasets and an observational study of heart attack data are used to illustrate ideas. Issues of computability and regarding the interpretability of the resulting model are explored.

Keywords: Survival analysis, concordance index, support vector machines, kernel methods.

INTRODUCTION

The statistical modeling of failure time data constitutes one of the principal research tracks in theoretical and applied statistics (see e.g. [1] for a good survey). On the other hand, research on machine learning and pattern recognition complements in many areas classical results in statistics (see e.g. [2]). In the latter, one takes a quite different perspective on the problem of inference: instead of trying to uncover the generating probabilistic mechanism (distribution) underlying the data, the overall objective now is to learn a predictive rule based on the observed samples which will *generalize* well to future samples [3, 4]. This paper attempts to integrate this view in an approach for predictive modeling of failure time data. A particular fruitful area in machine learning is the research on Support Vector Machines (SVMs) and kernel methods (see e.g. [5, 6] for an overview). Main advantages of this methodology are the integration of the framework in a context of statistical learning theory [4], convex optimization [7], and the straightforward extension to nonlinear models making use of Mercer kernels [5]. In addition to these theoretical and methodological advantages, countless case studies have shown the excellent performance in different areas.

The aim of this work is the design and analysis of a learning machine in the context of failure time data which characterizes the occurrence of failures of new subjects as well as possible, based on a set of the corresponding covariates. Therefore, one seeks an appropriate function of the covariates - denoted here as the health index in analogy to medical indices as for example the Body Mass Index (BMI) - extracting all information of the covariates which appear to be useful for explaining the observed failure times. The use of a specific health index is characterized using the c-index [8], capturing the concordance between the health index and the observed failure times. Kendall's τ as reviewed in [9] and the closely related c-index [8] are found to be a proper choice for measuring the discriminative and predictive power of two random variates, closely related to the Mann-Whitney statistic, the AUC and Somers d statistic [10]. The relation with the AUC permits the translation of recent advances in machine learning in a context of ordinal regression and ranking [11, 12, 13] and information retrieval [14, 15]. The c-index is particularly useful in the case of right censoring [8, 16], and was already used for designing an optimal neural network in [17].

This paper is organized as follows. Next section discusses the formal setup of this paper. A discussion on the resulting learning machine for extracting the optimal health index based on censored observations follows. Finally results on artificial and real datasets are reported.

PREDICTIVE MODELING OF FAILURE TIME DATA

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 (X, T) is a random vector with a joint distribution function F_{XT}. 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. Ties can be handled as in e.g. [15]. Let $C > 0$ denote the time of censoring following F_C, which is assumed to be independent of both X and T. Now let $\delta \in \{0, 1\}$ denote a binary random variable indicating wether the subject is withdrawn from the study before its actual failure (right censoring). If $\delta = 1$, let $Y \geq 0$ denote the time of failure. Formally

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

Let the set $\mathcal{D} = \{(X_i, Y_i, \delta_i)\}_{i=1}^n$ denote the observed i.i.d. samples from a distribution $F_{XY\delta}$. Let the random variable Z denote the random triple (X, Y, δ) with i.i.d. samples $\{Z_i\}_{i=1}^n$.

The paper uses the following notations for vectors and matrices: a lower case b will denote a vector or a scalar (clear from context), and $\mathbf{b}$ will denote a matrix. In particular, the data samples $\{X_i\}_{i=1}^n \subset \mathbb{R}^d$ can be organised

in a matrix $\mathbf{x} \in \mathbb{R}^{n\times d}$ as $\mathbf{x}_i = X_i^T$ for all $i = 1, \ldots, n$, whereas the variables $\{y_i\}_{i=1}^n \subset \mathbb{R}$ can be organised in a vector y. Furthermore, let $1_n \in \mathbb{R}_n$ and $0_n \in \mathbb{R}_n$ be a constant vector with constants respectively 1 and 0, and let $\mathbf{I}_n \in \mathbb{R}_{n\times n}$ be the identity matrix of size n such that $\mathbf{I}_{n,ij} = I(i = j)$ for all $i, j = 1, \ldots, n$, where $I(z > 0) = 1$ if $z > 0$, and zero otherwise.

Case Without Censoring

In the case no censoring occurs $Z = (X, T)$. For notational convenience, define the kernel $q_u : \mathcal{Z} \times \mathcal{Z} \to \{0, 1\}$ for any $Z, Z' \in \mathcal{Z}$ as follows

$$q_u(Z, Z') = I((u(X) - u(X'))(T - T') > 0)\,. \quad (2)$$

The key objects of study of this paper are the health index $u : \mathbb{R}^d \to \mathbb{R}^+$, the c-index between two covariates, its empirical version and the indexed random variable $M_u \in [0, 1]$.

Let $u : \mathbb{R}^d \to \mathbb{R}^+$ be a health function mapping a covariate $x \in \mathbb{R}^d$ to $u(x) \in \mathbb{R}$. Let (X, T) be a random variable with joint distribution F_{XT}. The health index u is said to be concordant with respect to F_{XT} to a level

$$\begin{aligned} CI(u; F_{XT}) &= P(u(X) > u(X') \big| T > T') \\ &= P(I((u(X) - u(X'))(T - T') > 0)) \\ &= P(q_u(Z, Z'))\,. \end{aligned} \quad (3)$$

Informally, a health index is concordant with failure times encoded in T if the ordering of two randomly drawn subjects is with high probability similar in both failure time and health. This quantity can be estimated using an i.i.d. sample $\{(X_i, T_i)\}_{i=1}^n$ which is ordered as $T_i \leq T_j$ for all $i < j$, as follows

$$\begin{aligned} CI_n(u) &= \frac{2}{n(n-1)} \sum_{T_i<T_j} I(u(X_i) < u(X_j)) \\ &= \frac{1}{n(n-1)} \sum_{i\neq j} I((u(X_i) - u(X_j))(T_i - T_j) > 0) \quad (4) \\ &= \frac{1}{n(n-1)} \sum_{i\neq j} q_u(Z_i, Z_j)\,. \end{aligned}$$

The second equality holds by symmetry. The degree of concordance of a new sample Z with n i.i.d. copies $\{Z_i\}_{i=1}^n$ is denoted by

$$M_u(Z; Z_1, \ldots, Z_n) = \frac{1}{n} \sum_{i=1}^n I((u(X) - u(X_i))(T - T_i) > 0)\,. \quad (5)$$

Right-Censored Observations

In the context of right-censored observations Z denotes the triple (X, T, δ) with i.i.d. samples $\{(X_i, T_i, \delta_i)\}_{i=1}^n$. Let the quantity of interest be defined as

$$q_u(Z, Z') = I((T - T')(u(X) - u(X')) > 0)W(Z, Z')\,, \quad (6)$$

where the symmetric term $W(Z, Z') = W(Z', Z) \in \{0, 1\}$ denotes whether the couple (Z, Z') is comparable. In the case of right-censoring, a formal definition of this term becomes

$$\begin{aligned} W(Z, Z') &= I(Y < Y' \wedge \delta = 1) + \\ &\quad I(Y > Y' \wedge \delta' = 1) \end{aligned} \quad (7)$$

for any $Z, Z' \in \mathcal{Z}$. The c-index $CI_n(u)$ then becomes

$$\frac{2}{|W|} \sum_{T_i<T_j} I((T_j - T_i)(u(X_j) - u(X_i)) > 0)W(Z_j, Z_i) \quad (8)$$

where $|W|$ is defined as $|W| = \sum_{T_i<T_j} W(Z_j, Z_i)$. See e.g. [8, 16, 17].

LEARNING MACHINE FOR PREDICTIVE MODELING

We consider the following mechanism for learning a predictive model of failure time data. The following definition defines the objective of learning as well as a proper way to make predictions regarding failure time corresponding to a new subject. Let $\mathcal{H} \in \{u : \mathbb{R}^d \to \mathbb{R}\}$ be a class of plausible health indices. An empirical risk minimizer takes the following form

$$\hat{u} = \arg\min_{u\in\mathcal{H}} \mathcal{R}_n(u; \mathcal{D}) = \arg\max_{u\in\mathcal{H}} CI_n(u)\,. \quad (9)$$

This paper studies the class of health functions defined as

$$\mathcal{H}_u = \left\{ u_w : \mathbb{R}^d \to \mathbb{R}, w \in \mathbb{R}^d \middle| u_w(x) = w^T x, x \in \mathbb{R}^d \right\}, \quad (10)$$

An intercept term is not relevant in this context since it does not affect the ordering relation.

Pairwise Maximal Margin Machine

As minimizing $1 - CI_n(u)$ directly amounts to a hard combinatorial problem, a convex upper bound in terms of the hinge loss is generally used. The hinge loss of a term $e \in \mathbb{R}$ is defined as $\ell(e) = (1 - e)_+$ where $(\cdot)_+$ returns the positive part of the argument. Minimizing this upperbound will yield the optimal estimator $\arg\min_w 1 - CI_n(u_w(X), T)$ when $n \to \infty$ (see e.g. [13], section 4). This motivates the use of the linear estimator $\min_w \sum_{i<j} \ell(w^T X_j - w^T X_i)$. Adding a regularization term $w^T w$, and rewriting the estimator as a standard quadratic programming problem results in the following estimator. The empirical optimal health index $u_w : \mathbb{R}^d \to \mathbb{R}$ is found as follows

$$\begin{aligned} (\hat{w}, \hat{\xi}) &= \arg\min_{w,\xi} \frac{1}{2} w^T w + C \sum_{i<j,\delta_i=1} W_{ij}\xi_{ij} \\ &\text{s.t.} \quad w^T X_j - w^T X_i \geq 1 - \xi_{ij}, \\ &\quad \xi_{ij} \geq 0\,, \end{aligned} \quad (11)$$

where C is a positive real constant and $\{W_{ij}\}$ is a set with appropriate weighting terms $W_{ij} = W(Z_i, Z_j)$, dealing with the specific censoring pattern.

As for the case of support vector machines, this optimization problem can be motivated alternatively using the concept of maximal margin. The margin of a point $x \in \mathbb{R}^d$ to a hyperplane $x \in \mathbb{R}^d : w^T x = 1$ is defined as $\frac{|w^T x|}{||w||_2}$. The *signed* margin of a sample $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$ with this hyperplane is defined as $\frac{y(w^T x)}{||w||_2}$. Now, for any couple (X_i, X_j) with corresponding times $T_i < T_j (\text{or } i < j)$, one can define a hyperplane $\{w^T x + b_{ij} = 0 | (X_i, X_j), i < j\}$. Maximizing the margin between the two points can be done as $\max_{w,b_{ij}} \frac{-(w^T X_i + b_{ij})}{||w||_2} + \frac{(w^T X_j + b_{ij})}{||w||_2} = \frac{w^T (X_j - X_i)}{||w||_2}$. Or equivalently as $\min_w ||w|| \text{ s.t. } w^T (X_j - X_i) \geq 1$. Maximizing all margins between any couples (X_i, X_j) for all $i < j$ is done by the following optimization problem

$$\min_w \frac{1}{2} w^T w \text{ s.t. } w^T (X_j - X_i) \geq 1, \forall\, i < j\,. \quad (12)$$

It is clear that if such a health index u_w can be found, the corresponding concordance index equals one exactly. In general, it is advantageous to allow for violations in the constraints using slack variables. Doing so, and minimizing the one-norm of the slack variables $\{\xi_{ij}\}_{i<j}$ results again in the estimator (11).

A Primal-Dual Maximal Margin Kernel Machine

It is shown how this formulation can be rewritten as a nonlinear kernel-based version based on an argument from convex optimization theory. Let $N \in \mathbb{N}_0$ be defined as the number of elements is the set $\{(i, j) : i < j = 1, \ldots, n\}$. The Lagrangian to (11) becomes

$$\mathcal{L}(w, \xi; \alpha, \beta) = \frac{1}{2} w^T w + C \sum_{i<j} W_{ij} \xi_{ij} - \sum_{i<j} \alpha_{ij} (w^T X_j - w^T X_i - 1 + \xi_{ij}) - \sum_{i<j} \beta_{ij} \xi_{ij}\,, \quad (13)$$

with multipliers $\alpha, \beta \in \mathbb{R}^N_+$. Switching the order $\max_{\alpha,\beta} \min_{w,\xi} \mathcal{L}(w, \xi; \alpha, \beta)$ to $\min_{w,\xi} \max_{\alpha,\beta} \mathcal{L}(w, \xi; \alpha, \beta)$ (Slater's condition is trivially satisfied) and taking the first order conditions for optimality yields:

$$\begin{cases} \frac{\partial \mathcal{L}(w,\xi;\alpha,\beta)}{\partial w} = 0 \;\rightarrow\; w = \sum_{i<j} \alpha_{ij} (X_j - X_i) \\ \frac{\partial \mathcal{L}(w,\xi;\alpha,\beta)}{\partial \xi_{ij} b} = 0 \;\rightarrow\; C W_{ij} = \alpha_{ij} + \beta_{ij}, \quad \forall\, i < j\,. \end{cases} \quad (14)$$

Hence the dual problem becomes:

$$\min_\alpha \frac{1}{2} \sum_{i<j} \sum_{k<l} \alpha_{ij} \alpha_{kl} (X_j - X_i)^T (X_l - X_k) - \sum_{i<j} \alpha_{ij} \quad \text{s.t. } 0 \leq \alpha_{ij} \leq C W_{ij}, \forall\, i < j\,. \quad (15)$$

Note the relationship with the dual expression corresponding with a maximal margin classifier (e.g. the SVM). This can be rewritten in matrix-vector notation by introducing the matrix $D \in \{-1, 0, 1\}^{N \times n}$, such that

$$D\mathbf{x} = \begin{bmatrix} X_1 - X_2 \\ \vdots \\ X_1 - X_n \\ \vdots \\ X_{n-1} - X_n \end{bmatrix}. \quad (16)$$

The dual problem of (11) is given as

$$\hat{\alpha} = \arg\min_\alpha \frac{1}{2} \alpha^T D \mathbf{x}\mathbf{x}^T D^T \alpha - 1_N^T \alpha \quad \text{s.t. } 0_N \leq \alpha \leq C\mathbf{w}\,, \quad (17)$$

where $\mathbf{w} \in \mathbb{R}^N$ is defined as $\mathbf{w} = (W_{12}, \ldots, W_{(n-1)n})^T$. The corresponding optimal solution $\hat{w}$ can be evaluated in a new point $X_* \in \mathbb{R}^d$ as

$$\begin{aligned} u_w(X_*) &= \hat{w}^T X_* \\ &= \sum_{i<j} \hat{\alpha}_{ij} (X_i - X_j)^T X_* \\ &= (\mathbf{x} X_*)^T D^T \hat{\alpha}\,. \end{aligned} \quad (18)$$

Application of the kernel trick gives the following result. Given a positive definite kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ inducing a feature map $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}^{d_\varphi}$ such that $K(X, X') = \varphi(X)^T \varphi(X')$ for any $X, X' \in \mathbb{R}^d$. Let the kernel matrix $\Omega \in \mathbb{R}^{n \times n}$ be defined as $\Omega_{ij} = K(X_i, X_j)$ for all $i, j = 1, \ldots, n$, and let $\Omega(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^n$ be defined as $\Omega(X) = (K(X_1, X), \ldots, K(X_n, X))^T \in \mathbb{R}^N$.

Given an appropriate kernel function and its corresponding kernel matrix, the optimal empirical failure time predictor is found by

$$\hat{\alpha} = \arg\min_\alpha \frac{1}{2} \alpha^T D \Omega D^T \alpha - 1_N^T \alpha \quad \text{s.t. } 0_N \leq \alpha \leq C\mathbf{w}\,, \quad (19)$$

and the corresponding optimal solution can be evaluated in a new point $X_* \in \mathbb{R}^d$ as

$$u_w(X_*) = \Omega(X_*)^T D^T \hat{\alpha}\,. \quad (20)$$

Remark that the formulation corresponds exactly with the so-called rank SVMs as introduced in [11], except for the terms $\{W_{ij}\}$ coping properly with right-censoring.

Visualization of results

For the representation of the results, two figures are introduced. Since the goal of this paper is to use the c-index in order to predict the correct order of events, the ranks of the observed events are plotted against the ranks of the predicted events. A model perfectly predicting the event order results in a straight line beginning at the bottom left and ending at the top right. This corresponds with a c-index equal to 1. The prediction of the reverse order leads to a straight line perpendicular to the one mentioned before. A random order is represented by a c-index of 0.5 and a horizontal line as is showed in Figure 1. To obtain such a figure a spline smoother is used.

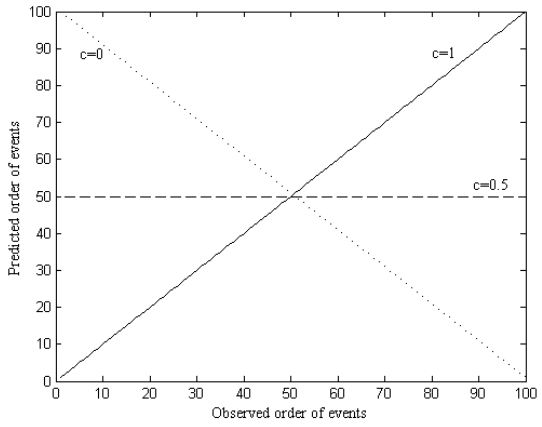

Fig. 1: Illustration of the link between the c-index and the relation between ranks of observed and predicted events. If the predicted order equals the observed order a line at 45 degrees is obtained, corresponding to a c-index of 1.

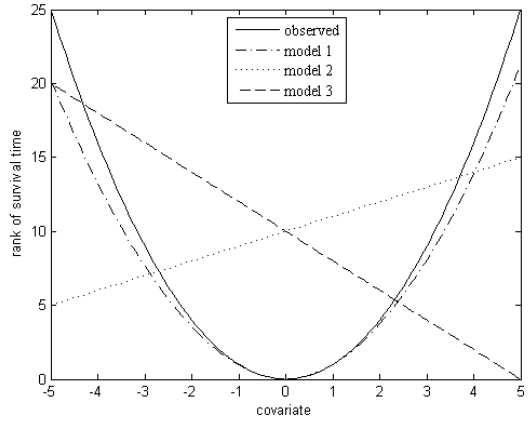

Fig. 2: Illustration of the predicted relation between a covariate and ranks of the survival time. Models predicting the observed relationship are preferred. Model 1 is selected above the other models.

In practical applications it is important to estimate the correct relation between variables and survival time. This property is checked by plotting the predicted and observed survival ranks against the covariate values after smoothing with a spline. An example is given in Figure 2.

RESULTS

To compare the proposed method with standard models for survival analysis, artificial data were created. The performance of the model based on the c-index is compared with the performance of the Cox proportional hazard model [18] and the accelerated failure time model [1]. Finally a heart failure dataset is used for comparison using real data.

Artificial Data

Artificial data is created to compare prediction of survival time for the described method with more common methods as there are the proportional hazard model and the accelerated failure time model [1, 18]. The first artificial dataset (data 1) contains 500 samples and six covariates sampled from a random distribution with zero mean and a unit variance. Two survival times are generated. The first survival time T_1 is sampled from a Weibull distribution

$$F(t|X) = 1 - \exp(-[\lambda(X)t]^{\alpha}), \qquad (21)$$

with $\lambda(X_i) = \exp(\sum_{k=1}^{d} X_{i,k})$, α was chosen to be 1.5 and d represents the number of covariates. $X_{i,k}$ represents the kth covariate of patient i. The second survival time is calculated as

$$T_{i,2} = \exp(\sum_{k=1}^{d} X_{i,k} + \varepsilon), \qquad (22)$$

where ε represents gaussian noise with unit variance. Ten percent of all datapoints were censored . A second artificial dataset (data 2) contains fourteen covariates. Survival times are calculated as before. All mentioned performances are calculated on an independent test set.

Each model uses all variables. Training was done on a training set of 300 samples, an equally sized test set was used to test performance. The models based on the c-index used 200 samples out of the training as a validation set. Hyperparameters giving the best c-index on the validation set were selected. Calculation of the c-index on test and validation set was performed as

$$CI_n^t = \frac{1}{n_t(n_t-1)} \sum_{i \neq j} q_u(Z_i, Z_j) W(Z_i, Z_j), \qquad (23)$$

for samples $i,\ j$ of the test set and

$$CI_n^v = \frac{1}{n_v(n_v-1)} \sum_{i \neq j} q_u(Z_i, Z_j) W(Z_i, Z_j), \qquad (24)$$

for samples $i,\ j$ of the validation set respectively. n_t denotes the number of samples in the test set and n_v the number of samples in the validation set.

The variables and survival time were first normalized to zero mean and unit standard deviation. This normalization didn't seem to influence the results.

Weibull distributed survival time

As can be expected the proportional hazard model and the accelerated failure time model perform well in predicting a Weibull distributed survival time. Since the accelerated failure time model is a proportional hazard model in case of a Weibull distribution it is not surprising that both methods are performing equally well. On a test set of data 1 the proportional hazard model and the accelerated failure time model have a c-index of 0.9201. The presented model has a c-index of 0.9300 when using a linear kernel and one of 0.9350 when using a radial basis

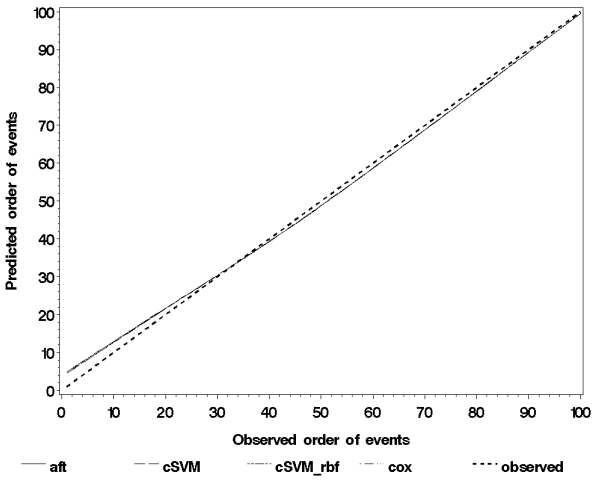

Fig. 3: Ordering of events for a Weibull distributed survival time. A model predicting the order of the events correctly results in a straight line beginning at the bottom left and ending at the upper right. All models preserve the observed order relatively well.

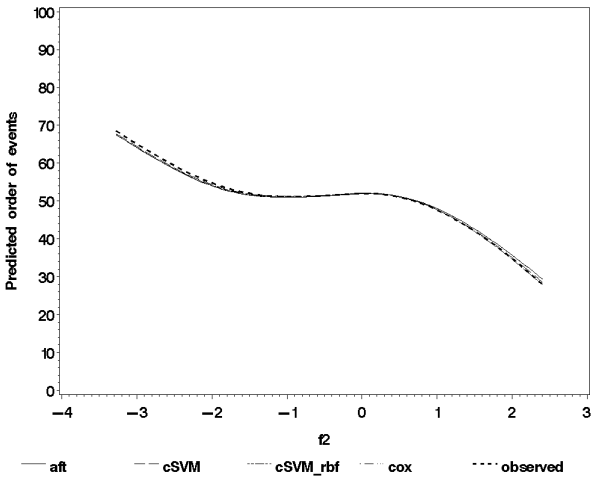

Fig. 4: Effect of the second covariate (f2) on the prediction of the event time. The observed relationship between f2 and the Weibull distributed event time is indicated as observed. All methods are able to predict the correct relation well.

function (RBF) as a kernel (Table 1). No statistically significant difference is observed between the predicted survival curves using the log rank test. Figure 3 shows that all the models are able to predict the ordering of events relatively well.

Identical analysis is performed on dataset 2. For a survival time with weibull distribution, the proportional hazard and accelerated failure time model obtained a c-index of 0.9501. The cSVM model with a linear kernel had a c-index of 0.9543. The model with an rbf kernel had a c-index equal to 0.9539 (Table 1). The differences in c-index between the models are not significant.

A second important issue in modeling survival time concerns the estimation of the correct relation between different covariates and the predicted survival time. To investigate this property, the second variable (f2) is plotted against the predicted ordering of events as well as the observed ordering of events. Figure 4 represents the observed and predicted relationships on the first dataset when predicting the Weibull distributed survival time. All models obtain the correct relation between covariate and survival time.

dataset		T_1 (eq. (21))	T_2 (eq. (22))
data1	Cox	0.9201	0.8870
	aft	0.9201	0.8870
	cSVM	0.9300	0.8885
	cSVM_rbf	0.9350	0.8874
data2	Cox	0.9501	0.8907
	aft	0.9501	0.8907
	cSVM	0.9543	0.8877
	cSVM_rbf	0.9539	0.8838

Table 1: c-index on artificial data for the investigated models on an independent test set for two different survival times T_1 and T_2. The proportional hazard model is referred to as the Cox-model, the accelerated failure time model as aft, the method using support vector machines with a linear kernel as cSVM and the one with an rbf kernel as cSVM_rbf. The upper part of the table contains the results for a dataset with only six variables. The lower part resumes the results for a dataset with fourteen variables.

Survival time sampled from an accelerated failure time distribution

The log rank test for testing differences between survival curves does not show evidence for a significant difference in survival curves when predicting the second survival time. All models show a lower c-index than in the previous example. The obtained results are analogous to the results of the first example.

Heart Failure Data

The Worcester Heart Attack Study WHAS500 Data was used to compare the different models on a real life dataset [20]. These data are from Dr. Robert J. Goldberg of the Department of Cardiology at the University of Massachusetts Medical School and contains data of 500 patients with 14 relevant covariates. All patients were hospitalized after an acute myocardial infarction between 1975 and 2001. Patients were admitted to hospitals in the Worcester, Massachusetts Standard Metropolitan Statistical Area. The dataset was randomly divided into a training and test set, each containing 100 patients. Remaining datapoints were used as a validation set to tune hyperparameters for the cSVM models.

To select variables for the proportional hazard model

and the accelerated failure time model a univariate analysis was performed. Only variables with a p-value less than 0.1 were selected for multivariate analysis. Backward selection was used to obtain the final model. Both the proportional hazard and accelerated failure time model retained patient's age and the factor indicating wether he or she had congestive heart complications (chf). In a second experiment all variables were used.

Table 2 summarizes the results on the heart failure data. When using all variables the proportional hazard and accelerated failure time model perform better than the models based on the c-index, as illustrated in Figure 5. The lower c-index of the cSVM model with an rbf kernel is confirmed by the curve lying further away from the ideal curve than for all other models. According to the log rank test there is no significant difference between the obtained survival curves. When only age and history of cardiovascular disease are used the c-index is highest for the cSVM model with rbf kernel, but the differences are not significant.

Comparison of the proportional hazard, accelerated failure time and cSVM model can be done by looking at the weights these models obtain for each of the covariates (Table 3). An indication for the relationship between the variable and the survival time is the sign of the weight. A positive weight in the accelerated failure time and cSVM model or a negative weight in the proportional hazard model indicate an increasing survival time for an increasing covariate. The signs of the weights differ for atrial fibrillation and history of cardiovascular disease. The proportional hazard and accelerated failure time point out a lower survival time for patients with atrial fibrillation. The cSVM model points out an increasing survival time for patients with atrial fibrillation. Figure 6 shows the observed and predicted survival curves stratified for atrial fibrillation. The proportional hazard and accelerated failure time model underestimate survival of patients with atrial fibrillation leading to a significant difference in survival curves stratified for atrial fibrillation. The predicted survival curves of the cSVM model lie closer to the observed survival. As is the case for the observed survival, no significant difference in survival curves stratified for atrial fibrillation is detected.

	variable selection	all variables
Cox	0.7176	0.7512
aft	0.7160	0.7502
cSVM	0.7125	0.7251
cSVM_rbf	0.7199	0.6796

Table 2: c-index for the investigated models on time to failure of the heart tested on an independent test set of 100 datapoints. After variable selection only age and history of cardiovascular disease were kept. cSVM with an rbf kernel has the largest c-index after variable selection. The differences in c-index between the models is higher when all variables are used; the proportional hazard model performs best.

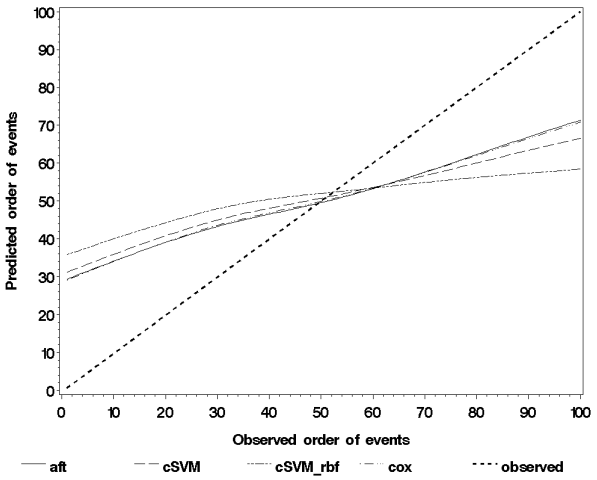

Fig. 5: Predicted and observed ranks of survival on an independent test set for heart failure data. The curve for the cSVM model with an rbf kernel lies further away from the ideal curve, due to a lower c-index compared to the other models.

The relationship between the ranks of the survival times and the initial heart rate of the patient is investigated in Figure 7. The cSVM model best estimates the relationship the best for heart rates below 90 beats per minute. For higher heart rates the proportional hazard and accelerated failure time are preferred.

CONCLUSIONS

This paper showed how modified support vector machines can be used for survival analysis. The proposed method is able to predict the ordering of complex survival times, which is made possible by the ordering encoded in the c-index. From the experiments it turned out that even when the data were generated from a classical parametric model the results obtained with cSVM models were comparable to results of the proportional hazard and

	Cox	aft	cSVM
age	0.05054	-0.0873	-0.0637
gender	-0.35556	0.6508	0.3428
heart rate	0.01328	-0.0241	-0.0006
systolic blood pressure	0.01103	-0.0194	-0.0069
diastolic blood pressure	-0.02689	0.0478	0.0129
body mass index	-0.00037	0.0003	0.0542
history of cardiovascular disease	-0.24532	0.4028	-0.2211
atrial fibrillation	0.39205	-0.7032	0.1538
cardiogenic shock	1.14483	-2.5863	-0.9196
congestive heart complications	0.70369	-1.2799	-0.9248
complete heart block	1.41930	-2.6097	-0.4868
myocardial infarct order	0.15357	-0.1989	-0.0856
myocardial infarct type	-0.55150	1.0194	0.4489
length of hospital stay	-0.00518	0.0124	0.0282

Table 3: Weight given for each variable of the heart failure dataset by the different models.

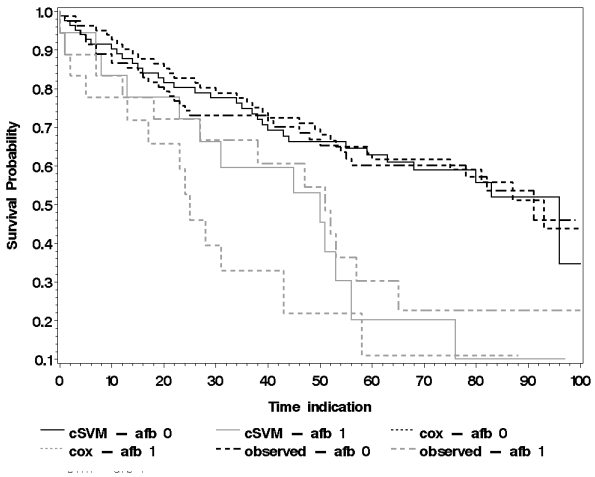

Fig. 6: Predicted survival curves on an independent test set, stratified for atrial fibrillation (afb). The proportional hazard model and the accelerated failure time model underestimate survival for patients with atrial fibrillation. The cSVM and rbfcSVM models give a better indication of survival. The difference is not significant according to the log rank test. All models perform well for patients without atrial fibrillation.

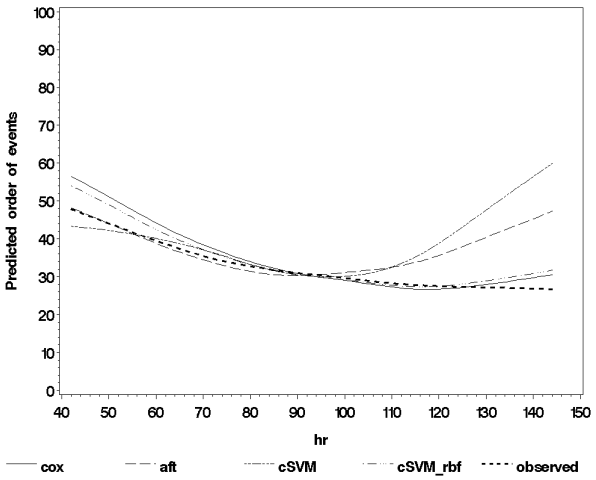

Fig. 7: Investigation of the relation between heart rate and predicted and observed rank of survival time, on an independent test set. The cSVM model predicts the relationship the best for heart rates below 90 beats per minute. For higher heart rates the proportional hazard and accelerated failure time are preferred.

accelerated failure time model. The performance of the current formulation can be improved by further optimization of the hyperparameters. In order to generate more interpretable models different variable selection methods will be explored.

Currently work is done to decrease calculation complexity in order to make calculations on larger datasets possible. We started looking for similar formulations in the context of least squares support vector machine, reducing the complexity of the algorithm from $O(n^2)$ to $O(n)$.

Once larger datasets can be handled experiments will be set up in cooperation with the Multidisciplinary Breast Centre (MBC) U.Z. Leuven for testing these models on a large breast cancer dataset.

ACKNOWLEDGMENTS

This research is funded by a PhD grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). This research work was carried out at the ESAT laboratory of the Katholieke Universiteit Leuven. It is supported by grants from several funding agencies and sources: EU: BIOPATTERN (FP6-2002-IST 508803); Research Council KUL: GOA AMBioRICS, CoE EF/05/006 Optimization in Engineering, several PhD/postdoc & fellow grants; Flemish Government: FWO: PhD/postdoc grants, projects, G.0407.02 (support vector machines), G.0197.02 (power islands), G.0141.03 (Identification and cryptography), G.0491.03 (control for intensive care glycemia), G.0120.03 (QIT), G.0452.04 (new quantum algorithms), G.0499.04 (Statistics), G.0211.05 (Nonlinear), G.0226.06 (cooperative systems and optimization), G.0321.06 (Tensors), G.0302.07 (SVM/Kernel, research communities (ICCoS, ANMMM, MLDM); IWT: PhD Grants, McKnow-E, Eureka-Flite2 Belgian Federal Science Policy Office: IUAP P6/04 (Dynamical systems, control and optimization, 2007-2011) ; EU: ERNSI. JS is a professor and SV is a full professor at KU Leuven, Belgium.

REFERENCES

[1] J.D. Kalbfleisch, R.L. Prentice. *The Statistical Analysis of Failure Time Data.* Wiley series in probability and statistics. Wiley (2002).

[2] T. Hastie, R. Tibshirani, J. Friedman. *The Elements of Statistical Learning.* Springer-Verlag (2001).

[3] L. Devroye, L. Györfi, G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag (1996).

[4] V. Vapnik. *Statistical Learning Theory.* Wiley and sons (1998).

[5] J. Shawe-Taylor, N. Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press (2004).

[6] J.A.K. Suykens, T. Van Gestel, J. De Brabanter and B. De Moor, J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, Singapore (2002).

[7] S. Boyd and L. Vandenberghe. *Convex Optimization*, Cambridge University Press (2004).

[8] Jr. Harrell, K. Klee, R. Califf, D. Pryor, R. Rosati. ”Regression modeling strategies for improved prognostic prediction.” *Statistics in Medicine*, **95**, pp.634-635 (1984).

[9] W.H. Kruskal. ”Ordinal measures of association.” *Journal of the American Statistical Association*, **53**, pp. 814-861 (1958).

[10] M.G. Kendall, J.D. Gobbons *Rank Correlation Methods (5th edition)*. London: Arnold 1990.

[11] R. Herbrich, T. Graepel, K. Obermayer. ”Large margin rank boundaries for ordinal regression.” *Advances in Large Margin Classifiers*, pp. 115-132 (2000).

[12] K. Crammer and Y. Singer. ”Pranking with ranking.” *In Proceedings of the conference on Neural Information Processing Systems (NIPS)* (2001).

[13] S. Clemencon, G. Lugosi, N. Vayatis, P. Aurer, R. Meir. ”Ranking and scoring using empirical risk minimisation.” *18th Annual Conference on Learning Theory (COLT)* (2005).

[14] A. Rakotomamonji. ”SVMs and area under the ROC curves.” *Technical Report PSI-INSA Rouen 2004* (2004).

[15] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, D. Roth. ”Generalizing bounds for the area under the ROC curve.” *Journal of Machine Learning Research*, **6** pp. 393-425 (2005).

[16] L. Antolini , P. Boracchi , E. Biganzoli. ”A time-dependent discrimination index for survival data.” *Statistics in Medicine*, **24** (24), pp. 3927-3944 (2005).

[17] L. Yan, D. Verbel and O. Saidi. ”Predicting prostate cancer recurrence via maximizing the concordance index.” *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 479-485 (2004).

[18] D.R. Cox. ”Regression Models and Life-Tables (with Discussion).” *Journal of the Royal Statistical Society*, Series B, **34**, pp. 187-220 (1972).

[19] K. Pelckmans, J.A.K. Suykens and B. De Moor. ”Predictive Modeling of Failure Time Data.” *Internal Report 07-35, ESAT-SISTA, K.U.Leuven (Leuven, Belgium)* (2007).

[20] http://www-unix.oit.umass.edu/∼statdata/statdata/stat-survival.html