On the use of a clinical kernel in survival analysis

V. Van Belle[1], K. Pelckmans[2], J.A.K. Suykens[1] and S. Van Huffel[1]

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

2- Department of Information Technology,
University of Uppsala, SE-751 05 Uppsala, Sweden
kp@it.uu.se

Abstract. Clinical datasets typically contain continuous, ordinal, categorical and binary variables. To model this type of datasets, linear kernel methods are generally used. However, the linear kernel has some disadvantages, which were tackled by the introduction of a clinical one. This work shows that the use of a clinical kernel can improve the performance of support vector machine survival models. In addition, the polynomial kernel is adapted in the same way to obtain a clinical polynomial kernel. A comparison is made with other non-linear additive kernels on six different survival data. Our results indicate that the use of a clinical kernel is a simple way to obtain non-linear models for survival analysis, without the need to tune an extra parameter.

1 Introduction

Kernel based methods find more and more applications within medical decision making, e.g. prediction of malignancy of tumors, classification of tumors, prediction of viability of pregnancies, etc. In such medical problems, different types of information are provided. Some variables will be continuous, e.g. patient's age, binary, e.g. smoking, ordinal, e.g. a performance score (low, moderate, high, excellent) and others will be nominal variables, e.g. cell type. However, the linear kernel does not take the type of variable into account. The linear kernel is calculated as the inner product of the normalized variable values. Although this is an easy way to calculate similarity and the interpretation of results is straightforward, some disadvantages remain. First, the similarity between whatever value of a variable and a value of zero will always be zero, whether both values are close or not. Second, for ordinal data, the similarity between values of two adjacent classes depends on the total number of levels of the variable. Third, for nominal variables, dummy variables need to be calculated, in order to be able to treat them as non-related.

To tackle the problems described above, a clinical kernel was proposed in [1] for classification problems. The clinical kernel is an additive kernel, as the linear one, but instead of calculating cross-products, it calculates the relative difference of differences in function values for continuous and ordinal data. For nominal data, the value of the kernel is set to 1 for values that are exactly the same, and

0 otherwise. In this way it is no longer necessary to make $k-1$ dummy variables for $k-$level nominal variables.

In this paper, the clinical kernel is applied to support vector machine survival models [2, 3, 4]. In addition, the proposed adaption of the linear kernel to obtain the clinical one, is adapted to the polynomial kernel. We investigate whether both kernels can improve the performance of kernel based survival models. The rest of the paper is organized as follows. Section 2 starts by describing the kernel based model for survival analysis. In Section 3 the clinical kernel is compared with the linear kernel and the polynomial kernel is adapted towards a clinical polynomial kernel. Section 4 illustrates the use of different kernels on 6 clinical survival datasets. Finally, Section 5 gives some conclusions.

2 Support vector machines in survival analysis

Building survival models with kernel based methods is based on the empirical maximization of the concordance index (c-index) [5]. The c-index measures the percentage of comparable pairs which is concordant. A pair of observations is considered to be concordant whenever (i) the pair is comparable and (ii) the difference in observed failure time and the difference in model outcome have the same sign. A comparable pair is a pair for which the time order is known. Non comparable pairs are: (i) one observation with an event after x years and the other observation with a right censored event time at y years with $y < x$; (ii) two right censored observations.

In addition to the empirical optimization of the c-index, the model outcome is targeted at the true event time for events, and at a value larger than the censoring time for right censored observations. Let $u(x) = w^T\varphi(x)$. Let Y be the vector containing the sorted failure times and $\Phi = [\varphi(x_1)\dots\varphi(x_n)]^T \in \mathbb{R}^{n\times n_\varphi}$, with feature map φ, the matrix containing the corresponding features. The model formulation then becomes (see [4] for more details)

$$\min_{w,\epsilon,\xi,\xi^*} \frac{1}{2}w^Tw + \gamma 1^T\epsilon + \mu 1^T(\xi+\xi^*) \quad \text{s.t.} \quad \begin{cases} D\Phi w + \epsilon \geq DY \\ \Phi w \geq Y - \xi \\ -R\Phi w \geq -RY - \xi^* \\ \epsilon \geq 0 \\ \xi \geq 0 \\ \xi^* \geq 0\,, \end{cases} \tag{1}$$

where $R = \mathrm{diag}(\delta)$, with δ a censoring indicator vector ($\delta = 1$ for events, 0 otherwise). The matrix D ensures the comparison between pairs:

$$D = \begin{bmatrix} -1 & 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 & 0 \\ \vdots & & & & & & \vdots \\ 0 & \dots & 0 & 0 & 0 & -1 & 1 \end{bmatrix}. \tag{2}$$

The data are sorted from the beginning such that $DY \geq 0$. Therefore the c-index would be optimized by $D\Phi w \geq 0$. Targeting $D\Phi w$ at DY instead of at a positive value, indicates that the difference in failure should be considered [2]. After formulating the Lagrangian of (1) and taking the KKT conditions, the solution if obtained from

$$\begin{aligned}\min_{\alpha,\beta,\beta^*} \quad & \tfrac{1}{2} \begin{bmatrix} \alpha^T & \beta^T & \beta^{*T} \end{bmatrix} \begin{bmatrix} DKD^T & DK & -DKR \\ KD^T & K & -KR \\ -RKD^T & -RK & RKR \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \beta^* \end{bmatrix} \\ & + \begin{bmatrix} -Y^TD^T & -Y^T & Y^TR \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \beta^* \end{bmatrix} \quad \text{s.t.} \begin{cases} 0 \leq \alpha \leq \gamma \\ 0 \leq \beta \leq \mu \\ 0 \leq \beta^* \leq \mu \,. \end{cases}\end{aligned} \tag{3}$$

The outcome $u(x^*)$ for a new observation x^* can then be found as $u(x^*) = (D^T\alpha + \beta - R\beta^*)^T K_n$, with $K_n = [\varphi(x_1)^T\varphi(x^*) \dots \varphi(x_n)^T\varphi(x^*)]^T$.

3 Kernel functions

In kernel based methods one does not have to specify $\varphi(\cdot)$ explicitly, but the mapping $\phi(\cdot)$ of a covariate is induced implicitly by defining the innerproduct $K(x_i, x_j) = \varphi(x_i)^T\varphi(x_j)$. This in turn requires that the user specifies a suitable kernel function, rather than a (high-dimensional) representation $\varphi(x)$ of x. Generally one chooses one of the following kernels: (i) a linear kernel $K(x, x^*) = x^Tx^*$; (ii) a polynomial kernel of degree a $K(x, x^*) = (\tau + x^Tx^*)^a$, where $\tau \geq 0$ or (iii) an RBF kernel $K(x, x^*) = \exp(-||x - x^*||_2^2/\sigma^2)$, with τ and σ tuning parameters.

An alternative to the linear kernel was proposed in [1]. This clinical kernel is an additive kernel $K_{\text{clin}}(x_i, x_j) = \sum_{p=1}^{d} K_{\text{clin}}^{(p)}(x_i^{(p)}, x_j^{(p)})$, where the componentwise kernel $K^{(p)}$ is calculated differently for different types of covariates. For continuous and ordinal variables, the kernel is defined as [1]

$$K_{\text{clin},1}^{(p)}(x_i^{(p)}, x_j^{(p)}) = \frac{(\max^{(p)} - \min^{(p)}) - |x_i^{(p)} - x_j^{(p)}|}{\max^{(p)} - \min^{(p)}}\,, \tag{4}$$

where $\min^{(p)}$ and $\max^{(p)}$ are the minimal and maximal value of the covariate p, evaluated on training data. For nominal variables, the kernel is defined as

$$K_{\text{clin},2}^{(p)}(x_i^{(p)}, x_j^{(p)}) = \begin{cases} 1 \text{ if } x_i^{(p)} = x_j^{(p)} \\ 0 \text{ if } x_i^{(p)} \neq x_j^{(p)}\,. \end{cases} \tag{5}$$

Since the polynomial kernel has the same disadvantages as the linear one, we adapted the polynomial kernel in the same way as the linear kernel: $K_{\text{poly}-\text{clin}} = (\tau + K_{\text{clin}})^d$ and call it the clinical polynomial kernel. Due to Mercer's condition, the kernel needs to be positive definite for the kernel trick to be applicable in

(1) and (3). One can prove that the clinical and polynomial clinical kernel are both positive definite kernel using kernel properties (see [6]).

4 Results

This Section describes the comparison of six clinical survival datasets: one dataset concerning leukemia [7], two lung cancer datasets [8, 9], one breast cancer dataset [10], one about prostatic cancer [11] and a sixth dataset on kidney transplants [12]. More information on the datasets can be found in Table 1 and in the references. All datasets were 100 times randomly divided into training and test sets. Half of the training data were used as a validation set in the tuning phase. Coupled simulated annealing [13] was used to tune the regularization and/or kernel parameters.

Table 1: Description of the six clinical survival datasets.

dataset	# test	# training	# nominal	# cont/ordinal
leukemia (LE)	43	86	3	5
lung cancer (1) (LC1)	46	91	4	3
lung cancer (2) (LC2)	56	111	1	6
prostatic cancer (PC)	161	322	3	5
breast cancer (BC)	229	457	2	6
kidney transplant (KT)	288	575	2	1

In Figure 1 the differences in estimated functional forms for the variables age and Karnofsky score (indicating how a cancer patient is functioning on a scale from 0 to 100 percent) on one particular test set of the LC1 dataset are shown for the linear and clinical kernel. This Figure clearly illustrates the non-linear behavior of the clinical kernel. Therefore, the clinical kernel is compared with the linear one, and with other non-linear kernels.

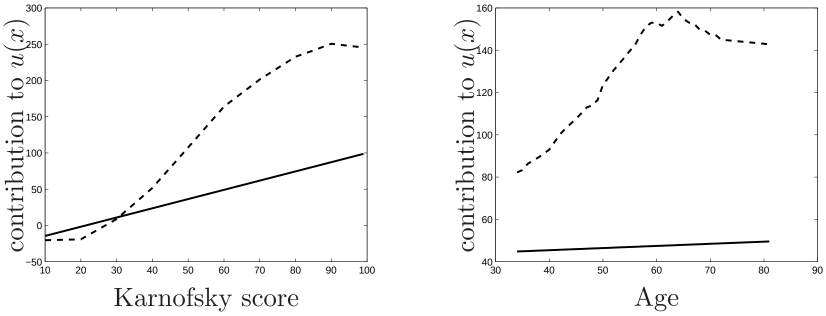

Figure 1: Estimated effects of 2 covariates for one particular training test set split of the LC1 dataset for the linear (solid line) and clinical kernel (dashed line).

Tables 2 and 3 summarize the performance on the test sets of all six datasets and 5 additive kernels. Only polynomial kernels of the second degree were considered. The logrank χ^2 statistic expresses the ability of the generated prognostic index to separate two groups. The median value of the prognostic index is taken as the threshold between the two groups. Comparing the linear and clinical kernel, clearly favors the latter. Comparing both polynomial kernels does not reveal a large improvement. Comparing the clinical kernel with the polynomial and RBF kernels, shows that the clinical kernel is able to obtain a performance which is comparable to that of other non-linear kernels. However, the clinical kernel has the advantage that it has no kernel tuning parameter.

Table 2: Median concordance index on 100 randomizations between training, validation and test set. The best performing model is indicated in bold. Statistical significant differences between the clinical and all other kernels were tested with the Wilcoxon rank sum test and indicated as: *,^{o} if $p < 0.05$, **,^{oo} if $p < 0.01$ and ***,^{ooo} if $p < 0.001$. Differences in favor of the clinical kernel are indicated with *, differences in favor of the other kernels are indicated with o.

data	lin	clin	poly	poly-clin	RBF
LE	0.65±0.05***	0.70±0.06	0.69±0.06	0.70±0.04	**0.71±0.05**
LC1	0.69±0.05	**0.70±0.05**	**0.70±0.04**	**0.70±0.04**	0.68±0.05***
LC2	**0.62±0.05**^{o}	0.61±0.05	0.57±0.05***	0.60±0.05	0.61±0.05
PC	0.73±0.05***	**0.78±0.03**	0.76±0.03**	**0.78±0.03**	0.76±0.03**
BC	0.62±0.03***	**0.68±0.02**	**0.68±0.02**^{o}	**0.68±0.02**	0.67±0.02
KT	0.55±0.12***	0.64±0.04	0.65±0.07	0.64±0.04	**0.66±0.03**^{o}

Table 3: Median logrank χ^2 on 100 randomizations between training, validation and test set. The best performing model is indicated in bold. Statistical significant differences between the clinical and all other kernels were tested with the Wilcoxon rank sum test and indicated as: *,^{o} if $p < 0.05$, **,^{oo} if $p < 0.01$ and ***,^{ooo} if $p < 0.001$. Differences in favor of the clinical kernel are indicated with *, differences in favor of the other kernels are indicated with o.

data	lin	clin	poly	poly-clin	RBF
LE	2.07±4.03***	**8.17±6.69**	4.25±3.95***	5.88±3.78***	6.50±4.67
LC1	3.78±5.88***	7.95±6.42	7.19±5.99	**8.06±6.67**	4.64±5.37***
LC2	**2.87±3.30**^{oo}	1.93±2.22	1.04±2.47	1.78±2.44	2.01±2.91
PC	5.06±5.08***	12.88±6.02	10.54±5.63*	**13.27±5.79**	10.36±5.87*
BC	7.16±5.88***	17.71±8.90	**25.50±8.96**^{ooo}	20.14±8.02^{oo}	19.14±7.35
KT	3.92±6.09***	10.79±5.52	**11.29±5.30**	9.32±5.56	11.84±5.13^{o}

5 Conclusions

This work compared the performance within a kernel based survival model of the linear versus the clinical kernel. On the 6 datasets used here, the performance was improved by using the clinical kernel. However, in contradiction to the linear kernel, the clinical kernel is a non-linear kernel and clinical interpretation becomes more difficult. The polynomial kernel was adapted in the same way as the linear one, to obtain a clinical polynomial kernel. After comparison of linear, clinical, polynomial and RBF kernels, we conclude that the clinical kernel is an easy and handy kernel which can be used to obtain non-linear models in survival analysis without the need to tune a kernel parameter.

Acknowledgments

This research is supported by Research Council KUL: GOA-AMBioRICS, GOA MaNet, CoE EF/05/006, IDO 05/010, IOF-KP06/11, IOF-SCORES4CHEM; Flemish Government: G.0407.02, G.0360.05, G.0519.06, G.0321.06, G.0341.07, G.0452.04, G.0499.04, G.0211.05, G.0226.06, G.0302.07, McKnow-E, Eureka-Flite; Belgian Federal Science Policy Office: IUAP P6/04; EU: FP6-2002-LIFESCIHEALTH 503094, IST-2004-27214, FP6-MC-RTN-035801; Prodex-8 C90242; EU: ERNSI. V. Van Belle is supported by a grant from the IWT.

References

[1] Daemen A. and De Moor B. Development of a kernel function for clinical data. In *the 31th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5913–5917, Minneapolis, Minnesota, September 2009.

[2] Van Belle V., Pelckmans K., Suykens J.A.K., and Van Huffel S. Learning Transformation Models for Ranking and Survival Analysis. Technical report, 09-45, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2009.

[3] Van Belle V., Pelckmans K., Suykens J.A.K., and Van Huffel S. Additive survival least squares support vector machines. *Statistics in Medicine*, 29(2):296 – 308, 2010.

[4] Van Belle V., Pelckmans K., Suykens J.A.K., and Van Huffel S. Support vector methods for survival analysis in clinical applications: a combined ranking-regression approach. Technical report, 09-235, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2009.

[5] Harrell F., 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.

[6] Genton M. Classes of kernels for machine learning: A statistics perspective. *Journal of Machine Learning Research*, 2:299–312, 2000.

[7] Emerson S.S. and Banks P.L.C. *Case Studies in Biometry*, chapter Interpretation of a leukemia trial stopped early, pages 275–299. Wiley-Interscience, 1994.

[8] Prentice R.L. A log gamma model and its maximum likelihood estimation. *Biometrika*, 61(3):539–544, 1974.

[9] Therneau T.M. and Grambsch P.M. *Modeling Survival Data: Extending the Cox Model*. Springer, 2 edition, 2000.

[10] 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.

[11] Byar D. and Green S. Prognostic variables for survival in a randomized comparison of treatments for prostatic cancer. *Bulletin du Cancer*, 67:477–490, 1980.

[12] Klein J.D. and Moeschberger M.L. *Survival Analysis. Techniques for censored and truncated data.* New York: Springer, 1997.

[13] Xavier de Souza S., Suykens J.A.K., Vandewalle J., and Bolle D. Coupled simulated annealing. *IEEE Transactions on Systems, Man, and Cybernetics - Part B.* In press.