On the Use of Artificial Neural Networks for the Analysis of Survival Data

Stephen F. Brown, Alan J. Branford, and William Moran, *Member, IEEE*

***Abstract*— Artificial neural networks are a powerful tool for analyzing data sets where there are complicated nonlinear interactions between the measured inputs and the quantity to be predicted. We show that the results obtained when neural networks are applied to survival data depend critically on the treatment of censoring in the data. When the censoring is modeled correctly, neural networks are a robust model independent technique for the analysis of very large sets of survival data.**

***Index Terms*— Backpropagation, neural-network applications, statistics, survival analysis.**

I. Introduction

SURVIVAL analysis is the study of experiments which are performed to measure the amount of time that elapses until a particular event occurs. Examples are measurements of the lifetimes of industrial components or measurements of the time between onset of a particular disease and the patient's death from that disease. The time event can only occur once for a given subject and is usually described as the subject's *failure time*. Survival analysis is normally performed to study how the measured properties of each subject, conventionally called their inputs in the neural-network literature or covariates in the statistical literature, affect their survival time and/or can be used to predict the survival time for new subjects.

The analysis of survival data is usually more complicated than it might first appear because of the presence of *censored* data. Ideally, each subject would be observed until they fail; however, this is not always possible. For example, some subjects may not have reached their failure time when the study is terminated or some patients in a medical study may die from causes unrelated to the disease being studied. The time at which a subject ceases to be observed for some reason other than failure is called the subject's *censoring time*. All that can be inferred about the failure time of a censored subject is that it is greater than their censoring time. There are well studied methods for the statistical analysis of survival data and we present the basic statistics principles needed for the remainder of the paper in Section II.

Artificial neural networks (ANN's) have been applied to an increasing number of prediction and classification problems in recent years (see, for example, Hertz *et al.* [5]) and there is some interest in how ANN's can be used to predict survival times. In Section III we examine a published approach, [2], [4] to applying ANN's to survival analysis, which uses an *ad hoc* method to deal with censored subjects, and show it has serious shortcomings in its handling of censored data. In Section IV we describe a new approach for predicting survival times using neural networks that borrows several significant ideas from survival statistics and handles censored data in a natural way. In Section V the method is applied to some example data sets.

Manuscript received May 11, 1995; revised May 16, 1996 and April 2, 1997.

The authors are with the Department of Mathematics and Statistics, Flinders University, and the Cooperative Research Centre for Sensor Signal and Information Processing, Adelaide SA 5001, Australia.

Publisher Item Identifier S 1045-9227(97)04908-4.

II. Some Results from Survival Statistics

We present only the very basic principles of the statistical analysis of survival data here. The interested reader is referred to Kalbfleisch and Prentice [6] and Cox and Oakes [3] for a comprehensive treatment of the subject.

The failure time for each subject is modeled as a nonnegative random variable T, the distribution of which is statistically independent of the failure times for the other subjects. The distribution of T will depend on the values of the inputs; thus two subjects with the same values for the inputs will have the same probability distribution for their failure times. Subjects with different values for the inputs will, in general, have different failure time distributions. The only restriction that is placed on the censoring mechanism is that censored subjects be representative of the population at risk [6]. For example, subjects cannot be removed from a study simply because it appears that they will shortly fail.

A useful characterization of the distribution of failure times is the survival function which is defined as

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

and is the probability of the patient surviving to time t. It is a nonincreasing function of t with $S(0) = 1$ and $S(\infty) = 0$. Also of interest is the hazard function, $h(t)$, which is defined as

$$h(t) = \lim_{\Delta t \to 0+} \frac{\Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}. \tag{2}$$

The hazard function can be interpreted as the instantaneous rate of failure at time t given that the subject has survived to time t. It is a nonnegative function of t. The survival function can be written in terms of the hazard function as

$$S(t) = \exp\left\{ - \int_0^t h(t')dt' \right\}. \tag{3}$$

For a homogenous population, the survival curve can be estimated using the maximum likelihood approach first given by Kaplan and Meier [7]. If we order the failure times in

ascending order so that t_k is the kth failure time then the Kaplan–Meier estimate of the survival function is

$$\hat{S}(t) = \prod_{k \mid t_k < t} \left(\frac{n_k - d_k}{n_k} \right) \tag{4}$$

where d_k is the number of subjects that fail at time t_k and n_k is the total number of subjects that fail or are censored at time t_k or later.

For studies undertaken with subjects drawn from an inhomogeneous population, the most widely used statistical approach is Cox's proportional hazards model in which the hazard function is modeled as the product of an arbitrary baseline hazard function, $h_0(t)$, and exponential terms for each input. Each input is assumed constant with respect to time. If we let $\mathbf{x}$ be a vector of inputs and β be a vector of parameters then the model hazard function is given by

$$h(t; \mathbf{x}) = h_0(t) \exp(\beta^T \mathbf{x}). \tag{5}$$

The resulting survival curves can be written as

$$S(t; \mathbf{x}) = [S_0(t)]^{\exp(\beta^T \mathbf{x})} \tag{6}$$

where

$$S_0(t) = \exp\left\{ - \int_0^t h_0(t') dt' \right\}. \tag{7}$$

The regression parameters β are estimated using the method of partial likelihood without the need to know or estimate the baseline hazard function. Once the regression parameters have been determined, the baseline hazard can be constructed using a maximum likelihood approach. It is not necessary for us to study the details of these fitting procedures here.

III. Previous ANN Approaches

A previous study of survival data with ANN's has attempted to predict the subject survival time directly from their given inputs. The procedure used to train the neural network incorporates an *ad hoc* technique to treat the censored subjects. Before examining this scheme, we will consider a simple test case to illustrate the biases that can be introduced when a poor choice is made for the treatment of censored data.

We consider a hypothetical set of failure and censoring times drawn from observations of a homogeneous population, that is, the ANN input is the same for each subject. In this case, the ANN will produce the same output, t_{out}, for all the subjects. During training, the ANN weights are adjusted to minimize the difference between the net output and the training outputs, which are simply taken to be the sample failure and censoring times. The error to be minimized during training is usually

$$E = \frac{1}{2} \sum_{i=1}^{m} (t_i - t_{\text{out}})^2 \tag{8}$$

where the t_i are the training set outputs and m is the sample size. It is straightforward to show that the minimum occurs when

$$t_{\text{out}} = \sum_{i=1}^{m} t_i / m. \tag{9}$$

Clearly, the ANN has learned the mean of the training outputs, that is, the mean of the observed failure and censoring times. As a subject's censoring time is less than its failure time, the net will have learned a crude lower bound for the mean failure time of the sample. The magnitude of this bias in the net output will depend on the amount and distribution of the censoring and cannot be estimated from the data. Note that the bias is introduced because we are combining two different quantities, censoring and failure times. Using an error criteria other than least squares may change the neural-network estimate but cannot remove the bias introduced by the censoring process.

Choong *et al.* [2] and deSilva *et al.* [4] have used ANN's to study skin and breast cancer mortality. They train the network using both the failure and censoring times. However, if a subject is censored and the ANN predicts a survival time greater than the censoring time, the error for the prediction is taken to be zero and the network weights are not updated. This will remove some of the bias introduced by the censored data but we still have no way of assessing the size of the bias that remains.

To see more clearly the effect of the biases discussed here, we have performed a simulation where subjects were randomly drawn from a population with an exponential survival time distribution

$$S(t) = e^{-t/\tau} \tag{10}$$

where $\tau > 0$ is an arbitrary parameter. The value of τ was varied from five to 150 in steps of five. For each value of τ, 100 failure times were drawn randomly from the distribution and 100 censoring times randomly drawn from numbers uniformly distributed between zero and 150. If the censoring time for a given subject is less than the failure time then the subject is considered to be censored. Furthermore, any subjects that have not failed or been censored before the elapsed time reaches 100 are censored at that time.

The results of the simulation are shown in Fig. 1. For very small values of τ, say $\tau < 30$, very few subjects are censored and the ANN estimates are a reasonable estimate of the true mean of the distribution. For larger values of τ, censoring begins to play a larger part in the observations and the biases in the ANN estimates are significant. For values of τ between 30 and 60, the ANN estimates deviate from the true mean but remain close to the true median. However, for values of τ above 60, the ANN estimates have diverged from both the true mean and median. Also plotted in the diagram are the mean of the failure and censoring times from each data set. As expected, the ANN lifetime estimates are less biased than those obtained by simply taking the mean of all the observed times.

It should be noted that, for the exponential and other skewed distributions, even if the mean could be accurately calculated from the sample times, the median is a much better point lifetime estimator. For the exponential distribution, only 37% of the subjects will have a failure time greater than the mean.

The biases we have shown here were not picked up by the cross validation undertaken in [2] and [4] because their cross validation score contained many of the same biases as the fitting procedure.

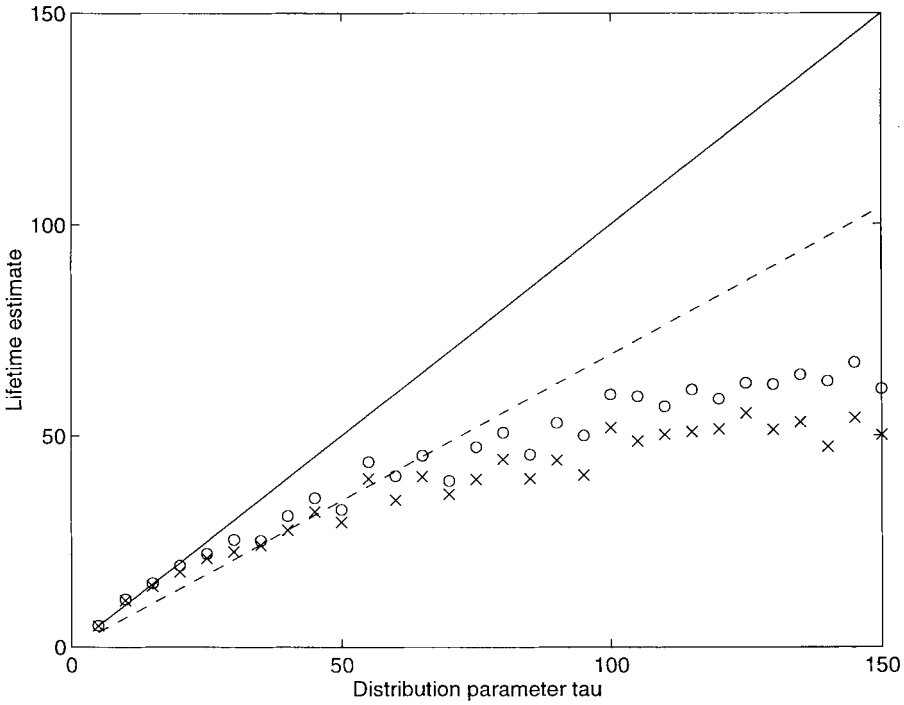

Fig. 1. An example of the biases present in existing ANN estimates of survival times.The solid and dashed lines are, respectively, the true mean and median of the distribution. The circles are the lifetimes as estimated using the method of Choong *et al.* [2] and the crosses are the means of the observed failure and censoring times.

IV. A New Approach

The results in the previous section clearly illustrated that, if ANN's are to become competitive with survival statistics, we must use a rigorous approach making no *ad hoc* assumptions about the failure times of the censored subjects.

Instead of attempting to find a point estimator of a subject's lifetime, we estimate the complete survival curve $S(t)$ for the subject. The survival curve estimate is constructed from a hazard function, calculated using an ANN. To make the system tractable for ANN's, we discretize the elapsed time in units of width Δt. The inputs for the ANN are the subject's inputs and the jth output is the estimated hazard at time $j\Delta t$.

We begin by writing the discretized version of (3) as

$$\tilde{S}(t_j) = \exp\left\{-\sum_{k=1}^{j} \tilde{h}_k\right\} \tag{11}$$

where the discretized hazards $\tilde{h}_k$ are all nonnegative. We transform (11) to a product form as follows:

$$\tilde{S}(t_j) = \prod_{k=1}^{j}(1-h_k) \tag{12}$$

$$= \tilde{S}(t_{j-1})(1-h_j) \tag{13}$$

where the hazard components $h_j = 1 - \exp(-\tilde{h}_j)$ are in the range [0,1] and $\tilde{S}(0) = 1$. Because these hazard components are bounded above and below, they are much more suitable for neural-network training than the unbounded hazards in (11). All that remains is to specify how the observed failure and censoring times can be turned into hazard components that can be used to train the ANN.

For each subject i we associate an empirical survival function $S_i(t)$. If the subject fails at time t_f, then

$$S_i(t) = \begin{cases} 1 & t \le t_f \\ 0 & t > t_f. \end{cases} \tag{14}$$

Comparing with (13) leads to the requirement that the empirical hazard components for this subject, h_j^i, must be zero for $j < j_{\text{crit}}$ and one for $j = j_{\text{crit}}$, where j_{crit} is the smallest value of j such that $t_f < t_j$. There is no constraint on the hazard components at later times because the survival function is already zero and is unchanged by further terms in the product formula.

If the subject is censored at time t_c, the empirical survival function is

$$S_i(t) = \begin{cases} 1 & t \le t_c \\ \text{unknown} & t > t_c. \end{cases} \tag{15}$$

The hazard components, h_j^i, are therefore zero for $j < j_{\text{crit}}$, where j_{crit} is the smallest value of j such that $t_c \le \frac{1}{2}(t_{j-1} + t_j)$, that is, if a subject has survived for more than half the time interval before being censored, we consider them to have survived to the end of the interval. The empirical survival curve imposes no constraint on the hazard components at later times.

When the network is being trained, the error at any output node presented with an undefined empirical hazard is set to zero, preventing the undefined hazards from updating the network weights.

Once the network has been trained, the estimated survival curve can be calculated for any given set of subject inputs. If a point lifetime estimate is required, the estimated median

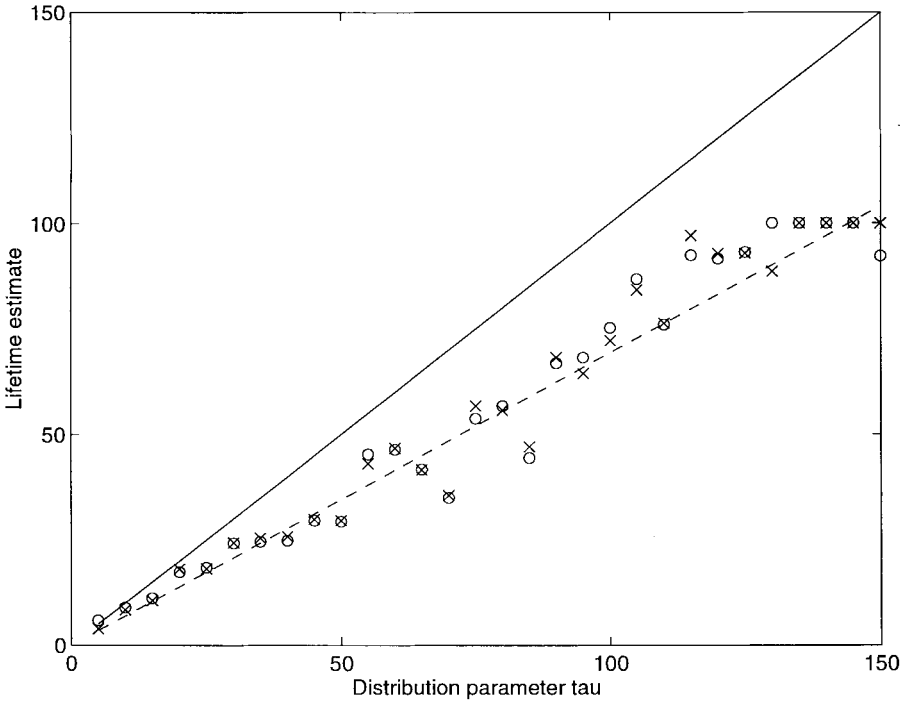

Fig. 2. The solid and dashed lines are, respectively, the true mean and median of the distribution. The circles are the median lifetimes as estimated using our ANN analysis and the crosses are the median lifetimes as estimated using Kaplan–Meier analysis.

lifetime is simply the time at which $\tilde{S}(t)$ equals 1/2. If the derived probability of survival at the maximum time is greater that 1/2, then the median must occur at an unknown time past the experiment duration.

To illustrate how this formulation works, we will again consider the case of a hypothetical set of subjects drawn from the same population and assume that the ANN used to estimate the hazard function has found the true minimum of the error function. At each output node j the calculated hazard h_j will be the minimum of the following error term:

$$E = \frac{1}{2}\sum_{i=1}^{m}\left(h_j^i - h_j\right)^2 \tag{16}$$

where m is the sample size. Recall that subjects that have failed before time $j\Delta t$ or been censored before time $(j-\frac{1}{2})\Delta t$ have unconstrained hazard components h_j^i that are deemed to give zero error. We can therefore expand (16) as

$$E = \left(\frac{1}{2}\sum_{k=1}^{n_f}(1-h_j)^2 + \frac{1}{2}\sum_{k=1}^{n_c}(0-h_j)^2 + \frac{1}{2}\sum_{k=1}^{n}(0-h_j)^2\right) \tag{17}$$

where n_f is the number of subjects with failure times between $(j-1)\Delta t$ and $j\Delta t$, n_c is the number of subjects that are censored between times $(j-\frac{1}{2})\Delta t$ and $j\Delta t$, and n is the number of subjects that have not failed or been censored by time $j\Delta t$. At the minimum we have

$$h_j = \frac{n_f}{n + n_f + n_c}. \tag{18}$$

This gives a survival curve that is very similar to the Kaplan–Meier maximum likelihood estimate for the survival function. It is commonly called the life table estimate. Breslow and Crowley [1] have shown that, provided the elapsed time over the duration of the experiment is divided into at least ten intervals, the bias in the life table analysis is negligible.

For an inhomogeneous population, there is no corresponding theoretical result for us to use to authenticate our method and we must rely on empirical evidence.

It should be noted that the only conditions we have placed on the ANN architecture is that the output neurons have the conventional sum of squares error criterion, as given by (16), and that there are no weight updates when the hazard component is unconstrained by the subject's failure or censoring time. The ANN outputs can lie between any two limits during training but must be scaled into the range [0,1] before the survival function is calculated. If the ANN's output layer activation function has asymptotes at either zero or one, it is important that the training outputs be scaled into the region between the asymptotes otherwise the ANN may be very slow to converge during training. All of the ANN's used in the next section were multilayer perceptrons with a single hidden layer and sigmoidal activation functions. They were batch trained using a conjugate gradients algorithm.

V. Evaluation of the Method's Performance

In Section III we demonstrated the biases present in an earlier ANN technique by applying it to data simulated from a homogenous population. In Section IV we argued theoretically that our new approach should be able to produce the life table estimate, an approximation to the Kaplan–Meier estimate, in this case. We now apply the new method to the data set

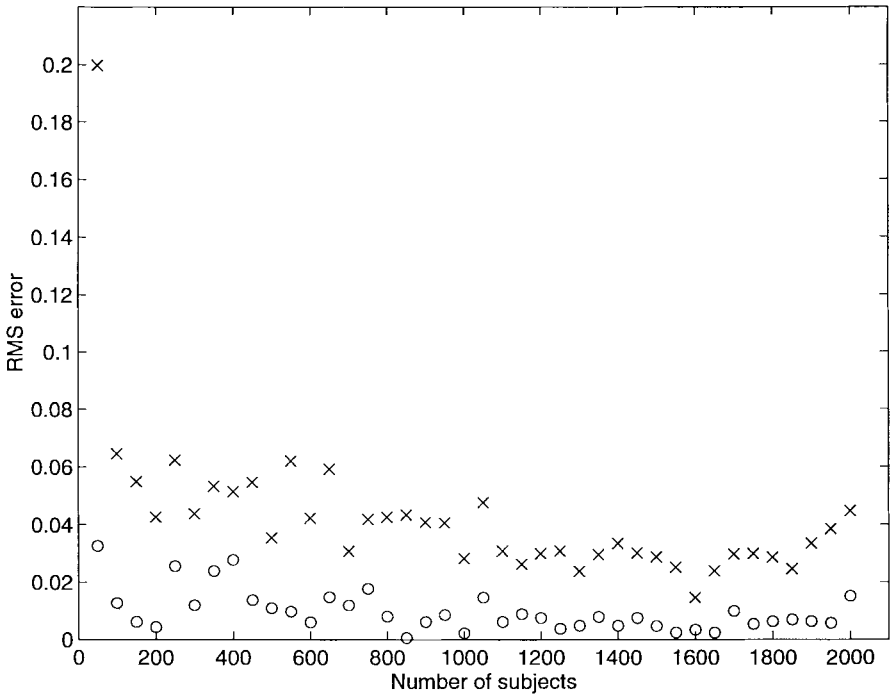

Fig. 3. Fitting to a proportional hazards model. The crosses are the rms errors between the ANN survival estimate and the true survival distribution. The circles are the rms errors between the Cox regression estimate and the true survival distribution. The number given on the abscissa is the total number of subjects: the ANN and Cox model were trained using half this number of subjects.

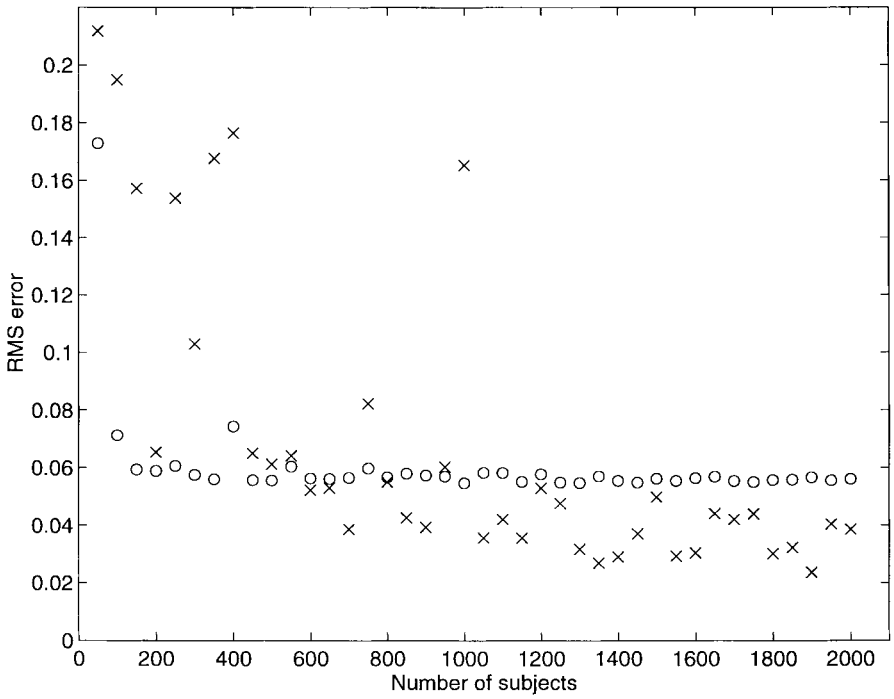

Fig. 4. Fitting to a nonproportional hazards model. The crosses are the rms errors between the ANN survival estimate and the true survival distribution. The circles are the rms errors between the Cox regression estimate and the true survival distribution. The number given on the abscissa is the total number of subjects: The ANN and Cox model were trained using half this number of subjects.

analyzed in Section III. The ANN used here had a single-hidden-layer neuron, ten output neurons, and was trained for ten epochs. As the subjects are assumed to be from a homogeneous population they all have the same inputs to the ANN. For convenience, each subject had a single ANN input which was set to zero. Fig. 2 is a plot of the median lifetime as estimated by the ANN and by Kaplan–Meier analysis. Comparison with Fig. 1 shows that the performance of the new ANN method is clearly superior to that of the previous ANN approach. The estimated median reliably tracks the true median

over the full range of distributions modeled. The difference between the ANN and Kaplan–Meier estimates of the median is very small, as predicted by Breslow and Crowley [1].

We now test the method using subjects drawn from populations with a single input x, which is constrained to lie between -1 and one. We first consider a population which has a survival function given by

$$S(t,x) = [\exp(-t^2/\tau)]^{\exp(x)} \tag{19}$$

where τ was set to 5000. Observe that this distribution has the proportional hazards property. Subjects were drawn from this distribution and censored according to the scheme given in Section III. The number of subjects used was increased from 50 to 2000 in steps of 50. Each set of subjects was split randomly into two sets, a training and a testing set, each containing half of the total sample. ANN's with ten output neurons and from one to six hidden layer neurons were then trained using the training set. The optimal number of hidden layer neurons and training epochs for each data set was determined by cross validating against the testing set and choosing the configuration with the smallest testing set error. For the data sets with a small number of subjects, the optimum ANN typically had from one to three hidden layer neurons. For the larger data sets, those with more than 300 training subjects, the optimum ANN typically had from three to six hidden layer neurons. The ANN with the smallest cross validation error was used to generate survival curves for 21 equally spaced values of x and each evaluated at elapsed times of 10, 20, $\cdots$, 100 units. We then calculate the rms difference between these survival estimates and the true distribution. As a check, we also used Cox regression to fit the training data and then calculated the rms error between the Cox estimate of the survival distribution and the true distribution.

Fig. 3 shows the rms errors of the ANN and Cox model as a function of the number of subjects. The survival distribution used to generate the data is of the correct form for the Cox model and the fit between the model and true distributions is correspondingly good. The quality of the ANN estimate depends on the number of subjects available, with the estimate improving as the number of subjects increases, and even with 2000 subjects, the rms error can be as large as 0.05.

The previous test was then repeated with subjects drawn from a population that has a survival distribution given by

$$S(t,x) = [\exp(-t^2/\tau)]^{\exp(2x^2-1)} \tag{20}$$

where τ was again set to 5000. Observe that this distribution does not have the proportional hazards property. The results from this experiment are shown in Fig. 4. Because the survival function is more complicated than in the previous case, the ANN survival estimates need more samples to reach a given rms error level than before. However, when a sufficiently large training set is used, the ANN estimates converge to the solution as before. The Cox model fit to the data is poor for this case, as would be expected. However, it must be pointed out that in practice an analysis of the residuals after fitting the Cox model would lead the experimenter to try nonlinear factors in the proportional hazards model, usually including a quadratic term, which would then produce a very good fit to the sample data.

These two experiments show that, provided a sufficiently large data set is available, the ANN approach described here can model inhomogenous survival distributions.

VI. Conclusion

We have shown that the correct treatment of censored data is crucial when analyzing survival data. ANN approaches for survival analysis that use *ad hoc* methods to treat censored data can lead to significant biases in the estimated subject lifetimes. Our approach of learning an approximation to the hazard function has been shown to give the optimum results in simple cases, giving confidence in the ANN survival estimates obtained in more general applications.

Nonetheless, statistical analysis of survival data, using Cox regression, performs much better than the ANN analysis, particularly when the number of subjects in the sample is small, say less than 1000. Therefore, we recommend that model-based analysis in general, and Cox regression in particular, remain the method of choice for survival analysis and that the ANN techniques we have derived here only be used when a large sample is available and an acceptable model cannot be found.

References

[1] N. Breslow and J. Crowley, "A large sample study of the life table and product limit estimates under random censorship," *Ann. Statist.*, vol. 2, pp. 437–453, 1974.
[2] P. L. Choong, C. J. S. deSilva, J. Taran, P. Heenan, and H. Dawkins, "Survival analysis using artificial neural networks," in *Proc. 1st Australia and New Zealand Conf. Intell. Inform. Syst.*, 1993, pp. 283–287.
[3] D. R. Cox and D. Oakes, *Analysis of Survival Data*. London: Chapman and Hall, 1984.
[4] C. J. S. deSilva, P. L. Choong, and Y. Attikiouzel, "Artificial neural networks and breast cancer prognosis," *Australian Comput. J.*, vol. 26, pp. 78–81, 1994.
[5] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation*. Redwood City, CA: Addison-Wesley, 1991.
[6] J. D. Kalbfleisch and R. L. Prentice, *The Statistical Analysis of Failure Time Data*. New York: Wiley, 1980.
[7] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *J. Amer. Statist. Association*, vol. 53, pp. 457–481, 1958.

Stephen F. Brown received the B.Sc. degree (Hons.) in physics from the University of Tasmania, Australia, in 1984. He received the M.Sc. and Ph.D. degrees from the University of Sydney, Australia, in 1989 and 1993, respectively, for research into plasma processes on the surface of the Sun.

He is currently a Research Scientist at the CSIRO Division of Telecommunications and Industrial Physics, Sydney, Australia. His research interests include medical image processing and computer-aided diagnosis.

Alan J. Branford received the B.Sc. degree (Hons.) in applied mathematics from the University of Adelaide, Australia, in 1979. He received the M.Sc. degree in 1980, also from the University of Adelaide, for research in applied probability in the area of birth-and-death processes, and the Ph.D. degree from the Statistical Laboratory of the University of Cambridge, U.K., in 1983, studying paroxysmal phenomena in stochastic models involving indirect feedback.

He is a Senior Lecturer in Statistical Science at Flinders University, Adelaide, Australia, and is a Member of the Cooperative Research Centre for Sensor Signal and Information Processing (CSSIP). His interests include biostatistics, particularly the analysis of survival data, and image processing.

William Moran (M'95) received the B.Sc. degree in mathematics from the Universityof Birmingham, U.K., and the Ph.D. degree from the University of Sheffield, U.K.

He is Professor of Mathematics at Flinders University, Adelaide, Australia, and Leader of the Analytical Techniques and Medical Diagnostics programs at the Cooperative Research Centre for Sensor Signal and Information Processing (CSSIP).