Proceedings of the 29th Annual International
Conference of the IEEE EMBS
Cité Internationale, Lyon, France
August 23-26, 2007.

Continuous and Discrete Time Survival Analysis: Neural Network Approaches

A. Eleuteri, M. S. H. Aung, A. F. G. Taktak, B. Damato, P. J. G. Lisboa

Abstract— **In this paper we describe and compare two neural network models aimed at survival analysis modeling, based on formulations in continuous and discrete time. Learning in both models is approached in a Bayesian inference framework. We test the models on a real survival analysis problem, and we show that both models exhibit good discrimination and calibration capabilities. The C index of discrimination varied from 0.8 (SE=0.093) at year 1, to 0.75 (SE=0.034) at year 7 for the continuous time model; from 0.81 (SE=0.07) at year 1, to 0.75 (SE=0.033) at year 7 for the discrete time model. For both models the calibration was good ($p<0.05$) up to 7 years.**

I. Introduction

Survival analysis is used when we wish to study the occurrence of some event in a population of subjects and the time until the event is of interest. This time is called *survival time* or *failure time*. Survival analysis is often used in industrial life-testing experiments and in clinical follow-up studies.

In literature there are many different modelling approaches to survival analysis. Parametric models based on specified families of distributions may involve too strict assumptions on the failure times which usually extremely simplify the experimental evidence, particularly in the case of medical data [1]. Semiparametric models do not make assumptions on the distributions of failures, but make instead assumptions on how the system features influence the survival time (the usual assumption being the proportionality of hazards); furthermore, usually these models do not allow for direct estimation of survival times. Finally, nonparametric models usually only allow for a qualitative description of the data on the population level.

Neural networks have recently been used for survival analysis; for a survey on the current use of neural networks we refer to [2], [3]. The only neural network architectures aimed at survival analysis and trained in a Bayesian framework are described in [4], [5], [6], [7].

Manuscript received April 2, 2007. This project is funded by the Biopattern Network of Excellence FP6/2002/IST/1; proposal N. IST-2002-508803; Project full title: Computational Intelligence for Biopattern Analysis is Support of eHealthcare; URL:www.biopattern.org.

A. Eleuteri and A. F. G. Taktak are with the Department of Clinical Engineering, Royal Liverpool University Hospital, Liverpool, UK (phone: 44-0151-706-4214; fax: 44-0151-706-5803; e-mail: antonio.eleuteri@gmail.com).

M. S. H. Aung and P. J. G. Lisboa are with the School of Computing and Mathematical Sciences, Liverpool John Moores University, Liverpool, UK.

B. Damato is with the St. Paul's Eye Unit, Royal Liverpool University Hospital, Liverpool, UK.

Neural networks provide efficient parametric estimates of survival functions, and, at least in principle, the capability to give personalised survival predictions. In a medical context, such information is valuable both to clinicians and patients. It helps clinicians to choose appropriate treatment and plan follow-up efficiently. Patients at high risk could be followed up more frequently than those at lower risk in order to channel valuable resources to those who need them most. For patients, obtaining information about their prognosis is also valuable for planning their lives.

In this paper we introduce a new class of models for continuous time modelling, and compare its performance with an existing model which shares some similarities but is aimed at discrete time modelling [6], [8].

In Section II we describe the survival analysis problem in discrete and continuous time; in Section III we provide details about the two neural network models and the Bayesian approach to modelling, whereas in Section IV a sample application on real data is shown.

II. Continuous And Discrete Time Survival Analysis

A. Definitions

Let T denote a positive random variable, with distribution function P, representing the time of occurrence of an event. The survival function, $S(t)$, is defined as:

$$S(t)=\Pr(T>t), \tag{1}$$

that is, the probability of not experiencing the event up to (and including) time t. We shall assume that the survival function also depends on a set of covariates, represented by the vector x. Depending on whether T is absolutely continuous or discrete, we can introduce the *hazard rate* and *hazard probability* functions [1]:

$$h_r(t)= P(t)'/(1-P(t)) \tag{2}$$

$$h_p(t)= (P(t)- P(t-1))/(1- P(t)) = \Pr(T=t \mid T>=t) \tag{3}$$

where P' is the first derivative of the distribution function P.

B. Censoring

In many survival analysis applications we do not directly observe realisations of the random variable T; therefore we must deal with a missing data problem. The most common

form of is *right censoring*, i.e. we observe realisations of the random variable:

$$Z=\min(T,C) \tag{4}$$

where C is a random variable whose distribution is usually unknown. We shall use a censoring indicator d to denote whether we have observed an event (d=1) or not (d=0). It can be shown that inference does not depend on the distribution of C [1].

C. Likelihood functions

Log-likelihood functions for the discrete and absolutely continuous cases can be obtained in terms of hazard functions [1].

For the continuous and discrete time cases, we have respectively:

$$L_r = \sum_i d_i \log h_r(t_i, x_i) - \int_0^{t_i} h_r(u, x_i)\mathrm{d}u\,,$$

$$L_p = \sum_i d_i \log h_p(x_i, t_i) + (1-d_i)\log(1-h_p(x_i, t_i))\,. \tag{5}$$

In the latter case we recognize the log-likelihood function for a logistic regression problem, in which some pre-processing of the data is necessary. In particular, each pattern x_i is replicated a number of times equal to the cardinality of the set $\{1,2,\ldots,t_i\text{-}1\}$, whereas the target d_i is set to zero for each of these replications. Therefore, the index i in the summation above should be intended over the set of pre-processed data.

III. Neural Network Models

The basic neural network model for both approaches is the Multi-Layer Perceptron (MLP) [9]:

$$a(t,x;w) = b_0 + \sum_k v_k\, g(u_k^T x + u_0 t + b_k)$$

where $g()$ is a sigmoid function, and $w=\{b_0, v, u, u_0, b\}$ is the set of parameters. Depending on the problem, the MLP defines a model for the logarithm of the hazard rate function in the continuous time case:

$$a(t,x;w) \triangleq \log h_r(t,x)$$

whereas in the discrete case the MLP models the log-odds ratio of the hazards:

$$a(t,x;w) \triangleq \log \frac{h_p(t,x)}{1-h_p(t,x)}$$

We refer to the continuous time model as Conditional Hazard Estimating Neural Network (CHENN), whereas the discrete time model is called a Partial Logistic Artificial Neural Network (PLANN) [8]. In this paper, we use the Bayesian version of the PLANN model, called PLANN-ARD (PLANN with Automatic Relevance Determination) [6]. Previous attempts at modelling the continuous time survival function used an MLP with constraints on the weight space, so that the output of the network could be interpreted as a survival function [5]; the constraints on weight space however made the problem computationally difficult.

A. Bayesian learning

The Bayesian learning framework has several advantages over maximum likelihood methods [9], [10], since model overfitting is unlikely; the model is automatically regularized; and error bars can be obtained (at least in theory) to estimate the uncertainty in the predictions.

In the conventional maximum-likelihood approach to training, a single weight vector is found, which minimizes an error function; in contrast, the Bayesian scheme considers a probability distribution over weights w. This is described by a prior distribution $p(w)$ which is modified when we observe the data $D=\{(x,t)\}$. This process can be expressed by Bayes' theorem:

$$p(w \mid D) = \frac{p(D \mid w)\, p(w)}{p(D)}\,. \tag{6}$$

To evaluate the posterior distribution, we need expressions for the prior $p(w)$ (which is itself parameterized by *hyperparameters*) and for the likelihood $p(D|w)$.

The posterior distribution is usually very complex and multimodal, and the determination of the normalization factor (also called the *evidence*) is very difficult. Furthermore, the hyperparameters must be integrated out, since they are only used to determine the form of the distributions.

A solution is to integrate out the parameters separately from the hyperparameters, by making a Gaussian approximation; then, searching for the mode with respect to the hyperparameters. It turns out, as noted in [9], [10], that this procedure gives a good estimation of the *probability mass* attached to the posterior, in particular for distributions over high-dimensional spaces.

The full Bayesian treatment of inference implies that we do not simply get a pointwise prediction for functions $f(x,t;w)$ of a model output, but a full distribution. Such predictive distributions have the form:

$$p(f(x,t \mid D)) = \int f(x,t \mid w)\, p(w \mid D)\, dw\,. \tag{7}$$

The above integral is in general not analytically tractable,

even when the posterior distribution over the parameters is Gaussian. However, it is usually enough to find the moments of the predictive distribution, in particular its mean and variance, which can usually be obtained by approximation [9], [10]. We emphasize that it is important to evaluate first and second order information to understand the overall *quality* and *reliability* of a model's predictions. Error bars also provide hints on the distribution of the input patterns [11] and can therefore be useful to understand whether a model is extrapolating its predictions.

In our case, error bars on the hazard and survival functions can only be obtained for the CHENN model. For the PLANN-ARD model, current approximations only allow error bars on the log-odds ratio, which is formally equivalent to a regression network; for a complete description, see [6], [9], [10].

IV. An Application: Intraocular Melanoma Prognosis

A. The Problem

Intraocular melanoma occurs in a pigmented tissue called the uvea, with more than 90% of tumours involving the choroid, beneath the retina. About 50% of patients die of metastatic disease, which usually involves the liver. In this application the event of interest is all-cause mortality.

The data used to test the model were selected from the database of the Liverpool Ocular Oncology Centre [12]. The dataset was split into two parts: 1823 patterns for training, 781 patterns for test. Nine prognostic factors were used. For the PLANN-ARD model, the time to event was discretised in 1-year blocks as required by the model definition.

B. Survival Predictions

Both neural network models were trained and tested on the same data. The PLANN-ARD model has 10 hidden units, whereas the CHENN model has 4 hidden units; these numbers were chosen to allow moderate flexibility in the models, on the ground that the functions modeled do not typically exhibit high frequency behaviour.

In Fig. 1 we show the estimated population survival curves of the test data, for both neural networks; for comparison, we also show the Cox and KM estimates. As can be seen, the agreement with the KM estimate is quite good up to about 7 years for both models, whereas the Cox estimate seems to show an optimistic bias. In practice the main interest is in relatively short time predictions, from 1 to 7 years, so these performances can be deemed acceptable from an application point of view.

C. Discrimination and Calibration

The performance of survival analysis models can be assessed according to their discrimination and calibration capabilities. Discrimination is the ability of the model to correctly separate the subjects into different prognostic groups. Calibration is the degree of correspondence between the estimated probability produced by the model and the actual observed probability [13].

One of the most used methods for assessing discrimination in survival analysis is Harrell's C index [13], [14], which is an extension to survival analysis of the Area Under the Receiver Operator Characteristic. Calibration is usually assessed by goodness-of-fit testing procedures based on a chi-square statistic. This method however does not take into account censoring; therefore, we applied the less known

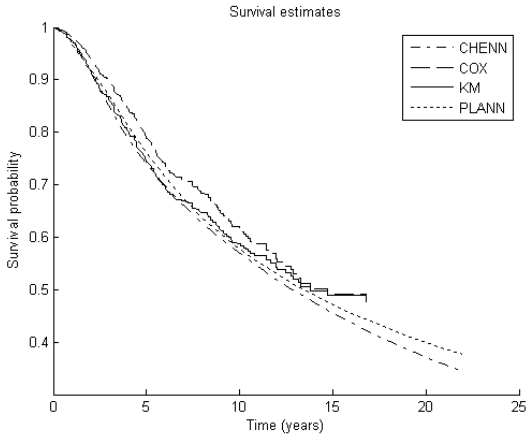

Fig. 1. Population survival probability estimates for different methods. For the CHENN and PLANN-ARD models these were evaluated as the average of the survival predictions for each pattern in the test set.

Kolmogorov-Smirnov (KS) test with corrections for censoring [15].

The C index was evaluated over the set of years {1,3,5,7}. The values are reported in Table I. As can be seen, there is no statistically significant difference between the CHENN and PLANN-ARD discrimination performance.

The KS test with corrections for censoring was applied for the same set of years, and up to the maximum uncensored time (16.8 years); the confidence level was set as usual at

TABLE I
C-Index (with standard error)

Year	CHENN	PLANN-ARD
1	0.80 (0.093)	0.81 (0.07)
3	0.80 (0.043)	0.79 (0.04)
5	0.77 (0.036)	0.77 (0.036)
7	0.75 (0.034)	0.75 (0.033)

For each year the discrimination capability of the two models have been assessed. There does not seem to be a statistically significant difference between the models.

0.05. For both the PLANN-ARD and CHENN models the null hypothesis that the modelled distributions follow the empirical estimate cannot be rejected for years 1 to 7, whereas it is rejected if we compare the distributions up to 16.8 years; the null hypothesis is always rejected for the Cox model.

V. Conclusion

In this paper a new neural network model for survival

analysis in a continuous time setting has been proposed, which approximates the logarithm of the hazard rate function. The model formulation allows computation of error bars on both hazard rate and survival predictions. This model is compared with a neural network, which models the log-odds ratio of the hazard probability in a discrete setting; the latter model however does not allow evaluation of error bars on survival predictions, but only on the log-odds ratios. Both models are trained in the Bayesian framework to reduce the risk of overfitting. The models have been tested on real data, to predict survival from intraocular melanoma; by using formal discrimination and calibration tests we have shown that both models have good performance within a time horizon of 7 years, which is found useful for the application at hand; whereas Cox's model exhibits an optimistic bias which might be dangerous in applications.

REFERENCES

[1] D. R. Cox, D. Oakes. *Analysis of Survival Data*. Chapman and Hall, 1984. Ch 1-4.

[2] B. D. Ripley, R. M. Ripley. "Neural Networks as Statistical Methods in Survival Analysis". *Artificial Neural Networks: Prospects for Medicine* (R. Dybowsky and V. Gant eds.), Landes Biosciences Publishers, (1998).

[3] G. Schwarzer, W. Vach, M. Schumacher. "On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology". *Statistics in medicine* **19**, pp. 541-561, (2000).

[4] B. Bakker, T. Heskes. "A neural-Bayesian approach to survival analysis". *Proceedings IEE Artificial Neural Networks*, pp. 832-837, (1999).

[5] A. Eleuteri, R. Tagliaferri, L. Milano, S. De Placido, M. De Laurentiis. "A novel neural network-based survival analysis model", *Neural Networks*, **16**, pp. 855-864, (2003).

[6] P. J. G. Lisboa, H. Wong, P. Harris, R. Swindell. "A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer". *Artificial intelligence in medicine*, **28**, pp. 1-25, (2003).

[7] R. M. Neal. *Survival Analysis Using a Bayesian Neural Network*. Joint Statistical Meetings report, Atlanta, (2001).

[8] E. Biganzoli, P. Boracchi, L. Mariani, E. Marubini. "Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach". *Statistics in Medicine*. **17**, pp. 1169-86, (1998).

[9] C. M. Bishop. *Neural networks for pattern recognition*. (Oxford University Press Inc. New York. 1995). Ch. 4, 10.

[10] D. J. C. MacKay, "The evidence framework applied to classification networks". *Neural Computation*, **4 (5)**, pp. 720-36, (1992).

[11] C. K. I. Williams, C. Qazaz, C. M. Bishop, H. Zhu. "On the relationship between Bayesian error bars and the input data density". *Proceedings of the 4th International Conference on Artificial Neural Networks*, pp. 160-165, Cambridge (UK), (1995).

[12] A. F. G. Taktak, A. C. Fisher, B. Damato, "Modelling survival after treatment of intraocular melanoma using artificial neural networks and Bayes theorem". *Physics in Medicine and Biology*. **49**, pp. 87-98, (2004).

[13] S. Dreiseitl, L. Ohno-Machado. "Logistic regression and artificial neural network classification models: a methodology review". *J.Biomed.Inform.*, **vol. 35, no. 5-6**, pp. 352-359, (2002).

[14] F. E. Harrell Jr., R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati. "Evaluating the yield of medical tests". *JAMA.*, **vol. 247, no. 18**, pp. 2543-2546, 1982.

[15] J. Koziol. "Goodness-of-fit tests for randomly censored data". *Biometrika* **67 (3)**, pp. 693-696, (1980).