Neural network analysis to predict treatment outcome in patients with ovarian cancer

Marc Theeuwen†, Bert Kappen†, Jan Neijt‡
† RWCP[1] Novel Functions SNN[2] Laboratory, Laboratory for Medical Physics and Biophysics, Geert Grooteplein Noord 21, 6524 EZ Nijmegen, The Netherlands, E-mail marct@mbfys.kun.nl or bert@mbfys.kun.nl, Telephone +31-24-3614245
‡ Department of Internal Medicine, Utrecht University Hospital, PO Box 85500, 3508 GA Utrecht, The Netherlands

Abstract

The traditional technique to model survival probabilities is the Cox regression analysis [Cox and Oakes, 1984]. Recently, also neural networks have been applied for survival analysis and the prediction of prognosis in cancer treatment [Liestøl K, 1994]. The main advantages of the neural network approach are the relative ease with which time dependencies in prognostic factors can be obtained, the improved prediction performance on independent test data for large numbers of input parameters [Kappen and Neijt, 1993], and the potential to model non-linear relations using hidden units. Although this last point has only been established on artificial data [De Laurentiis and Ravdin, 1994], it probably constitutes the most important difference for future applications.

Neural networks are different from Cox's survival analysis in both the cost criterion that is optimized, the data model that is being used, and the treatment of censored patients. In this paper we analyze the differences between the two approaches theoretically. In addition, we illustrate the time dependent influence of the prognostic factors on a data base of 917 patients on patients with ovarian cancer.

[1] Real World Computing Partnership
[2] Dutch Foundation for Neural Networks

Survival Analysis

Consider a number of patients $\mu = 1, \ldots, N$. Let $[t_I, t_F]$ be the time interval during which the study takes place. Let t_μ^- denote the time at which patient μ enters the study ($t_I < t_\mu^- < t_F$). Let t_μ^+ denote the time at which patient μ dies. If $t_\mu^+ < t_F$, denote by $t_\mu = t_\mu^+ - t_\mu^-$ the *survival time* of patient μ. If $t_\mu^+ > t_F$, t_μ^+ is unknown and patient μ is called *censored* and t_μ is defined as $t_\mu = t_F - t_\mu^-$. Let each of the patients be characterized by some properties $\vec{z}_\mu$ at $t = t_\mu^-$. Survival analysis attempts to model a relation between patient characteristics $\vec{z}_\mu$ and survival time t_μ.

Without loss of generality, we can order the patients such that $t_1 < \ldots < t_N$. Denote by $t_\alpha, \alpha = 1, \ldots, M$ a number of observation times ($t_1 \leq t_\alpha \leq t_N$). Observations can be made at regular time intervals, as we will assume for neural networks, or can coincide with the survival times t_μ (M=N) as is usually the case for standard survival analysis.

Standard survival analysis

To obtain the influences of the prognostic factors and survival probabilities, the maximum of the log likelihood K_{SSA} is determined (see e.g. [Kalbfleisch and R.L., 1980]):

$$K_{SSA} = \sum_\alpha \left[\sum_{\mu \in D_\alpha} \log\left(p_{\alpha-1\mu} - p_{\alpha\mu}\right) + \sum_{\mu \in C_\alpha} \log\left(p_{\alpha\mu}\right) \right]. \qquad (1)$$

Here the index α labels the observation times t_α and SSA stands for standard survival analysis. D_α and C_α are the sets of patients, which die in the interval $(t_{\alpha-1}, t_\alpha)$, or are for the last time reported alive at t_α, respectively. Similarly, A_α is the set of patients who survive the interval $(t_{\alpha-1}, t_\alpha)$ and are either censored at some time $\tau_\mu > t_\alpha$ or survive the whole interval of interest. Note that for all α $A_\alpha + \sum_{\beta=1}^{\alpha}(C_\beta + D_\beta)$ is the total set of patients. The probability $p_{\alpha\mu}$, which is the survival function, models the probability to be alive at t_α. Therefore $p_{\alpha-1\mu} - p_{\alpha\mu}$ gives the unconditional probability to die in the interval $(t_{\alpha-1}, t_\alpha)$. Maximizing K_{SSA} is thus equal to maximizing the probability to die for patients in the interval in which they actually die, and to maximize the probability for censored patients to be alive at their last date of observation. Using the proportional hazards approach, $p_{\alpha\mu}$ is modelled as

$$p_{\alpha\mu} = p_{\alpha 0}^{exp(\vec{\beta}\cdot\vec{z}_\mu)}. \qquad (2)$$

$p_{\alpha\mu}$ is called the baseline survival probability and is only time dependent, $\vec{\beta}\vec{z}$ is called the prognostic index and determines the increased ($\vec{\beta}\vec{z} < 0$) or decreased ($\vec{\beta}\vec{z} < 0$) survival probability relative to the baseline. For an optimal estimation of survival probabilities a joint estimation of $p_{\alpha 0}$ and $\vec{\beta}$ in equation 1 should be carried out. In practice, however, one is normally more interested in the relative order in which patients die, which is given by $\vec{\beta}$ and therefore uses a simpler cost criterion which is independent of $p_{\alpha 0}$ to estimate $\vec{\beta}$. Afterwards this estimation of $\vec{\beta}$ might be used in equation 1 to estimate $p_{\alpha 0}$ (see e.g. [Kalbfleisch and R.L., 1980]).

Neural networks survival analysis

In neural networks survival analysis, fixed observation times are considered. For each observation time, the population of patients is divided in those that are alive, A_α, those that are dead, $\sum_{\beta=1}^{\alpha} D_\beta$, and those that are censored $\sum_{\beta=1}^{\alpha} C_\beta$. This defines M classification problems, each of which can be approximated with neural networks.

Recently, new classes of neural networks have been developed that allow a probabilistic classification for Multi Layer Perceptrons [Levin et al., 1990] and for Boltzmann Perceptrons [Kappen, 1995]. The criterion that is minimized during training, and that is used for evaluating the predictive quality of the network solution on the independent test set, is the Kullback divergence, which is the optimal criterion for classification according to information theory, between the desired probability and the probability given by the network:

$$K_\alpha = \frac{1}{P} \sum_\mu \left(P_{\alpha\mu} \log \frac{P_{\alpha\mu}}{p_{\alpha\mu}} + (1 - P_{\alpha\mu}) \log \frac{(1 - P_{\alpha\mu})}{(1 - p_{\alpha\mu})} \right) \tag{3}$$

where $P_{\alpha\mu}$ and $p_{\alpha\mu}$ are the survival probabilities as given by the data and by the neural network, respectively. Note that K_α has a lower bound equal to zero if these probabilities are equal. During the learning phase of the neural network it tries to minimize K_α. $P_{\alpha\mu} = 0, 1$ for patients μ that have died and are alive at t_α. For patients that are censored before t_α we do not know whether they are alive or dead at t_α. We estimate the probability that they are dead at t_α from the fraction of patients that are dead at t_α and that are alive at t_μ: $P_{\alpha\mu} = \frac{1}{N_{A_\alpha}} \sum_{\nu \in A_\alpha} P_{\alpha\nu}$. Since the right hand side may also contain censored patients, this procedure must be applied to the longest living censored patients first.

Only for the extreme of only one observation α, is the same optimisation done by both the standard survival analysis approach with log likelihood and the neural network approach. For all other cases both approaches are inherently different. Also the probability functions that are used differ. In stead of the proportional hazards assumption, we used a Boltzmann distribution in which the influences of the prognostic factors for each of the M classification problems are determined independently. Thus the time dependent influences of these prognostic factors could be studied.

Comparison between standard survival analysis and neural networks

The differences between standard survival analysis and the neural network approach are several. First, the use of the cost criterion is different. The reason is that SSA models one survival process in time, and the neural network approach models M classification tasks. Both approaches attempt to model the survival function $p_{\alpha\mu}$. Secondly, although both approaches attempt to model the survival function $p_{\alpha\mu}$, SSA often makes the proportional hazard assumption and subsequently only calculates $\vec{\beta}$. Thus, most available implementations of SSA do not give the survival function. This makes comparison with neural networks complicated. Neural networks model the survival function for each value of α by a different network. This has as advantages, that 1) very complex relations can in principle be modeled 2) no temporal structure (such as proportional

hazard) is assumed, which allows for flexible modeling of temporal aspects. Disadvantages are that censored patients are treated in a less elegant manner and the survival function is not guaranteed to be non-increasing as a function of α. However, we do not consider these as major disadvantages. Neither of these aspects has lead to complications in our practical studies.

Results

We constructed a database including 917 patients from 4 studies, two studies from The Netherlands Joint Study Group for Ovarian Cancer, (see respectively [Neijt et al., 1984, Neijt et al., 1987]) and two from the Gynecological Cancer Cooperative Group of the European Organisation for Research and Treatment of Cancer (EORTC). For each year we trained a Boltzmann Perceptron to classify the patients into the classes alive and dead, thereby modelling the survival probability for each year. As error criterion the Kullback divergence was used. In Table 1 the results on the training and independent test set are listed. In addition the results of the 0-hypothesis that there is no relationship between the patient characteristics and the survival probability is shown as well.

A pilot comparison between Cox's survival analysis method and the Boltzmann Machine showed that the neural network performed slightly better. This can mainly be attributed to the paradigm of training and test set, which leads to a better prediction on independent data than standard statistical tests assuming Gaussian distributions of errors. Another reason for the better performance of neural networks is that the proportionality condition of the hazard is not assumed (see e.g. Liestøl et al. 1994).

The robustness of the method was demonstrated by training the Boltzmann Perceptron on the Dutch data base and using this network for prediction of the survival of the EORTC data. The predictions of the neural network were equally good for the EORTC data as for the Dutch data.

We carefully reduced the set of prognostic factors in order to obtain a minimal set of prognostic factors with a maximal predictive value. First all patient characteristics were normalized such that the magnitudes of their weights could be compared. The neural network was trained until convergence was obtained, after which the patient characteristic with the smallest absolute weight was removed. This procedure was repeated until the performance on the independent test set deteriorated. The minimal set obtained in this way (consisting

Kullback/1000	1 y	2 y	3 y	4 y	5 y	6 y	7 y	8 y
train	563	481	399	360	322	312	297	281
test	575	502	414	365	344	325	320	289
0	632	577	482	434	399	380	361	338
% correct	1 y	2 y	3 y	4 y	5 y	6 y	7 y	8 y
train	65.1	72.2	77.6	81.3	83.9	84.3	85.2	85.2
test	64.5	70.4	77.6	80.9	83.1	84.6	84.6	85.9
0	54.4	66.1	75.6	79.4	82.0	83.2	84.2	84.8

Table 1: The Kullback divergence and the percentage of correct survival predictions as a function of time (years).

of performance, grading, histological cell type, number of leucocytes, number of residual tumors after debulking, diameter of residual tumors after debulking and figo stage) had a higher predictive accuracy than the complete set of 76 prognostic factors. The performance on the training set, however, was less than with the complete data set. This reduction to the minimal set of prognostic factors is important for diagnostic purposes.

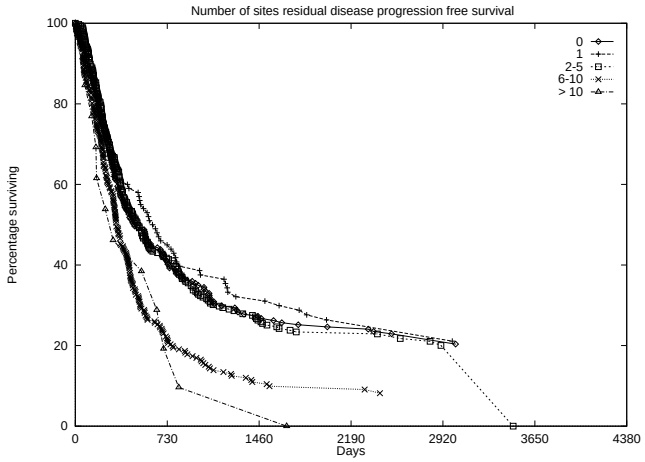

Figure 1: Survival curves for different number of sites of residual disease.

Additional information is given by the time dependence of the importance of these prognostic factors. This effect of time-dependent prognosis can be seen in the survival curves. For example the variable representing the number of sites of residual disease has only a small impact on the survival probability after 1 year, but a significant impact on the later years (see figure 1).

Conclusions

Using Boltzmann Perceptrons to model survival probability for each year independently, time-varying prognostic factors can successfully be modelled. In principle this can also be incorporated into proportional hazards models, but in practice this is never done. We selected a minimal set of prognostic factors with maximal predictive value using the confidence intervals of the prognostic factors as selection criterium. We concluded that improved short-term survival is determined by a good performance status, a low number of leucocytes, a small diameter of redidual disease and a low FIGO stage (stage III). Long term survival is mainly determined by FIGO status (stage III), grade (well differentiated), cell type (serous) and low number of sites of residual disease. A pilot comparison between our approach and a standard Cox proportional hazard model indicated that neural networks are slightly better performing.

References

[Cox and Oakes, 1984] Cox, D. and Oakes, D. (1984). *Analysis of Survival Data*. Monographs on Statistics and Applied Probability. Chapman and Hall, London.

[De Laurentiis and Ravdin, 1994] De Laurentiis, M. and Ravdin, P. (1994). A technique for using neural network analysis to perform survival analysis of censored data. *Cancer Letters*, 77:127–138.

[Kalbfleisch and R.L., 1980] Kalbfleisch, J. and R.L., P. (1980). *The statistical analysis of failure time data*. Wiley, New York.

[Kappen, 1995] Kappen, H. (1995). Deterministic learning rules for Boltzmann machines. *Neural Networks*, 8:537–548.

[Kappen and Neijt, 1993] Kappen, H. and Neijt, J. (1993). Neural network analysis to predict treatment outcome. *Annals of Oncology*, 4:31–34. Suppl. 4.

[Levin et al., 1990] Levin, E., Tishby, N., and Solla, S. (1990). A statisitical approach to learning and generalization in layered neural networks. *Proceedings IEEE*, 78:1568–1574.

[Liestøl K, 1994] Liestøl K, Kragh Andersen P, A. U. (1994). Survival analysis and neural nets. *Stat in Medicine*, 13:1189–1200.

[Neijt et al., 1984] Neijt, J., Ten Bokkel Huinink, W., and Van der Burg, M. e. a. (1984). Randomised trial comparing two combination chemotherapy regimens (hexa-caf vs chap-5) in advanced ovarian carcinoma. *The Lancet*, 2:594–600.

[Neijt et al., 1987] Neijt, J., Ten Bokkel Huinink, W., and Van der Burg, M. e. a. (1987). Randomized trial comparing two combination chemotherapy regimens (chap-5 vs cp) in advanced ovarian carcinoma. *J. Clin. Oncol.*, 5:1157–1168.