ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

Survival Analysis with Cox Regression and Random Non-linear Projections

Samuel Branders[1], Benoît Frénay[2] and Pierre Dupont[1]

1- Université catholique de Louvain - ICTEAM/INGI - Machine Learning Group
Place Sainte Barbe 2, 1348 Louvain-la-Neuve - Belgium

2- Université de Namur - Faculty of Computer Science
Rue Grandgagnage 21, 5000 Namur - Belgium

Abstract. Proportional Cox hazard models are commonly used in survival analysis, since they define risk scores which can be directly interpreted in terms of hazards. Yet they cannot account for non-linearities in their covariates. This paper shows how to use random non-linear projections to efficiently address this limitation.

1 Introduction

Survival analysis is a class of statistical methods for studying the occurrence and timing of events. Such events include e.g. relapse, metastasis or death in cancer studies. One of the specific issues with survival data is the censoring. A patient is censored if he disappears or leaves the study before the event of interest occurs. Cox models [1] can be used in such settings to relate covariates, such as gene expression values, to the time to occurrence. However, Cox models cannot handle as such non-linearities in their covariates, what may restrict their usefulness in some settings.

This paper aims to show that Cox models can easily handle non-linear relationships if one uses random non-linear projections. Such tools have been used in extreme learning [2] to obtain results which are close to those of support vector machines, but at a much smaller computational cost. Random projections are used here with Cox models and a feasibility study is performed. Results are comparable to those of standard Cox models, but the proposed method can be used to handle data with non-linear relationships.

The remaining of this paper is organized as follows. Sections 2 and 3 review survival analysis and show the interest of Cox models for this problem. Section 4 explains what random non-linear projections are and how to use them. Section 5 details the proposed methodology to extend the Cox model, which is experimentally assessed in Section 6. We conclude this work in Section 7.

2 Survival Analysis

In survival analysis, each instance $i \in \{1, \dots, n\}$ is characterized by a 3-tuple $(t_i, \delta_i, \mathbf{x}_i)$ where $\mathbf{x}_i$ contains the d covariates and t_i is either the time of the event (such as metastasis or death) when $\delta_i = 1$ or the censoring time when $\delta_i = 0$. For each patient i, the objective is to model its associated hazard $h_i(t)$. This time depending function gives the probability of a patient i to have the event at time t knowing that he has not yet experienced the event before.

Such a framework is very common in cancer research and clinical studies. Many specific techniques exist to handle survival data such as survival-SVMs [3], partial logistic artificial neural networks [4], *etc.* The *standard* method to deal with survival data is the Cox proportional hazard model [1]. The Cox model is a generalized linear model. It has the advantages of being both efficient (fitting the model is a convex problem) and easily interpretable in terms of hazard.

3 Cox Regression

The Cox regression assumes the hazard $h_i(t)$ and $h_k(t)$ of any pair of instances (i, k) to be proportional [1]. The hazard of a patient can then be rewritten as the product of a baseline hazard $h_0(t)$ and a positive function of the covariates:

$$h_i(t) = h_0(t) \exp\left(\boldsymbol{\beta}^\top \mathbf{x}_i\right). \tag{1}$$

Consequently, the partial likelihood of the Cox model can be written as

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\exp\left(\boldsymbol{\beta}^\top \mathbf{x}_i\right)}{\sum_{k \in R(t_i)} \exp\left(\boldsymbol{\beta}^\top \mathbf{x}_k\right)} \right]^{\delta_i} \tag{2}$$

where $R(t_i) = \{k | t_k \geq t_i\}$ is the set of patients still at risk right before time t_i.

Here, the R package in [5] is used to train the Cox model, where an l_2 (ridge) penalty is added to the log-likelihood of the model to prevent overfitting. The risk score, $r_i = \boldsymbol{\beta}^\top \mathbf{x}_i$, is the final result of the Cox model. It can be directly interpreted in terms of hazard function for each patient : $h_i(t) = h_0(t) \exp(r_i)$.

4 Non-linear Random Projections

A potential limitation of the Cox model is that it cannot deal with non-linear relationships. Hence, a natural extension consists in adding support for features which must be non-linearly transformed to compute the hazard function. Many approaches exist in machine learning to obtain non-linear models like kernels or neural networks. However, kernelized Cox models [6] and survival-SVMs [3] come with the additional complexity of defining an appropriate non-linear kernel, whereas survival neural networks [4] are slow to learn. This paper focuses on a different approach which allows one to keep the interpretability and simplicity of Cox regression.

In extreme learning, it has been shown that random non-linear projections of the inputs [2] can be used to achieve state-of-the-art results in both non-linear classification and regression [7]. Those non-linear projections are obtained independently from training data: only their dimensionality d and the number of non-linear projections m must be known. The p-th projection is defined as:

$$z_p(\mathbf{x}_i) = \sigma\left(\sum_{j=1}^{d} W_{jp} x_{ij} + b_p\right) \tag{3}$$

where σ is a non-linear function, W_{jp} is the weight between the j-th input x_{ij} and the p-th projection and b_p is the bias used for the p-th projection. Non-linear projections could be optimized but Huang et al. [2] have shown that one can simply (i) draw the weights and biases in Equation (3) randomly (*e.g.* from a uniform or Gaussian distribution) and (ii) keep them fixed during learning.

The advantage of the above strategy is that state-of-the-art results are obtained in non-linear classification and regression [7] at the cost of linear methods. Indeed, the matrix of inputs $\mathbf{X}$ is replaced by the matrix of random non-linear projections

$$\mathbf{Z} = \begin{pmatrix} z_1(\mathbf{x}_1) & \cdots & z_m(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ z_1(\mathbf{x}_n) & \cdots & z_m(\mathbf{x}_n). \end{pmatrix} \tag{4}$$

Afterwards, fast, linear methods (linear regression, linear SVMs, *etc*) can be used with $\mathbf{Z}$ instead of $\mathbf{X}$. Using random non-linear projections offers a good compromise between computation needs and prediction accuracy. This view has been popularized in [8, 9, 10] where it is shown that the number of random projections can be set to a large number (*e.g.* $m = 10^3$) or even be infinite [11] if regularization is used to control the model complexity.

5 Proposed Methodology

This paper proposes to use random non-linear projections as input to a Cox model rather than the original covariates. The main advantage is that non-linear relationships can now be modeled, while the interpretability of the Cox model output is preserved in terms of hazard. Also, contrarily to *e.g.* SVMs, we do not need to choose a kernel nor to tune its parameters. Since works like [8, 9, 10, 11] show that regularized linear methods work well with large numbers of random non-linear projections, l_2 regularization can be used to control the complexity of the resulting non-linear Cox model. Weights and biases are here drawn from a uniform distribution between -2 and 2, and the inputs are normalized before being non-linearly transformed. Section 6 assesses this methodology.

6 Experiments

This section validates the use of random non-linear projections with a Cox model. Experiments are performed on synthetic and real datasets; performances in survival regression are assessed according to the concordance index (C-index). The C-index lies between 0 and 1 and measures to which extent the risk scores are concordant with the time to event, that is, whether a patient with a higher risk actually experience the event before a patient with a lower risk [12]. A poor model is expected to have a C-index around 0.5. The l_2 regularization constant λ of the Cox model is tuned with 10-fold cross-validation on the training set.

A 10-fold cross-validation is used in all experiments with real datasets. All results are reported in forest plots containing: the average test performance in C-index for each model and the p-values of a paired t-test against the standard Cox proportional hazard model. The black squares are centered on the average C-index. The horizontal grey lines correspond to the 95% confidence intervals.

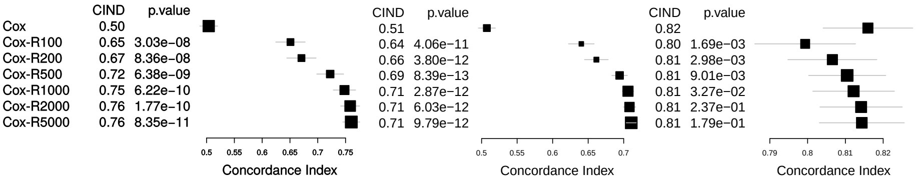

Figure 1: Results in C-index on synthetic data sets. Left, center and right plots respectively for f_1, f_2 and f_3.

6.1 Results on Artificial Non-linear Datasets

Artificial data are first considered to assess to which extent our approach is able to deal with non-linear features. A data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ is drawn from a standard distribution $\mathcal{N}(0, 1)$. The survival data (t_i, δ_i) are generated from two Weibull distributions (for event and censoring times) such that the hazard $h_i(t)$ depends on a combination $f(\mathbf{x}_i)$ of the features : $h_i(t) \propto \exp(f(\mathbf{x}_i))$. The Weibull shape parameters are set to 2.5 and 1 respectively for censoring and event times. The scaling parameters are 2914 and $20000 \exp(-1.5 f(\mathbf{x}_i)/\sigma)$, where σ is the standard deviation of $f(\mathbf{x}_i)$ over all generated samples[1]. Two non-linear combinations and one linear combination of features are considered here:

$$f_1(\mathbf{x}_i) = \sum_{j=1}^{d} x_{ij}^2, \quad f_2(\mathbf{x}_i) = \sum_{j=1}^{d} \exp(-x_{ij}^2) \quad \text{and} \quad f_3(\mathbf{x}_i) = \sum_{j=1}^{d} a_j x_{ij}. \tag{5}$$

Results are averaged in Figure 1 over 10 independent runs with $n = 1000$ instances (200 for training, 800 for validation) and $d = 5$ features. The Cox proportional hazard model is trained (i) on the 5 original features and (ii) using between 100 and 5000 random non-linear transformations of those features.

As expected, a standard (linear) Cox model is not able to deal with non-linear features (f_1 and f_2). The Cox model offers significantly better results when the original features are first transformed through non-linear random projections. Such a strategy is even not detrimental in the linear case (f_3). If a sufficiently large number (here 2000) of random projections is considered, the results are not significantly different from those of a standard Cox model. In general, the number of random projections to consider needs not be carefully tuned provided it is chosen large enough.

6.2 Results on Real-World Cancer Datasets

This section shows results for three real-world cancer datasets. The first real dataset is the flchain[2] dataset, which contains 8 features for 7874 patients. Multiple causes of death were recorded and the death due to a circulatory system diseases is considered here. Others causes of death are seen as censoring, which is one way to deal with competing risks [13]. The second dataset consists of

[1] Those values were chosen to produce events and censoring times similar to real data.

[2] available in the survival R package http://cran.r-project.org/web/packages/survival/

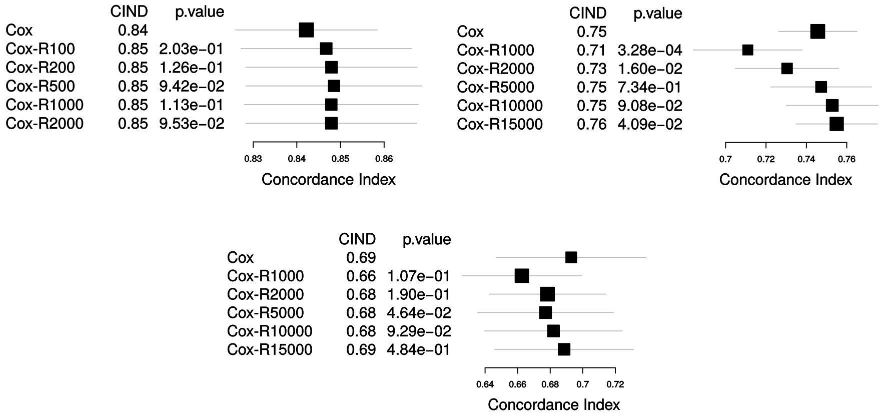

Figure 2: Results in C-index with the flchain, breast and colon datasets.

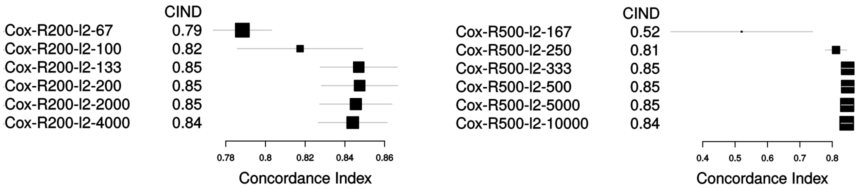

Figure 3: Results with the flchain dataset changing the l_2 regularization constant. Left and right plots respectively for 200 and 500 random projections.

five pooled breast cancer datasets from the GEO database (accession numbers: GSE2034, GSE5327, GSE7390, GSE2990, GSE11121 and GSE6532). 75% of the features with the lowest variances are removed, which is a standard pre-filtering of such high-dimensional data. The final dataset contains 1054 patients with 5571 features. The third dataset consists of seven pooled colon cancer datasets from the GEO database (accession numbers: GSE39582, GSE17536, GSE17537, GSE14333, GSE29621, GSE29623 and GSE38832). After a similar pre-filtering, the final dataset contains 1234 patients with 13669 features.

Figure 2 shows results obtained for the flchain, breast and colon datasets. The C-index reaches a plateau when the number of projections increases. Globally these results do not exhibit statistically significant differences with those of a standard Cox model. They illustrate that the proposed approach is effective even though explicit non-linearities are not required for these datasets.

The sensitivity of results to the choice of the l_2 regularization parameter λ is studied in Figure 3 using the flchain dataset. The number of random projections is fixed to 200 and 500 and results are reported with λ equal to $\{\frac{1}{3}, \frac{1}{2}, \frac{2}{3}, 1, 10, 20\}$ times the number of dimensions. Results are improving while increasing λ and reach a plateau, here when λ is roughly equal to the number of projections. The choice of λ seems robust and it does not seem difficult to tune.

7 Conclusion

This paper shows how random non-linear projections used in extreme learning can also be used to extend Cox models. Using the proposed methodology, survival analysis can be performed even with non-linear relationships between covariates and the associated risk scores. The computational cost is comparable to the cost of learning standard Cox models. Since Cox models are essentially used to compute risk scores, the results are still readily interpretable in terms of hazards. Such an approach avoids the additional complexity of defining an appropriate non-linear kernel or of training complex neural networks. Our future work includes testing the methodology of this preliminary work on survival data with significant non-linear effects.

References

[1] DR Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B*, 34(2):187–220, 1972.

[2] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1-3):489–501, 2006.

[3] V Van Belle, K Pelckmans, S Van Huffel, and J a K Suykens. Improved performance on high-dimensional survival data by application of Survival-SVM. *Bioinformatics (Oxford, England)*, 27(1):87–94, January 2011.

[4] Elia Biganzoli, Patrizia Boracchi, Luigi Mariani, and Ettore Marubini. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in medicine*, 17(10):1169–1186, 1998.

[5] Jelle J Goeman. L1 penalized estimation in the Cox proportional hazards model. *Biometrical journal. Biometrische Zeitschrift*, 52(1):70–84, February 2010.

[6] Hongzhe Li and Yihui Luan. Kernel Cox regression models for linking gene expression profiles to censored survival data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 65–76, 2003.

[7] G.B. Huang, D.H. Wang, and Y. Lan. Extreme learning machines: a survey. *International Journal of Machine Learning and Cybernetics*, 2(2):107–122, 2011.

[8] Qiuge Liu, Qing He, and Zhongzhi Shi. Extreme support vector machine classifier. In *Proceedings of the 12th Pacific-Asia conference on Advances in knowledge discovery and data mining*, pages 222–233, Berlin, Heidelberg, 2008.

[9] Y. Miche, A. Sorjamaa, P. Bas, O. Simula, C. Jutten, and A. Lendasse. Op-elm: Optimally pruned extreme learning machine. *Neural Networks, IEEE Transactions on*, 21(1):158–162, December 2009.

[10] Benoît Frénay and Michel Verleysen. Using svms with randomised feature spaces: an extreme learning approach. In *Proceedings of The 18th European Symposium on Artificial Neural Networks (ESANN)*, pages 315–320, 2010.

[11] Benoît Frénay and Michel Verleysen. Parameter-insensitive kernel in extreme learning for non-linear support vector regression. *Neurocomputing*, 74(16):2526 – 2531, 2011.

[12] Frank E Harrell, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15:361–387, 1996.

[13] RL Prentice and JD Kalbfleisch. The analysis of failure times in the presence of competing risks. *Biometrics*, 34(4):541–554, 1978.